Redefining gene distributions in *K. pneumoniae* and *E. coli* using large public datasets

Gal Horesh
Corpus Christi College, University of Cambridge

The dissertation is submitted for the degree of
Doctor of Philosophy

June 2020

To my parents, who got me here,

and to Harry, who walked with me every step of the way.

Declaration

The work presented was carried out at the Wellcome Sanger Institute between October 2016 and June 2020. This work is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. This work is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text.

This thesis does not exceed the prescribed word limit specified by the Biology Degree Committee.

Abstract

The work in this thesis is concerned with characterising genes and their distributions in *Escherichia coli* and *Klebsiella pneumoniae*. While both *K. pneumoniae* and *E. coli* are found in the guts of healthy individuals, as well as in animals and in the environment, they are particularly relevant organisms to study, as they represent key players in the dissemination of drug resistance and virulence in bacterial populations. Both organisms were given the highest priority by the World Health Organisation as organisms that pose the greatest threat to human health due to high levels of drug resistance. Additionally, they are both the leading cause of life-threatening extra-intestinal disease worldwide. Finally, some *E. coli* variants are also a major cause of severe diarrheal disease, most commonly in the developing world.

The phenomena that is driving these issues is horizontal gene transfer (HGT); the process by which new genetic material is introduced into a genome from an outside source. Drug resistance is most commonly driven by gene acquisition, and it is through the acquisition of virulence genes that *K. pneumoniae* and *E. coli* can cause disease. Indeed, HGT has been estimated to occur in high rates in *K. pneumoniae* and *E. coli*. Both are highly diverse organisms with very large gene pools and multiple co-circulating lineages. These facts make studying their gene pools on large scales highly relevant, as new genes and lineages are continuously discovered with the sequencing of new genomes.

The aim of this thesis was to utilise the availability of large public genomic datasets to study the gene pools of *K. pneumoniae* and *E. coli* on a scale and resolution not previously possible. Initially, the distribution of toxin-antitoxin (TA) systems was investigated in a collection of 259 *K. pneumoniae* isolates. TA systems are operons where one gene encodes for a toxin which inhibits a cellular process, and the other is an antitoxin which inhibits the toxin's activity. TA systems are relevant to study in the context of HGT as they have been shown to play a role in the maintenance of resistance and virulence genes and to contribute to antibiotic tolerance. The analysis on TA systems in *K. pneumoniae* revealed new insights regarding the distribution TA systems in the species. These insights were then expanded to examine the distribution of all genes of the *E. coli* gene pool in a collection of thousands of genomes. This analysis revealed that genes from different categories undergo different dynamics of gene gain and loss, as well as exposed *E. coli* lineages which may be important in their contribution to gene flow in the population. Due to the novelty and scope of the analyses presented, new computational tools and approaches were developed and are presented.

Acknowledgements

I would like to thank my supervisors, Nick Thomson and Eva Heinz, who have guided, inspired and encouraged me throughout my PhD. In particular, to Nick for teaching me that every situation, no matter how bad, is an opportunity, and to Eva for convincing me that there are no problems in life, only challenges.

I am also very grateful to Leopold Parts, for his honest feedback, useful advice and consistent support. The other members of my thesis committee: Andres Floto, Simon Harris and Jukka Corrander, have given me insightful discussions and constructive criticism. Julian Parkhill and Matt Berriman helped me shape this research by thoroughly questioning my PhD plans during my first year viva.

The work on toxin antitoxin systems would not have been possible without the contributions from my collaborators: Cinzia Fino, Alexander Harms and Kenn Gerdes. Cinzia worked extremely hard on all of the projects that we collaborated on. She was a great student and taught how much I love to teach. Matthew Dorman helped to coordinate the experiments and performed some experimental work himself. Additionally, he has been my considerate office neighbour and my PhD-cohort companion for the last three years.

Other collaborations, beyond the scope of this thesis, have furthered my learning and scientific progression. Gerry Tonkin-Hill and Neil MacAlasdair involved me in the Panaroo project and helped me with the pan-genome analysis. Grace Blackwell and Zam Iqbal entrusted me with the large scale transposon study which has been a good experience.

The work I have achieved would not have been possible on this scale without the help of the Pathogen Informatics team as well as the members of the Sanger Service Desk. I would particularly like to thank Martin Hunt who helped me to build SLING into a package.

For helping me get through my PhD with a smile on my face, with well needed walks and microwave chats, I would like to thank Alex Wailan, Kate Mellor, Alyce Tyler-Brown, Aline Cuenod, Ha My Pham, Sushmita Sridhar, Grace Blackwell and the rest of the members of Team 216 along the years. I am happy to say I have made some good friends along the way.

Last but not least, I would like to thank my partner, Harry Scholes, for unwavering support and all the work-related dinner conversations along the way. My PhD would not have been the same without him.

Finally, I would also like to extend my gratitude to Wellcome for funding my PhD.

Publications

SLING: a tool to search for linked genes in bacterial datasets, Horesh et al., Nucleic Acids Research, 2018, https://doi.org/10.1093/nar/gky738

Type II and type IV toxin–antitoxin systems show different evolutionary patterns in the global *Klebsiella pneumoniae* population, Horesh et al., *Nucleic Acids Research*, 2020, https://doi.org/10.1093/nar/gkaa198

A comprehensive and high-quality collection of *E. coli* genomes and their genes, Horesh et al., in preparation

A pan-genome analysis of 10,000 *E. coli* genomes reveals new patterns of gene sharing between lineages, Horesh et al., *in preparation*

Producing Polished Prokaryotic Pangenomes with the Panaroo Pipeline, Tonkin-Hill et al., bioRxiv, 2020, https://doi.org/10.1101/2020.01.28.922989

The distribution of toxins containing Gp49 across Gram-negative bacteria, Fino et al., *in preparation*

Horizontal and vertical spread of Tn1-related transposons in Gram-negative bacteria, Blackwell et al., *in preparation*

Table of Contents

List of Figures

List of Tables

Glossary

aa	amino acids
ACCTRAN	Accelerated Transformation
aEPEC	atypical Enteropathogenic *E. coli*
AIEC	Adherent Invasive *E. coli*
AMR	Antimicrobial Resistance
ANI	Average Nucleotide Identity
BLAST	Basic Local Alignment Search Tool
bp	basepairs
BSI	Bloodstream Infection
CDC	Centers for Disease Control and Prevention
CDS	Coding Sequence
COG	Clusters of Orthologous Groups
contig	contiguous assembled sequence
DAEC	Diffusely Adherent *E. coli*
EAEC	Enteroaggeragive *E. coli*
ECOR	*E. coli* Reference Collection
EHEC	Enterohaemorrhagic *E. coli*
EIEC	Enteroinvasive *E. coli*
ENA	European Nucleotide Archive
EPEC	Enteropathogenic *E. coli*
ESBL	Extended Spectrum Beta Lactams
ETEC	Enterotoxigenic *E. coli*
FDA	Food and Drug Administration
FDR	False Discovery Rate
GEMS	The Global Enteric Multicenter Study
HGT	Horizontal Gene Transfer
HMM	Hidden Markov Model
HUS	Hemolytic uremic syndrome
hvKp	hyper-virulent *K. pneumoniae*
ICE	Integrative and Conjugative Elements
IncA/C	Plasmid incompatibility type A/C
IPTG	isopropyl β-D-thiogalactopyranoside
LB	Lysogeny Broth
LEE	Locus of Enterocyte Effacement

Mbp	Million basepairs
MDR	Multidrug resistant
MFP	Membrane Fusion Protein
MGE	Mobile Genetic Element
MLST	Multi-locus Sequence Type
MSA	Multiple Sequence Alignment
NCBI	National Center for Biotechnology Information
ND	Not Determined
OMF	Outer Membrane Protein
PBS	phosphate-buffered saline
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PHE	Public Health England
PopPUNK	Population Partitioning Using Nucleotide K-mers
PSK	Post Segregational Killing
QC	Quality Control
RND	Resistance-Nodulation-Division
SNP	Single Nucleotide Polymorphism
SSN	Sequence Similarity Network
ST	Sequence Type
ST10	Assigned to Sequence Type 10
STEC	Shiga toxin-producing *E. coli*
TA	Toxin Antitoxin
TADB	Toxin Antitoxin Database
UTI	Urinary Tract Infection
VTEC	Verotoxigenic *E. coli*
WGS	Whole Genome Sequencing
WHO	World Health Organisation

1 Introduction

The *Enterobacteriaceae* are a taxonomic family of Gram-negative, rod shaped, facultative anaerobic bacteria, containing over 50 genera and 290 species [1]. This family includes some of the most clinically important genera of bacteria which are responsible for disease both in humans and other animals [2]. Two of these are *Escherichia* and *Klebsiella.* While the species *Escherichia coli* and *Klebsiella pneumoniae* share fewer than 55% genes and are only 82% identical in sequence [3], they share characteristics that make them highly relevant to study in today's world where the dissemination of drug resistance genes and virulence genes within bacterial populations is a growing issue. Both species are found across niches in the guts of healthy individuals, animals and the environment [4], however, both can cause severe life-threatening disease [5,6]. Indeed, both *K. pneumoniae* and *E. coli* are the leading causes of urinary tract infections (UTIs), bloodstream infections (BSIs) and meningitis [5,7,8]. Both species are ubiquitous worldwide, with multiple lineages co-circulating across different geographic locations [9,10]. Both species have large gene pools which further enhances the diversity in their populations as co-circulating lineages possess different sets of genes [9,11]. The availability of a large gene pool enables quick adaptation in fast-changing environments and in response to new pressures. The combination of these factors: an ability to adapt, global distribution and omnipresence both in the environment and in the human gut, have made both these two species prime players in the dissemination of genes that confer resistance and virulence worldwide. Indeed, both are included in the "ESKAPEE" pathogens, pathogens which are able to "escape" treatment with antibiotics [12,13]. They are also the highest priority on the World Health Organisation's (WHO) list of priority pathogens that pose the highest threat to human health due to high levels of resistance [14]. The availability of large datasets of these organisms provide the opportunity to examine their gene pools on a scale and resolution that was not possible before, along with the necessity to develop new approaches - these will be presented and explored in this thesis.

1.1 The organisms: *E. coli* and *K. pneumoniae*

1.1.1 The species *K. pneumoniae*

1.1.1.1 Taxonomy and classification

***K. pneumoniae* species complex**

K. pneumoniae was first described in 1882 by Carl Friedlander when the bacterium was isolated from patients who had died from pneumonia [15]. In the last two decades, the species definition has undergone major changes. Today it stands that the *K. pneumoniae* species complex contains five unique species and seven subspecies (Table 1.1). The three species in the species complex which have been known the longest are *K. pneumoniae* sensu stricto, *K. quasipneumoniae* and *K. variicola.* These were originally identified in 2001 by Brisse and Verhoef as three phylogroups that make up the *K. pneumoniae* population and were coined as phylogroups KpI, KpII and KpIII respectively [16]. Whole genome sequencing (WGS) of these has since proved that these are in fact separate species [9]. Since, *K. quasipneumoniae* and *K. variicola* were each divided to include two subspecies (Table 1.1) [17–20]. The remaining two species, KpVI and KpVII were only discovered in the last three years [20,21].

While two new species and one new subspecies have been described since 2017, these species are not well studied. There are five or fewer complete assemblies of these species currently on the National Center for Biotechnology and Information (NCBI) database (Table 1.1). In this thesis, the focus is on the isolates from the three most well studied subspecies; *K. pneumoniae* sensu stricto (KpI), *K. quasipneumoniae* subsp. *quasipneumoniae* (KpII) and *K. variicola* subsp. *variiloca* (KpIII). For clarity, I will refer to these three subspecies by the names *K. pneumoniae* sensu stricto, *K. quasipneumoniae* and *K. variicola* and to all three as *K. pneumoniae* or the *K. pneumoniae* species complex.

Typing *K. pneumoniae*

Traditionally, typing of *K. pneumoniae* was achieved using phenotyping methods, most commonly capsule serotyping of the polysaccharide capsule K antigen and bacteriocin typing [22–24]. Molecular typing methods were also developed, however they were not widely used due to lack of reproducibility [23]. In 2005, a multi-locus sequence typing scheme (MLST) based on seven chromosomal housekeeping genes was established and has since been widely used as the accepted scheme for typing *K. pneumoniae* [23,25]. MLST was proposed in 1998 as a standardised, deterministic and universal scheme for typing microorganisms [26]. *K. pneumoniae* sequence types (ST) are denoted at ST258, for instance for sequence type 258. The Pasteur Institute MLST database for *K. pneumoniae* contains 3,409 unique STs as of January 20th, 2020[1] (Table 1.1).

[1] https://bigsdb.pasteur.fr/klebsiella/klebsiella.html

Table 1.1: *K. pneumoniae* species and subspecies.

Species	Subspecies	Alt. name	Year identified	Number of assemblies on NCBI (on date 9.1.20)	STs
K. pneumoniae sensu stricto	NA	KpI	2001	8810	3132
K. quasipneumoniae	subsp. *quasipneumoniae* subsp. *similipneumoniae*	KpII-A / KpII KpII-B / KpIV	2001 (originally) 2005 (subspecies)	286	75 74
K. variicola	subsp. *K. variicola* subsp. *tropicalensis*	KpIII KpV	2001, 2004 (originally) 2017, 2019 (subspecies)	296	145 0
K. quasivariicola	NA	KpVI	2017	5	0
K. africanensis	NA	KpVII	2019	0	0

1.1.1.2 Pathogenicity and resistance

***K. pneumoniae:* an opportunistic pathogen and a leading cause of hospital acquired infections**

K. pneumoniae is, in most cases, an opportunistic pathogen causing infections in hospital settings amongst immunocompromised patients [5,24]. As of 2014, *K. pneumoniae* was considered the third cause of nosocomial infections in the United States and Europe, being the cause of 3%-8% of all hospital acquired infections [7]. An infection with *K. pneumoniae* can cause a range of severe life-threatening diseases including respiratory tract infections, meningitis, wound infections and blood-stream infections [5]. A recent incidence report by Public Health England (PHE), has shown that *K. pneumoniae* is the second leading cause for Gram-negative BSI in the UK and its prevalence has been increasing (Figure 1.1).

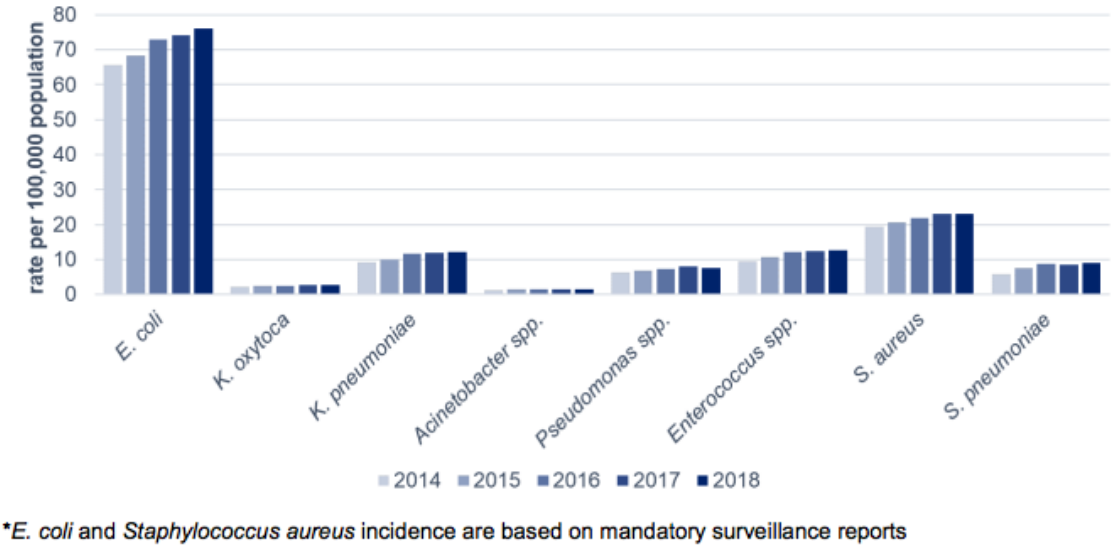

Figure 1.1: Incidence of bloodstream infections caused by eight major pathogens in England, *per 100,000 people from 2014 up to 2018. Taken from English Surveillance Programme for Antimicrobial Utilisation and Resistance (ESPAUR) Report 2018 – 2019.*

K. pneumoniae colonises both the nasopharynx and the intestinal tract of healthy individuals [5,24]. Carriage rates of *K. pneumoniae* have been reported to be higher amongst hospital personnel and patients than in the general public [24]. This leads to high rates of hospital acquired infections which are thought to originate from patients' own gastrointestinal tract, hospital equipment or hospital staff. Additionally, there is rapid transmission within hospital settings between carers, patients and the environment [24]. Therefore, major hospital outbreaks of *K. pneumoniae* commonly occur worldwide [27,28].

Hypervirulent community acquired *K. pneumoniae* infections

In the pre-antibiotic era, *K. pneumoniae* was considered a community acquired infection which predominantly affected alcoholics and diabetics, causing pneumonia, referred to as classical community acquired *K. pneumoniae* [29,30]. However, in the last three decades a hypervirulent *K. pneumoniae* (hvKp) has emerged which causes a new clinical manifestation of pyogenic liver abscess that is able to cause metastatic infections [30]. The first cases were reported in Taiwan in the 80s and 90s, and were followed by incidents in other countries in South-East Asia [29]. This form of community acquired *K. pneumoniae* infection has increased in incidence with time, as well as in other countries worldwide, including in North America and Europe [30,31]. Today it is considered a global threat. Unlike hospital acquired or classical community acquired *K. pneumoniae*, hvKp affects patients which are otherwise healthy [32]. Transmission of community acquired *K. pneumoniae* and hvKp is unclear. It is likely that

infection occurs from colonisation of patients' own gastrointestinal tract, and that transmission occurs via food, water, person to person or animal to person transmission [30].

K. pneumoniae sensu stricto: the leading cause of *K. pneumoniae* infections?

In the past it was considered that *K. pneumoniae* sensu stricto was the leading cause of severe infections by the *K. pneumoniae* species complex, whereas *K. quasipneumoniae* and *K. variicola* were less pathogenic species which were associated with carriage in humans and bovine [9]. However, in recent years it has become evident that isolates belonging to the other species of the *K. pneumoniae* species complex were often misidentified as *K. pneumoniae* sensu stricto [33,34]. Today it is understood that *K. variicola* and *K. quasipneumoniae* also cause severe infections and high mortality rates in humans [33–35]. With that being said, studies which are biased towards sampling from mammalian-infections consistently reveal a similar high prevalence of *K. pneumoniae* sensu stricto relative to isolates belonging to the other two species [36,37].

The new hypervirulence phenotype is mostly restricted to the *K. pneumoniae* sensu stricto, with hvKp mostly belonging to *K. pneumoniae* sensu stricto ST23 [31]. However, *K. quasipneumoniae* and *K. variicola* isolates have also been reported to cause community acquired hvKp infections in as much as 17% of cases [38–40].

Antimicrobial resistance in *K. pneumoniae*

One of the major challenges today is the increase in multidrug resistance which results in untreatable bacterial infections. Resistance has been predicted to be the number one cause of death in the world by 2,050 [41]. According to a 2019 Centers for Disease Control and Prevention (CDC) report, there are over 2.8 million antibiotic-resistant infections and 35,000 deaths as a result occurring in the United States every year[2]. *K. pneumoniae* is one organism for which resistance has increased in prevalence in the last few decades [5]. *K. pneumoniae* is often the first organism in which new antimicrobial resistance (AMR) genes are detected, before they are observed more widely across other pathogens [4]. In the last four years there have been reports worldwide of *K. pneumoniae* isolates which are pan-resistant and are not treatable with any available antibiotic [42–44].

Resistance in *K. pneumoniae* evolves with the use of new antibiotics and thus has been increasing with time [4]. *K. pneumoniae* is intrinsically resistant to beta-lactams such as ampicillin due to a chromosomally encoded beta-lactamases [9,17]. This has likely allowed *K.*

[2] ANTIBIOTIC RESISTANCE THREATS IN THE UNITED STATES, CDC, 2019

pneumoniae to initially spread in hospitals. Since then, resistance has emerged to every line of antibiotics; aminoglycoside in the 70s, followed by Extended Spectrum Beta-Lactams (ESBLs) and fluoroquinolones in the 80s, carbapenems in the 90s and most recently resistance to colistin [4,24]. Today, *K. pneumoniae* is the most common carbapenem resistant member of the *Enterobacteriaceae* causing infections [5].

Resistance has been reported across *K. pneumoniae* sensu stricto, *K. quasipneumoniae* and *K. variicola* lineages [25], with the highest level of resistance attributed to *K. pneumoniae* sensu stricto, particularly within specific MLST lineages such as STs 258, 11, 15 and 101 [5,25,36,37].

Hypervirulent and multidrug resistant *K. pneumoniae*

While resistance and hypervirulence are both major causes for concern within *K. pneumoniae*, these two phenotypes are mostly mutually exclusive, i.e. multidrug resistant (MDR) isolates are generally not hypervirulent [45,46]. In recent years, the emergence of isolates which are both resistant to last line antibiotics, including carbapenems, and are hypervirulent have been reported [45,47,48]. This convergence of resistance and hypervirulence is a major concern as the disease caused by hvKp is severe, affects healthy individuals and would be fatal without the ability to treat with antibiotics [5]. A recent study looking at the evolutionary dynamics of hypervirulent and MDR *K. pneumoniae* found that MDR isolates were highly diverse whereas hypervirulent isolates showed low levels of diversity, suggesting that the convergence of the two phenotypes occurs from the acquisition of virulence factors by resistant isolates, and not vice versa [46].

1.1.1.3 Genetics

***K. pneumoniae* has a large, open pan-genome**

An average *K. pneumoniae* genome is 5.5 Mbp and consists of approximately 5,500 genes [9,25]. Of these genes, fewer than 2,000 represent the core genome of the species, i.e. genes which are present in all isolates [9]. The remainder of the genes are part of the accessory genome, i.e. genes that are only present in some isolates and lineages. The sum of all the genes of the species, core and accessory, is termed the “pan-genome”. A “pan-genome” which increases in size with sampling of new genomes, i.e. more genes are discovered with new isolates is termed “open” (Figure 1.2). The *K. pneumoniae* pan-genome is open and has been estimated to contain at least 30,000 protein-coding genes [9]. Most of the genes making up the *K. pneumoniae* gene pool are rare and found in fewer than 5% of sequenced isolates [9].

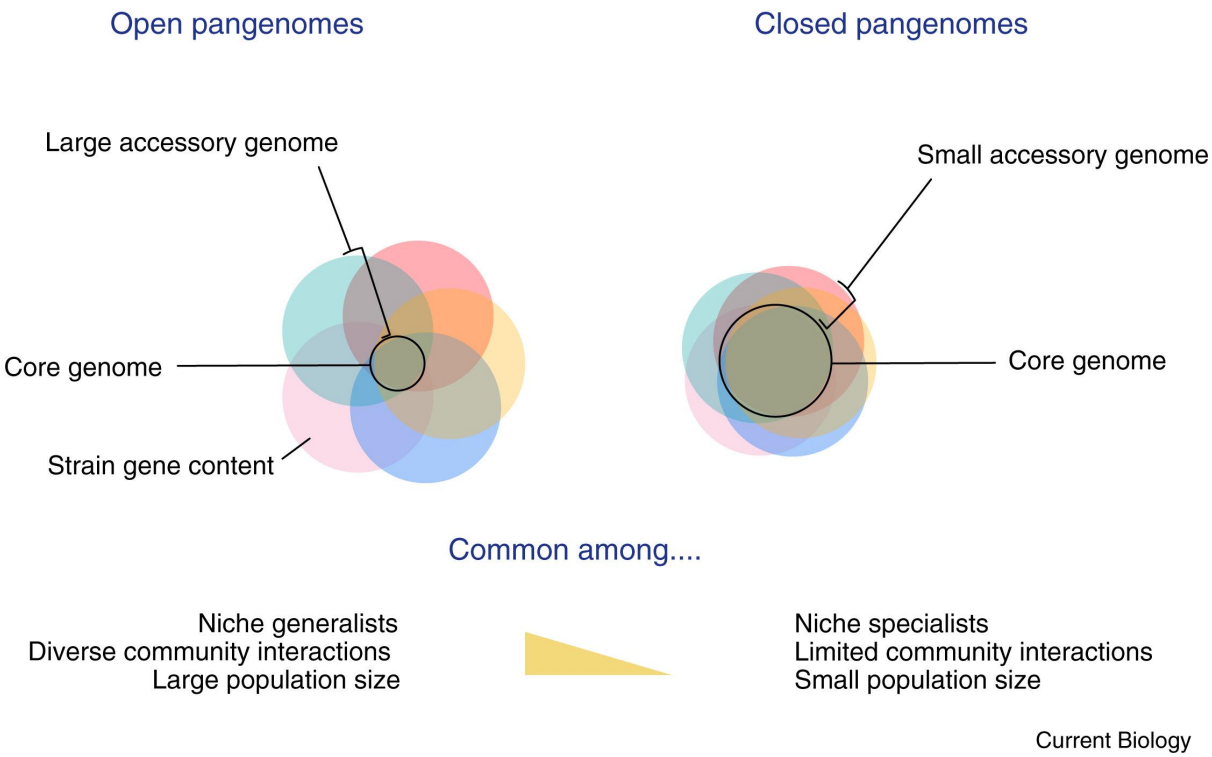

Figure 1.2: Pan-genome definition, *taken from [49]. The Venn diagrams represent the overlaps in gene content between isolates belonging to the same species. The set of all genes shared by all isolates is termed the "pan-genome". Genes which are shared by all isolates are termed "core" and the rest "accessory". The size of the pan-genome varies across species and provides an indication on lifestyle.*

K. pneumoniae sensu stricto, _K. quasipneumoniae_ and _K. variicola_ are separately evolving species.

The three species *K. pneumoniae* sensu stricto, *K. quasipneumoniae* and *K. variicola*, are separately evolving populations with barriers to gene flow between these closely related species [9]. This is indicated by high average nucleotide identity (ANI) within the species compared to between species (96%-97% compared with 99.5% within) [9]. An ANI cutoff of 95%-96% is usually used to define a species [50]. Indeed, there is little evidence of recombination between the three species as high sequence divergence has been shown to affect efficiency of recombination [9,51]. Finally, the accessory gene content has been found to often be species specific [9].

Genetic determinants of virulence

Virulence factors enable *K. pneumoniae* to colonise, multiply and survive within the host [5]. Unlike a "classical" pathogen which actively attacks the host, *K. pneumoniae* virulence is mostly driven by the evasion of the host immune response and biofilm formation [52]. The best

characterised and primary virulence factor is the polysaccharide capsule, which contains the K-antigen, and is determined by the *cps* locus [24,53]. Second, the lipopolysaccharide component (LPS) of the outer-membrane, which contains the O-antigen, is determined by the *rfb* locus [5,54,55]. Both the capsule and the O-antigen allow *K. pneumoniae* to evade immune defenses [5,53,55]. Pili (fimbriae) enable *K. pneumoniae* to adhere to host cells as well as to medical equipment such as urinary catheters [56,57]. Other virulence factors such as type VI secretion systems and siderophores for iron acquisition have also been described [5].

The presence of combinations of these virulence factors and more have been shown to be associated with the hypervirulent phenotype [30]. Multiple studies have found that the capsule type of hvKp is most commonly K1 or K2, suggesting it plays a role in hypervirulence [5,45]. However, there are K1/K2 capsule type isolates which are not hypervirulent, whereas some hypervirulent isolates are non K1/K2 [45]. Siderophores salmochelin and aerobactin, encoded by *iro* and *iuc* respectively, have also been implicated in hypervirulence [9,58,59].

Genetic determinants of resistance

Intrinsic resistance to beta-lactams within the *K. pneumoniae* species complex are orthologous to each *K. pneumoniae* species and has evolved from a common ancestor that diverged with the species; *bla*SHV in *K. pneumoniae* sensu stricto, *bla*OKP in *K. quasipneumoniae* and *bla*LEN in *K. variicola [9,17].* A number of chromosomal alterations lead to reduced susceptibility to antibiotics. These include induced expression of efflux pumps, reduced permeability by loss of outer membrane porins, or mutations in the antibiotic's target. These include mutations in *gyrA* and *parC*, targets of topoisomerase, confer reduced susceptibility to fluoroquinolones [60,61]. Resistance to colistin, a last resort treatment for MDR strains, has been reported due to mutations in genes of the PhoQ/PhoP system, a two component gene system that regulates several cellular activities [62–65]. Resistance to other antibiotics is mostly driven through the accessory genome, with resistance determinants being acquired mostly on plasmids [5,25]. Carbapenem resistance in *K. pneumoniae* is mainly driven by the presence of one of three carbapenemases; OXA-48, NDM-1 and KPC [4,66]. In total, over 400 unique AMR genes have been identified in *K. pneumoniae* [4,9]. In *K. pneumoniae*, plasmid mediated colistin resistance is mostly driven by mobilisable colistin resistance (*mcr*) genes *mcr-1* and *mcr-2* [5,25]. Worryingly, a hvKp carrying *mcr-1* has been reported in China [68].

1.1.2 The species *E. coli*

1.1.2.1 Taxonomy and classification

Typing of *E. coli*

E. coli was first described in 1885 by Theodor Escherich [69]. It has since been the most commonly used laboratory strain as it is easy to grow in culture and to apply genetic manipulations [70]. The diversity of the species in the early days was measured using typing methods similar to those used in *K. pneumoniae*: phage-typing and serotyping [71]. Antibodies against the O antigen on the lipopolysaccharide and the H antigen on the flagellar were the most commonly used approach to distinguish between *E. coli* isolates [71,72]. These serotypes are still used today to identify particular high-risk clones [73,74]. However, serotyping does not inform on the phylogenetic relationships between isolates as through horizontal gene transfer (HGT) two distantly related isolates may have the same serotype and vice versa [75]. In more recent years, multiple MLST schemes for *E. coli* have been proposed [76–79]. Two schemes, Achtman and Pasteur, are still maintained on the public MLST database (https://pubmlst.org/mlst/) [80]. Both schemes are based on different combinations of seven or eight housekeeping genes [79]. The Achtman scheme has proved to be the most widely used and thus will be used in this thesis [79].

Molecular methods define the *E. coli* phylogroups

The first molecular method used to subtype the natural *E. coli* population was multilocus enzyme electrophoresis, a method based on the mobility patterns of enzymes during electrophoresis [81–84]. In 1983, Ochman and Selander collected 72 *E. coli* isolates which they believed represented the diversity of the natural population [85]. The isolates were chosen based on a principal component analysis (PCA) of the allelic variation of 11 enzymes. Additionally, they were chosen to include those collected from different hosts, across geographic locations and to include pathogenic and non-pathogenic variants [85]. These were termed the *E. coli* reference (ECOR) collection. The first phylogenetic analyses of this collection revealed four major phylogenetic groups, termed “phylogroups”, named A, B1, B2, and D and two minor groups named C and E [86,87]. While these were the early days in our understanding of the *E. coli* population structure, the existence of these major phylogroups was confirmed by more advanced methods including WGS [75]. During the years, minor refinements were made to the phylogroups to include additional phylogroups F and more recently G [10,11,78,88,89]. Additionally, five monophyletic cryptic clades, which are indistinguishable from *E. coli* sensu stricto have also been identified [90,91]. More recent publications including even more genomes have suggested that these definitions require even

further expansion [92]. A phylogenetic analysis of 9,479 *Escherichia* genomes, taken from public databases and representing all sequenced STs, was proposed to represent the total known diversity of the genus and was suggested as a new reference collection termed the EcoRPlus collection [93]. Figure 1.3 presents the phylogenetic tree of *E. coli* phylogroups using the isolates from the ECOR collection. Importantly, there have been conflicting results regarding the relationship between the phenotypic characteristics of the phylogroups [11,94].

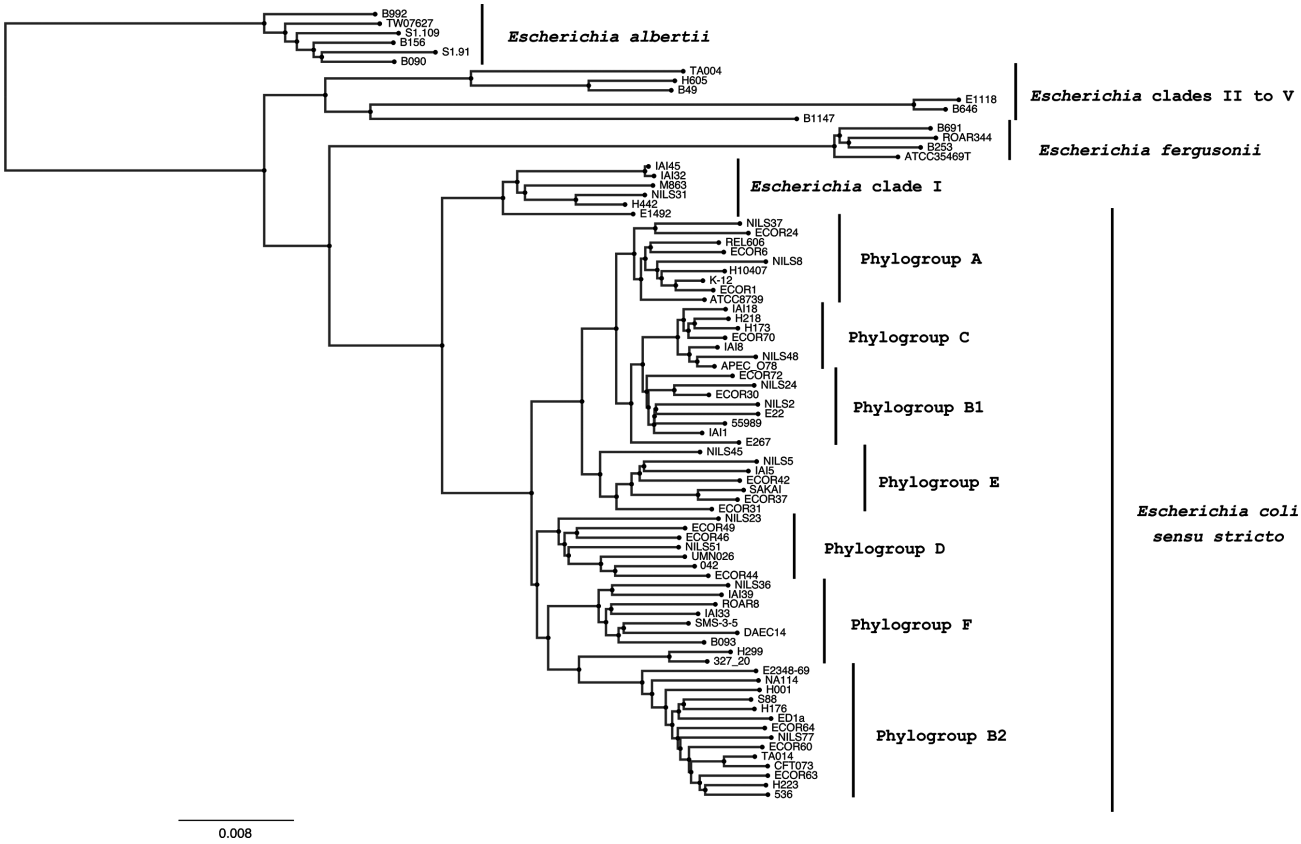

***Figure 1.3: Population structure of* E. coli**, *based on 83 isolates representing the* Escherichia *phylogroup diversity. Taken from [95].*

Shigella

Shigella, a closely related species, was first described in 1897 as the causative agent of dysentery [96]. In the late 19th century, *Shigella* and *E. coli* were determined to be two separate species based on phenotypic studies; *Shigella* is a non-motile, obligate pathogen whereas *E. coli* is motile and commonly found in the gastrointestinal tract of healthy individuals [97,98]. Moreover, the two organisms presented different metabolic profiles [94]. It has since been confirmed that *Shigella* in fact contains four species (*S. dysenteriae, S. flexneri, S. boydii*, and *S. sonnei*) which are nested within the *E. coli* phylogeny in different locations [99,100]. Hence, some *E. coli* variants are in fact more closely related to *Shigella*, than the different

Shigella species are to each other. Thus, *Shigella* spp., in regards to species definitions, are in fact *E. coli*s which have been given a different name for historical reasons.

1.1.2.2 Pathogenicity and resistance

E. coli is a common coloniser of the human gut in healthy individuals [10]. However, particular variants of *E. coli*, termed "pathotypes" or "pathovars", have specific properties which lead them to cause a range of diseases, both in humans and other animals [6,101,102]. These pathotypes are broadly divided into two categories depending on their site of infection.

Diarrheagenic *E. coli*

Diarrheagenic *E. coli* infect the gastrointestinal tract. Seven diarrheagenic pathotypes have been described: Enteropathogenic *E. coli* (EPEC), Enterotoxigenic *E. coli* (ETEC), Enterohaemorrhagic *E. coli* (EHEC), enteroaggeragive *E. coli* (EAEC), Enteroinvasive *E. coli* (EIEC), diffusely adherent *E. coli* (DAEC) and adherent invasive *E. coli* (AIEC) [6,101,102]. *Shigella* is also a diarrheagenic *E. coli* and is often classified as an EIEC [101,102]. Both EIEC and S*higella* are invasive and have an intra-cellular stage within the host cells [101,102]. The definition of these pathotypes is based on the presence (or absence) of particular known virulence factor genes, phenotypically or otherwise from the disease they cause [6,101,102]. These are detailed in Section 1.1.2.3 of this thesis. EPECs, ETECs and *Shigella* are prevalent in the developing world where they cause fatal diarrhea among infants and children [103,104]. ETECs, EAECs and *Shigella* are the most common causes for travellers' diarrhea [105]. EHECs are the only diarrheagenic *E. coli* that are cause for concern in the developed world as their major reservoir is in the gastrointestinal tracts of cattle [106,107]. Transmission occurs in the community through bovine-contaminated food, contaminated water or person to person transmission [101,106]. EHEC infections cause severe diarrhea and complications of an infection can lead to Haemolytic Uraemic Syndrome (HUS), a life-threatening condition which can lead to kidney failure [101,106]. EHEC serotype O157:H7 is the most common cause of diarrheagenic *E. coli* outbreaks in the developed world [106]. Importantly, while the pathotype definitions are useful, they do not encompass the complete possible range of *E. coli* pathogenicity. An example for this is a new pathotype, a hybrid between EAEC and EHEC typed as O104:H4, which emerged in 2011 and caused a large outbreak of bloody diarrhea and HUS in Germany [108].

Extra-intestinal *E. coli*

Extra-intestinal *E. coli* (ExPECs) infect other bodily sites besides the gastrointestinal tract [6,102]. The most common infections caused by ExPECs are UTIs, BSIs and meningitis [8]. ExPECs have been estimated to cause 80% of community-acquired UTIs [109]. According to

a recent report by Public Health England (PHE), *E. coli* is the leading cause of BSIs and UTIs in the UK, and the number of incidences continues to rise year by year[3] (Figure 1.1). Similar trends were observed in other developed countries worldwide [110–112]. Extra-intestinal infections, and in particular UTIs, are most commonly community acquired, but can also be transmitted in hospitals [113]. Often a BSI follows a UTI, suggesting transmission occurs between the two sites [8,113].

Lineage association with pathogenicity

The first study which expanded the ECOR collection to include diarrheagenic *E. coli* isolates demonstrated that the diarrheagenic pathotypes were spread across the phylogeny and do not cluster according to their pathogenicity [114]. This was the first indication that pathogenicity is most likely horizontally transferred between distantly related strains [75]. In the last decade, using larger WGS studies confirmed that diarrheagenic pathotypes span all the *E. coli* phylogroups [115–119]. ExPECs, on the other hand, are primarily associated with phylogroup B2 which represents 60%-70% of infections, however extra-intestinal infections are also caused from isolates from the other phylogroups [120,121]. Even more, there are four STs that represent the predominant lineages which cause almost half of all extra-intestinal infections in the developed world: ST131, ST73, ST95 and ST69 [120–122].

Antimicrobial resistance in *E. coli*

E. coli was added to the ESKAPE pathogens in 2012 to form the so called ESKAPEE pathogens, confirming its status as a major threat for treatment failure and the spread of AMR genes worldwide [12,13]. This was done due to the high prevalence of *E. coli* isolates in a study of MDR pathogens in an intensive care unit in Mexico, where more than 76% of sampled *E. coli* isolates were MDR [13]. Similar levels of multidrug resistance among *E. coli* isolates in hospital settings, most prominently resistance to ESBLs, have been reported in other studies since [123–127]. Carbapenem resistance among *E. coli* has been reported to be as high as 14% [127–129]. Worryingly, colistin resistance has also emerged in *E. coli*, originating from farm animals and more recently spread widely including to healthy individuals [67,130,131].

Varied resistance profiles have been observed across *E. coli* pathotypes and lineages [102]. In ExPECs, recurring UTIs are common, leading to periodical treatment with antibiotics within the same patient, a process which is likely contributing to the high levels of observed resistance within this pathotype [132]. Of particular concern is the global ExPEC lineage

[3] Annual epidemiological commentary: Gram-negative, MRSA and MSSA bacteraemia and *C. difficile* infection data, up to and including financial year April 2018 to March 2019, ESPAUR report

ST131 which is repeatedly reported as a leading cause of ESBL resistant UTIs and BSIs worldwide [120,121,133]. More recently, ST410 has also been labelled as a high risk ExPEC MDR lineage, resistant to both ESBLs and carbapenems [134,135]. Other STs which have been implicated in the dissemination of resistance genes among ExPECs are ST405, ST69 and ST101 [136]. As EPEC, ETEC, EAEC and EIEC (and *Shigella*) infections are treated with antibiotics, there have been increased incidences of resistance among these pathotypes in recent years [102,137,138]. Antibiotics are not the recommended treatment for EHEC as it causes Shiga toxin-mediated cytotoxicity, nor for DEAC infections which are treated by rehydration, however resistance genes to ampicillin, streptomycin, trimethoprim, sulfonamide and tetracycline are also common in these pathotypes [102,106].

Resistance to ESBLs is higher within hospital settings either due to a strong positive selection within hospitals or due to antibiotic treatment failure in the community[4]. However, community acquired ESBL resistant infections are increasing in prevalence, both for diarrheagenic *E. coli* and ExPECs [139–141] Carbapenem resistance has been reported both among ExPECs and diarrheagenic *E. coli*, predominantly amongst EPECs [102,127,142,143]. Even more, carbapenem resistance genes have been identified in the faeces of healthy individuals in the community, presenting the potential for these genes to transfer between *E. coli* pathotypes that occupy different niches as well as more broadly across other bacteria [127].

1.1.2.3 Genetics

***E. coli* as a model organism for studying the pan-genome**

Since *E. coli* is well characterised and arguably one, if not the most, important bacteria used to investigate bacterial evolution, laboratory strains of *E. coli* were some of the first fully sequenced genomes [144,145]. The availability of multiple whole genome sequences of this organism paved the path for some of the first comparative genomic studies, examining gene content across multiple isolates from different pathotypes. A comparison of three whole genomes of *E. coli* (EHEC, ExPEC and *E. coli* K-12) revealed the mosaic structure of the genome and the high diversity in gene content; only 39.2% of the genes identified were common to all three isolates [146]. Studies that followed using larger collections of up to 20 isolates, confirmed that an *E. coli* genome consists of approximately 5,000 genes, however only 2,000 are shared by all isolates [11,92,147]. The *E. coli* pan-genome is open; sampling additional genomes leads to further identification of novel rare genes [11,147,148] (Figure 1.2). While it was originally estimated to contain 17,000 genes, recent studies analysing over

[4] English Surveillance Programme for Antimicrobial Utilisation and Resistance (ESPAUR) Report, PHE, 2018-2019

10,000 genomes have estimated the size of the *E. coli* pan-genome at over 100,000 genes [92]. Furthermore, early studies were able to confirm that genes belonging to the core genome are most commonly of known functions and annotated to play a role in metabolism [11,147]. Genes in the accessory are mostly mobile elements, prophage remnants or genes that encode for outer-membrane proteins [11,147,149]. The vast majority of these accessory genes are rare and present in fewer than 15% of isolates [11,120,148].

Genetic determinants of virulence

There is a range of factors that confer virulence across the *E. coli* pathotypes, however, the strategies used for infection are broadly shared across the pathotypes. Often, adhesion to host cells is required, followed by undermining of the host cells' cellular processes by secreting proteins into the host cell [6,101,102]. The main virulence factors that are used to identify each of the pathotypes are detailed below. Additional virulence factors which have been found to be associated with the pathotypes have also been described, but are not detailed in this thesis [6,101,102].

EPEC: EPECs are defined by the presence of a pathogenicity island named "locus of enterocyte effacement" (LEE) [102,149,150]. The LEE encodes for an outer-membrane protein called intimin which enables adhesion (*eae* gene), a type III secretion system and effector proteins, one of which is called Tir [102,149]. The Tir effectors are injected into the host cell via the secretion system where they function as receptors for the intimin protein, allowing for stronger attachment to the epithelial cells [102,149,151]. EspB, a component of the type III secretion system, and intimin have been shown to be essential for EPEC virulence and therefore are most commonly used for EPEC identification [149,150,152]. An additional virulence factor, bundle-forming pilus (*bfp*), enables adhesion and has also been described as essential for virulence [153]. However, infections by EPECs that do not have *bfp* have been reported and are in fact common in both industrialised and non-industrialised countries [102,115,119,154]. These isolates have been termed atypical EPECs (aEPECs) [102,149,154].

EHEC: EHECs are defined by the presence of two virulence determinants: the LEE pathogenicity island and Shiga-toxin (Stx) (also known as verocytotoxin) [102,149]. Stx cleaves ribosomal RNA, thus killing the infected cell [101]. Stx is secreted in the colon and travels to the kidney where it causes damage that can lead to HUS. Stx also causes damage in the colon and thus leads to bloody diarrhea [102,155]. Isolates which are Stx-positive but do not possess the LEE pathogenicity island are generally termed Shiga-toxin producing *E. coli* (STEC) or verotoxigenic *E. coli* (VTEC) [102,149]. Atypical EHECs which produce Stx

and adhere to the epithelial cells via other adhesins have been also identified [149,156]. These include *E. coli* O104:H4 which produces Stx and has the adhesions of an EAEC [74].

ETEC: ETECs are defined by the ability to produce enterotoxins: heat-labile enterotoxin (LT) and heat-stable enterotoxin (ST) [102]. Additionally, ETECs possess a range of colonisation factors that enable them to attach to the epithelial cells [102,117,149]. The enterotoxins impact the electrolyte transport in the small intestine, leading to watery diarrhoea [157].

EAEC: EAEC were defined based on their phenotype when grown *in vitro* on tissue culture cells, as they create a "stacked brick" adherence pattern [102,158]. This phenotype has mostly been attributed to several different aggregative fimbriae, termed aggregative adherence fimbriae (AAFs), which are found on the pAA plasmid [149]. Identification of EAEC today is mostly determined genetically based on the presence of the *aatA* and *aaiC* genes, encoding for virulence protein transporter found on the pAA plasmid and a chromosomally encoded gene which has been associated with EAEC virulence [149,159]. However, the presence of these genes does not necessarily confer the adherence phenotype, thus there is currently no clear genetic definition of this pathotype [102,149].

DAEC: Similar to EAEC, DAEC were defined based on their adherence pattern *in vitro* on tissue cell culture [160]. The main genetic determinant identified for the adherence pattern are adhesins Afa/Dr and these are termed "typical DAEC" [102,161]. Atypical DAEC have the same adherence phenotype via the presence of other adhesins or otherwise, they possess Afa/Dr along with other virulence factors such as the LEE pathogenicity island or enterotoxins [102,149]. In those cases, the classification of an isolate would most likely be as an EPEC or an ETEC [149]. Thus, the classification of DAEC remains elusive [102].

EIEC/*Shigella*: The key virulence determinant in EIEC is the pINV plasmid, which is also found in *Shigella* and encodes for the proteins enabling the intracellular lifestyle [162]. Indeed, EIEC are phylogenetically and pathogenetically similar to *Shigella* [163]. The plasmid encodes for a type III secretion system which enables host-cell penetration, genes for movement within the cell, the invasion of neighbouring cells and evasion of the host immune system [162].

ExPEC: Unlike diarrheagenic *E. coli*, ExPECs are defined by their site of infection [102]. Therefore, there is no set of genes which is necessary or sufficient to cause extra-intestinal infections [164]. Rather, it is considered that most ExPEC infections are opportunistic, causing infections from the commensal *E. coli* population [149]. With that being said, there are particular factors that have been found to be enriched in ExPEC infections, suggesting that

there are factors that contribute to the ability of ExPECs to colonise other tissues [165]. These include type 1 fimbriae, pili, AfA/Dr adhesins, K1 polysaccharide capsule, toxins and genes involved in iron acquisition [102,149,165,166].

Unclear boundaries between the *E. coli* pathotypes

The pathotype definitions are highly valuable clinically in order to understand how to treat an infection and for epidemiological purposes for surveillance and measuring burden of disease of the different groups [149]. Therefore, identification of the mentioned virulence factors in clinical laboratories is useful. However, the boundaries between the pathotypes are not absolute. Virulence factors are often horizontally transferred, therefore there is no limitation for one of the pathotypes to acquire an additional virulence factor leading to “hybrid” pathotypes [102,149]. The most recent example of this is the O104:H4 *E. coli* outbreak in Germany in 2011 [108,167]. An EAEC had acquired Stx and thus became a Shiga-toxin producing enteroaggregative *E. coli* [167]. This pathogenic variant caused over 3,000 cases in healthy adults and led to 53 deaths [108]. The complexity of the pathotype definitions is depicted in Figure 1.4. Furthermore, ExPECs are defined by the site of infection. Thus, a strain that possesses the above-mentioned virulence factors which was not isolated from the gastrointestinal tract would be defined as ExPEC.

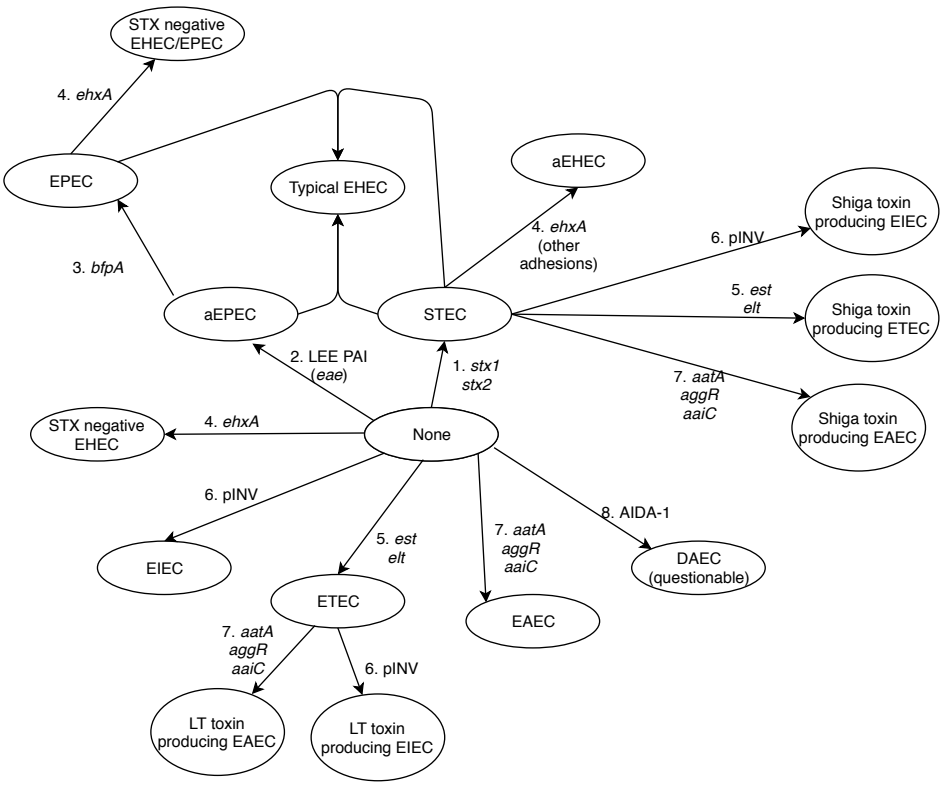

Figure 1.4: Decision network of the virulence factors defining the* E. coli *pathotypes. *Each node is an* E. coli *pathotype or combination of pathotypes. The edges define the presence of marker virulence factors. The complexity of the network expresses the ambiguity of the pathotype definitions.*

Genetic determinants of resistance

Resistance in *E. coli*, similar to *K. pneumoniae*, is predominantly driven by gene acquisition through HGT [168]. Chromosomal alterations that lead to resistance are similar to those in *K. pneumoniae* and include resistance to quinolones due to mutations in *gyrA* and *parC* genes [169]. Similar to *K. pneumoniae*, chromosomally encoded colistin resistance is mostly driven by mutations in the PhoQ/PhoP system [170]. Finally, chromosomally encoded ESBL resistance is driven by the overproduction of the AmpC beta-lactamase [136,171]. Otherwise, plasmid encoded resistance to ESBLs is most commonly conferred by the beta-lactamase CTX-M family of genes [136,168]. The most widespread CTX variants are CTX-M-14 and CTX-M-15 [136,168,172]. CTX-M-15 is most known for leading to ESBL resistance in the global MDR clone ST131 [173,174]. Plasmid mediated AmpC resistance is also common in *E. coli* [171]. Resistance to carbapenems in *E. coli* is mostly commonly conferred by the New Delhi metallo-β-lactamase (NDM-1) which is plasmid mediated [168,175]. Plasmid mediated colistin resistance in *E. coli* is encoded *mcr* genes, with five *mcr* genes having been identified in *E. coli* named *mcr* 1 to 5 [176–180].

1.2 The phenomena: Horizontal gene transfer

One of the main contributors to the pathogenicity, resistance and the large pan-genomes of *K. pneumoniae* and *E. coli* is HGT. HGT is the process by which new genetic material is introduced into a genome from an outside source, whether within the same species or between species [181]. Unlike the intrinsic accumulation of genetic mutations, gene acquisition is a rapid evolutionary process which enables immediate adaptation and propagation of genes across a whole population, including the spread of genes conferring resistance to antibiotics and virulence. HGT occurs at high rates in species that have large pan-genomes, such as *K. pneumoniae* and *E. coli* [182]. The increased pan-genome size and expansion of gene families within these species is driven primarily by HGT [182]. The outcome of gene flow between populations is most prominent when a phenotype is under strong selection in a specific environment, and thus genes disseminate quickly such has been the case for resistance genes in *K. pneumoniae* and *E. coli* [183,184].

1.2.1 Mechanisms of HGT

1.2.1.1 Inter-cellular mobility

Conjugation is the process by which genetic material is transferred between two cells through direct physical contact (Figure 1.5) [181,185]. A conjugation pilus is formed through which the genetic material is transferred [181]. Most commonly, plasmids transfer between cells via

conjugation [185]. In order for conjugation to occur, the presence of genes which encode for the conjugation machinery is needed. These primarily include the components of a type IV secretion system which forms the pilus through which the DNA is transferred, a relaxosome complex for processing the DNA and a type IV coupling protein which connects the relaxosome to the transport channel [186–188]. Plasmids are considered "conjugative" when the conjugation machinery is encoded on the plasmid itself. Plasmids are considered non-conjugative if they can exploit the conjugation machinery of other plasmids in the same cell to transfer between cells, and are these are termed "mobilisable" [185,187,189].

Transduction is the mechanism by which a phage acts as a vector for genetic transfer by incorporating genetic material from an infected bacterium during lysis or excision, and carries the material to another cell (Figure 1.5) [181]. Transduction can either be generalised or specialised. In generalised transduction, a random piece of the host DNA is incorporated into the phage DNA, whereas in specialised, the incorporation of host DNA in the phage DNA is driven by imprecise replication [181].

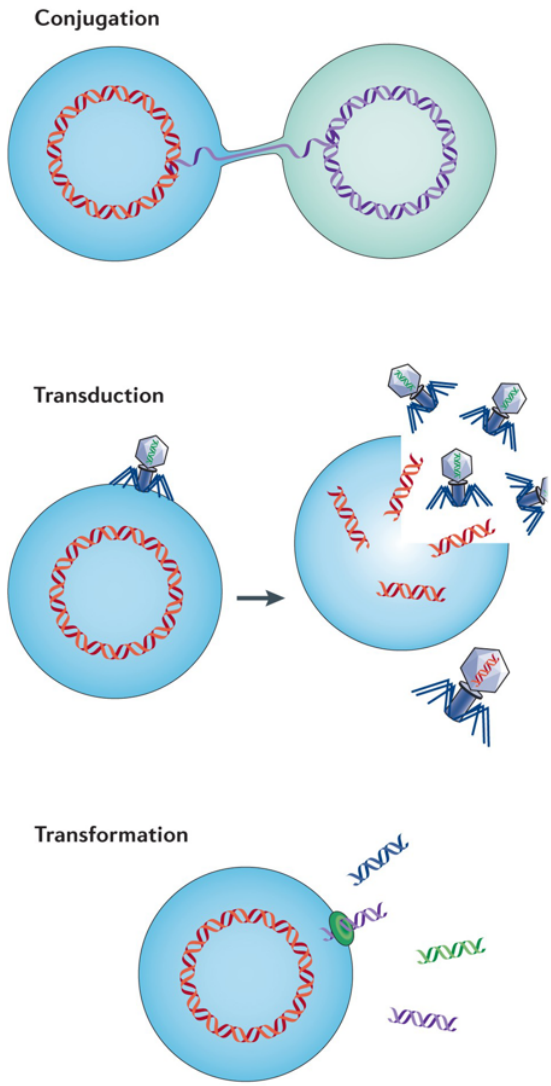

Figure 1.5: Main mechanisms of HGT. *Adapted from [181]. Conjugation occurs by direct physical contact between two cells. In transduction, phage act as vectors of gene transfer. Transformation is the uptake and integration of DNA from the environment.*

Transformation is the uptake and integration of DNA from the environment under natural growth conditions (Figure 1.5) [185]. This ability is driven by the presence of a set of genes which enable DNA uptake, termed “competence” [190]. Competence is often regulated based on the environmental conditions. Transformation has been observed across a range of bacteria and archaea, suggesting it is an important mechanism for introducing genetic variation [185,190].

1.2.1.2 Intra-cellular mobility

In addition to inter-cellular mobility of genetic material via the above-mentioned mechanisms, intra-cellular mobility by recombination and mobile genetic elements (MGEs) is a major contributor to the spread of genes in the population [189,191].

Recombination of genes, or fragments of genes, is the process by which genetic material is integrated into the chromosome [192]. This process is broadly divided into homologous recombination, which requires sequence homology between the two DNA segments involved, and non-homologous recombination [191]. Recombination is important in the integration of foreign DNA following inter-cellular transfer. For instance, following transformation, DNA molecules which share homology to the host genome can be integrated in the genome via homologous recombination. Homologous recombination is mostly driven by the presence of the RecA family of proteins [193]. Non-homologous recombination is driven by specialized enzymes encoded on particular MGEs and phages [191]. While recombination is widespread across bacteria, the level of recombination across species has been estimated to vary widely [192,194].

A range of MGEs have been described. Transposon and insertion sequences (IS) are DNA segments that encode all the enzymes required for their self transposition to different sites in the genome [189,195]. Integrons use site-specific recombination and encode for an integrase which enables the recombination of the integron gene cassette into these specific sites [196]. Genomic islands are defined as distinct regions of the chromosome which show the characteristic of having been horizontally transferred and are usually flanked by direct repeats [189]. Integrative conjugative elements (ICEs) are an example of a genomic island [183]. ICEs encode for their self conjugation, but unlike plasmids, they are integrated in the chromosome [183]. As they are chromosomal, they are vertically inherited, however, they can excise to form a circular DNA molecule, conjugate into another cell, and reintegrate in specific sites. The conjugative machinery of ICEs can also be exploited by non-conjugative plasmids. Other clinically important genomic islands include resistance islands and pathogenicity islands [189].

The presence of MGEs on a plasmid enables their transfer onto the chromosome without the need for the plasmid to replicate and persist, as the MGE can transfer to the chromosome even if the plasmid does not survive [189]. As coined in [189], the plasmids can act as “suicide vectors” for the horizontal spread of genes via MGEs. Additionally, their presence enables gene duplication within a genome. The duplication of the MGEs in genomes and plasmids leads to many homologous regions which can recombine leading to an even further increase in genetic diversity [189]. Importantly, this leads to many repeat regions within organisms which have a high load of MGEs in their genome. Highly repeated DNA makes the assembly of short read whole genome sequences challenging, as it is impossible to determine the order of fragments that are flanked by the same repeats [197]. This leads to highly fragmented assemblies with many short contigs.

1.2.2 Barriers to HGT

In order to investigate the flow of genes and the pan-genomes of *K. pneumoniae* and *E. coli*, it is essential to understand the barriers for transfer of genetic material between isolates and species.

1.2.2.1 Genetic barriers

There are genetic barriers that can limit the ability of genetic elements to transfer between cells. For instance, in order for homologous recombination to occur, there has to be a certain level of sequence similarity between the donor DNA and the host, or otherwise, particular DNA integration sites need to be present to be recognised by the integrase [185,198,199]. For transformation to occur, genes which enable competence must be present and active [190].

Plasmids depend on host replication systems for their own replication. Following conjugation, if the replication machinery (helicase, primase, polymerase) of the recipient cell is incompatible with the plasmid’s replication mechanism, the plasmid will not survive [189]. Plasmids which require the same replication factors from the host are termed “incompatible” as they compete for the same replication factors in a cell (including for copy-number control), and therefore cannot co-exist [189]. Plasmids have been typed into compatibility groups (Inc) based on their replicon [200,201]. Plasmids belonging to the same compatibility groups may also undertake surface exclusion to prevent a plasmid which uses the same replication factors to conjugate into the cell [185].

Transduction may not be possible between two bacteria which are susceptible to different phage. For instance, in order for a phage infection to occur, specific surface receptors must

be expressed on the outer membrane to enable the infection [202]. Otherwise, immunity to phage infection is given by genes which recognise phage associated motifs in the phage genome and prevent its replication [203]. These immunity genes may be encoded on other phage which are present in the host [204]. Other genetic immunological factors include restriction modification systems or CRISPR-cas systems which digest foreign DNA molecules [185,202,204,205]. Finally, abortive infections are the equivalent of programmed cell death, where a bacterium which has been infected by phage induces its own death to prevent its further propagation in the population [204].

1.2.2.2 Physical barriers

In addition to genetic barriers, some physical barriers exist which can prevent the transfer of DNA. One such barrier is the occupation of different niches [206]. Conjugation requires direct physical contact between two cells and transduction can only occur if two cells are in a shared ecological environment. For transformation to occur across ecological niches, the DNA must persist and be stable in the environment. Finally, studies have shown that the capsule could prevent the transfer of genetic material between cells, likely due to the physical barrier it creates [207]. However, a recent genomic analysis on a range of species showed a conflicting picture where species which encode for more capsules present higher levels of HGT [208].

1.2.3 HGT in *K. pneumoniae* and *E. coli*

K. pneumoniae and _E. coli_ cross niches, making them major gene-traffickers

In order for HGT to occur, the organisms taking part in the transfer of genetic material need to reside in the same habitats [4,206]. Organisms that can transfer through different niches, for instance, from a human to the environment and back to a human, would thus be major players in trafficking genes. Both *K. pneumoniae* and *E. coli* are ubiquitous across niches; they are colonisers of human guts as well as animal gastrointestinal tracts, found in the environment, in plants and in soil [4,209]. The level of movement of these two organisms between niches is still unclear, but genomic studies have shown that the same lineages exist across multiple niches, suggesting there is no clear separation between niches [4].

Genome plasticity in _K. pneumoniae_ and _E. coli_ is high

Genome plasticity is another prerequisite for a large pan-genome. An ability to lose and acquire genes enables the persistence of low frequency genes in the population [210]. Gene gain is most commonly driven by HGT [211].

High recombination rates in *E. coli* have been well established, and are estimated to be twice as likely to occur than mutation on any base pair in the genome, meaning that recombination is a major driving force in *E. coli* evolution [11,77]. Studies on recombination within and between the *E. coli* phylogroups have found that inter-phylogroup (and even inter-lineage) recombination is low whereas intra-phylogroup recombination is high, suggesting there are biological and ecological barriers to gene sharing between distant clades [212,213].

In *K. pneumoniae*, studies on recombination rates have been conflicting [25]. Recombination in *K. pneumoniae* has been estimated to be low [194]. On the other hand, an example of the contribution of recombination to genome plasticity and resistance in *K. pneumoniae* includes the emergence of the carbapenem resistance lineage ST258 which was driven by a large recombination event [214]. A more recent study found that large recombination events occur frequently in *K. pneumoniae* and is driving the emergence of novel STs [215].

High rates of gene gain and loss over time in *E. coli* are also indicated by the fact that very closely related isolates tend to share many genes, whereas distantly related isolates present a wide distribution in the number of genes they share [11]. The high number of genes shared between closely related isolates indicates recent gene acquisition on short timescales. On longer timescales, genes are lost and regained, therefore gene sharing varies and does not follow the phylogeny [11].

K. pneumoniae has been associated with hundreds of distinct plasmids spanning many Inc types and has been shown to possess, in most cases, between 2-5 plasmids per isolate [4]. On the other hand, *E. coli* has been shown to carry fewer plasmids, between 0-3 per isolate on average. The number of plasmids in each isolate can vary significantly, with some *E. coli* and *K. pneumoniae* isolates predicted to contain as many as 10 plasmids within a single isolate [4]. This large plasmid load suggests that *K. pneumoniae* and *E. coli* are particularly permissive to plasmids, and that these plasmids are able to persist for enough time to propagate in the population [4].

Another indication of high genome plasticity is the heterogeneity in genome sizes in both *E. coli* and *K. pneumoniae* as isolates vary by more than 1 Mb in genome size [4,216,217].

Interestingly, plasmid load, recombination and gene gain and loss rates differ between *E. coli* and *K. pneumoniae* lineages (and can be consistent within a lineage) [4,77,92,212]. This suggests different dynamics of gene acquisition and loss across the species.

1.2.4 Contribution of HGT to virulence and resistance

Resistance genes in *K. pneumoniae* and *E. coli* are often mobilised on plasmids

Resistance to beta-lactams, ESBLs, carbapenems and colistin are all present on plasmids [168]. As resistance is often encoded on plasmids, these often encode for additional genes that confer resistance to other antibiotics, disinfectants and heavy metals [5,201]. Hence, the gain of the resistance phenotype can generally improve fitness due to association with other genes on the same plasmid, and the selective pressure led by the presence of a single antibiotic leads to dissemination of multidrug resistance [25,136]. With that being said, a large resistance plasmid means greater DNA burden and inability to acquire other plasmids of the same incompatibility group. Therefore, there is a cost benefit in acquiring these resistance plasmids, probably explaining why resistance plasmids are not ubiquitous [189].

One major example is the ESBL resistance variant CTX-M-15 in the MDR *E. coli* lineage ST131, which is commonly found on IncFII plasmids and carries additional resistance genes [136,189,201]. Other examples are ST258 and ST11 of *K. pneumoniae* (of the same clonal group), which are commonly associated with carbapenem resistance and 50%-75% of the plasmids within these lineages harbour additional resistance genes [4,5,25]. Studies on ST258 and ST11 have confirmed that each isolate typically harbours 2-6 plasmids, thus isolates belonging to these lineages have been shown to contain between 12-15 distinct resistance genes [25,218].

Plasmid mediated colistin resistance by the *mcr* genes has been associated with a range of plasmids. The *mcr-1* gene has been found on multiple plasmid backbones including IncF, H, X and I-complex plasmids, *mcr-3* on IncHI plasmids, *mcr-2* on IncX plasmids and *mcr-4* and *mcr-5* were reported to be present non-conjugative ColE plasmids and are transmissible via transposon mediated transposition or through mobilisation on helper plasmids [176–178,189].

Hypervirulence genes in *K. pneumoniae* are horizontally transferred

A large plasmid, pLVPK, has been named as the main contributor to hypervirulence in *K. pneumoniae* [31,45]. The plasmid contains many of the virulence factors known to be associated with hypervirulence, including genes encoding for siderophores and RmpA, the regulator of capsule production [45]. Additionally, it has been shown that ST23 isolates, from the hypervirulent lineage, which do not have this plasmid show reduced virulence potential [219]. Other virulence genes have been associated with the presence of an ICE (ICE*Kp1*) or otherwise present in genomic islands [31,220]. WGS comparisons identified additional regions which were associated with integrases and had a low GC content that are unique to the

hypervirulent strains, further emphasising the contribution of HGT to the hypervirulent phenotype [31].

Acquisition of virulence genes defines the *E. coli* pathotypes

The Stx, which is present in STEC and EHEC isolates, is acquired by transduction on a lambdoid bacteriophage [221–223]. The phage is capable of both lysogenic and lytic growth and production of infectious particles, thus Stx can be maintained stably within the bacterial host chromosome as well as easily be transmitted to other cells by transduction during lytic growth.

Many of the marker virulence genes are found on plasmids and transmitted via conjugation. Most prominently these include the LT and ST genes in ETECs which are almost exclusively found on plasmids, the pINV in EIEC and *Shigella* which contains all the genes required for invasion and intracellular survival and the pAA plasmid in EAEC [224,225]. Typical EPEC isolates possess the pEAF plasmid which contains the genes for forming the bundle-forming pilus [102]. In addition to the virulence factors presented in Section 1.1.2.3, many other factors have been described to contribute to pathogenicity, such as fimbriae and toxins, and these are found on large virulence plasmids and genomic islands [6,102]. Most prominent are plasmids pO157 in EHECs, pB171 in EPECs and PcOO in ETECs [102,226]. Pathogenicity islands, which are an indication of an HGT event, are also common among pathogenic *E. coli*. One main example is the LEE pathogenicity island found in EPECs and in typical EHECs [102]. Pathogenicity islands have also been described in UPECs, EAECs, ETECs and EIECs [6].

Contribution of recombination to pathogenicity in *E. coli*

There have been conflicting results regarding the contribution of recombination in the core genome to the pathogenicity in *E. coli*. Early studies measuring recombination in MLST genes found that recombination rates among pathogenic variants of *E. coli* are 5-6 times higher than recombination rates in commensal non-pathogenic *E. coli [77]*. Following studies using whole genome sequences found that in fact recombination rates among the most virulent lineages, ExPECs including ST131 and EHECs, were lower relative to commensal *E. coli* [212,213]. The earlier studies suggested that pathogenic *E. coli* needed to adapt to changing environments, and thus variants that were able to adapt quicker were selected amongst pathogens [77]. More recent studies suggest that pathogenic *E. coli* are highly adapted and sexually separated from the rest of the population, thus undergo less recombination [213].

1.3 The genes: Toxin antitoxin systems

Virulence and resistance plasmids in *K. pneumoniae* and *E. coli* are often very large, and therefore are a metabolic burden on their host as their replication requires energy and metabolic sources. Hence, with no selection pressure, there are mechanisms in place for their maintenance [189]. Toxin-antitoxin (TA) systems likely play a role in *K. pneumoniae* and *E. coli* pathogenicity and resistance by the maintenance of these plasmids.

TA systems are bicistronic operons composed of a toxin gene and antitoxin gene [227]. The toxin inhibits cellular processes, such as translation or transcription, and thus leads to growth arrest or cell death. The antitoxin is co-transcribed with the toxin and inactivates the toxin's activity by different modes of inhibition. Typically, the antitoxin is less stable than the toxin as it is targeted to be degraded by cellular proteases such as Lon or Clp [228]. Thus, when a cell does not inherit the TA system post division, the existing antitoxin will degrade prior to the toxin and the cell will cease to grow or die [229]. This process was termed "post segregational killing" (PSK) [230]. Therefore, in the absence of selection, these plasmids are thought to be maintained by addiction to the TA system [226]. The existence of the phenomena of PSK has been under debate recently, as there are concerns that over-production of any protein would lead to cell death and that the expression levels used *in-vitro* are biologically irrelevant [231]. Nonetheless, there is clear evidence that TA systems play a role in the maintenance of plasmids, simply not necessarily by PSK but by growth inhibition [231].

1.3.1 Classification

Seven types of TA systems have been described based on the properties of the antitoxin and its mode of inhibition of the toxin (Figure 1.6). Type I system antitoxins are antisense RNAs which inhibit the toxin's activity by binding to the toxin's mRNA [232]. Antitoxins of type II systems are proteins which bind directly to the toxin, thus inhibiting its activity [227]. The antitoxins of type III systems are RNAs that form a pseudoknot that binds directly to the toxin [233]. Antitoxins of type IV systems are proteins that interact directly with the toxin's target [234]. Antitoxins of type V systems are RNAse that cleave the toxin mRNA [235]. Antitoxins of type VI systems are proteins that promote the degradation of the toxin by a protease [236]. Type VII TA systems have only more recently been described, where the antitoxin is a protein that leads to the modification of a cysteine residue on the toxin and inactivates it [237].

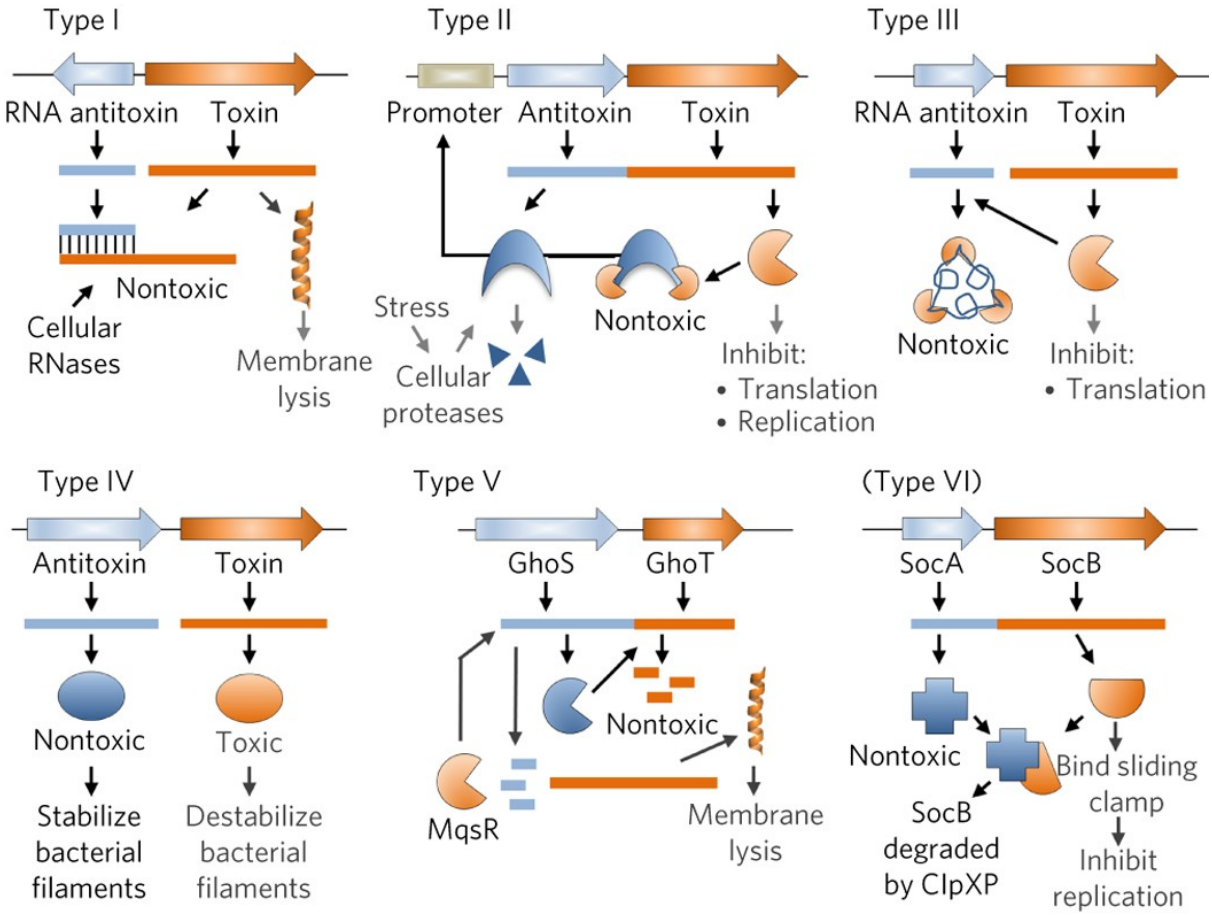

Figure 1.6: Types of TA systems. *Taken from [238]. Toxins are shown in orange and antitoxins in blue. The newly described Type VII system is not shown.*

1.3.2 Mechanisms

Toxins have been shown to have a variety of cellular targets. The toxins of type I systems, which are most commonly short proteins of only 60 aa, have a specific structure which localises in the inner membrane leading to membrane damage [232,239].

The proteins of type II systems are the most varied and have a range of cellular processes that they target [232]. CcdB and ParE toxins target DNA gyrase [240,241]. The Zeta toxin of a three component TA system ω-ε-ζ interrupts peptidoglycan synthesis [242]. VapC inhibits transcription by site specific cleavage of tRNAs [243]. HipA phosphorylates and inactivates GltX, the protein in charge of charging glutamate onto tRNAs [244]. Doc targets the ribosome leading to inhibition of protein synthesis [245]. RelE and MazF are RNAses that inhibit translation by cleavage of mRNA [246,247].

There are only a few representatives of the less well studied TA systems, types III-VII.The type III toxin, toxN, is RNAse and inhibits translation [248]. The type IV toxin, CbtA, inhibits

the polymerisation of cytoskeletal proteins MreB and FtsZ [249]. The type V toxin, GhoT, is a hydrophobic peptide similar to that of type I systems that disrupts the cell membrane [235]. The VI toxin, SocA, blocks replication elongation [236]. Finally, the target of the most recently discovered type VII toxin, Hha, is still unknown [237].

1.3.3 Role in resistance and pathogenicity

Maintenance of virulence and AMR genes

The most common type of TA systems thought to maintain virulence and resistance plasmids are type II systems, with some plasmids containing more than one TA system [226]. The VapBC system is found on the EIEC/*Shigella* pINV plasmid and has been shown to be essential for the maintenance of the plasmid in *Shigella* [250]. The CcdAB and RelBE TA systems are found on *E. coli* virulence plasmids pO157 and pB171 (described in Section 1.2.4.2) [226]. Computational analyses on TA systems found that beyond plasmids, TA systems are often found with other MGEs such as ICEs and pathogenicity islands, suggesting they play a role in maintenance of other MGEs [227,251].

Persistence

Bacterial cells enter a state of dormancy termed “persistence” in response to stress, for instance, in the presence of antibiotics or lack of nutrients [238,252]. Unlike resistance, persistence is a state which a cell can enter and exit, rather than a change in phenotype conferred by mutation or gene acquisition [238,252]. During this state, antibiotics are less effective as the bacterium is not growing. Thus, this phenomenon is thought to be the cause of failure of antibiotic treatment and recurrent infections, as once the antibiotic is removed the persisters can revert to a growing state and repopulate the population [238,252]. It has been established that TA systems, particularly type II TA systems, contribute to this phenotype. For instance, knock out of the *hipA* TA systems leads to reduced persistence, and expression of TA systems has been observed in persisters in macrophages [244,253]. The exit and entrance from the persister state is thought to be achieved by type II TA systems through a process called “conditional cooperativity” [254]. The toxin and antitoxin form a complex that regulates the expression of the TA operon. Depending on the stoichiometry of toxin and antitoxin, the complex of toxin and antitoxin either induces further expression of the antitoxin or represses it leading to growth arrest and persistence.

Modification of regulation

TA systems can affect gene regulation, and thus increase virulence. This may occur via a transcriptional read-through: the TA is inserted in the promoter region of another gene leading

to changes in the gene's expression [255]. Alternatively, antitoxins of type II systems often serve as transcriptional regulators, thus they may affect the regulation of other genes [227]. Finally, toxins of type II systems are often RNAses which may lead to selective degradation of mRNAs and thus to changes in gene expression [227].

Virulence caused by toxins

Another potential role of TA systems in virulence is that the toxins of the TA system can themselves be toxic to the host and thus increase virulence [227]. For instance, in the case of intracellular pathogens, toxins of type I systems may interfere with the host cell membrane, or otherwise toxins of type II systems that are mRNAses may degrade host mRNA.

TA systems as barriers to transduction

TA systems have been shown to play a role in inhibition of phage propagation. A type II system MazF/MazE, type III system ToxN/ToxI and type IV system AbiEii/AbiEi were all shown to reduce propagation of phage [256–258]. This is likely achieved by arrested growth which prevents further propagation of the phage [231]. As transduction with phage is one of the major mechanisms of HGT, the presence of different TA systems may act as defense systems against phage and may affect transduction rates and possibility of phage to infect different species and transfer genetic material between species, including resistance and virulence genes.

Biofilm formation

Deletion of TA systems in *E. coli*, *V. cholerae* and *P. aeruginosa* have been shown to reduce biofilm formation [259–262]. Biofilms are communities of microorganisms that attach to surfaces. They are extremely important clinically as they form the bacterial communities that contaminate catheters and medical implants [263]. Additionally, biofilms have been shown to be less sensitive to antimicrobial treatment [264].

1.4 The approach: comparative genomics using public databases

In 2001, two whole *E. coli* genomes were available for the first time with the sequencing of *E. coli* Sakai O157:H7, four years after the completion of the sequencing of the first *E. coli* genome K-12, enabling the first comparative genomic study [144,145]. With time and the development of sequencing technologies, the number of genomes available has been growing exponentially and thus the number of genomes used in these comparisons has increased

(Figure 1.7). A little over ten years ago, Touchon et al. investigated 20 complete *E. coli* genomes with the quote “We have thus taken advantage of the unprecedented availability of 20 completely sequenced genomes of the same species to analyse the evolution of the gene repertoire” [11]. Today, the largest published study on *E. coli* comparative genomics included 4,071 *E. coli* isolates from a single lineage ST131 [265]. The EcoRPlus collection, mentioned in Section 1.1.2.1, includes 9,479 *Escherichia* genomes [93]. As of today, EnteroBase, a database which performs daily scans of the Sequence Read Archive (SRA) to curate a collection of genomes of enteric pathogens, has over 130,000 *E. coli* and *Shigella* assemblies, and a recent preprinted study has analysed the *E. coli* population on that scale [92,93].

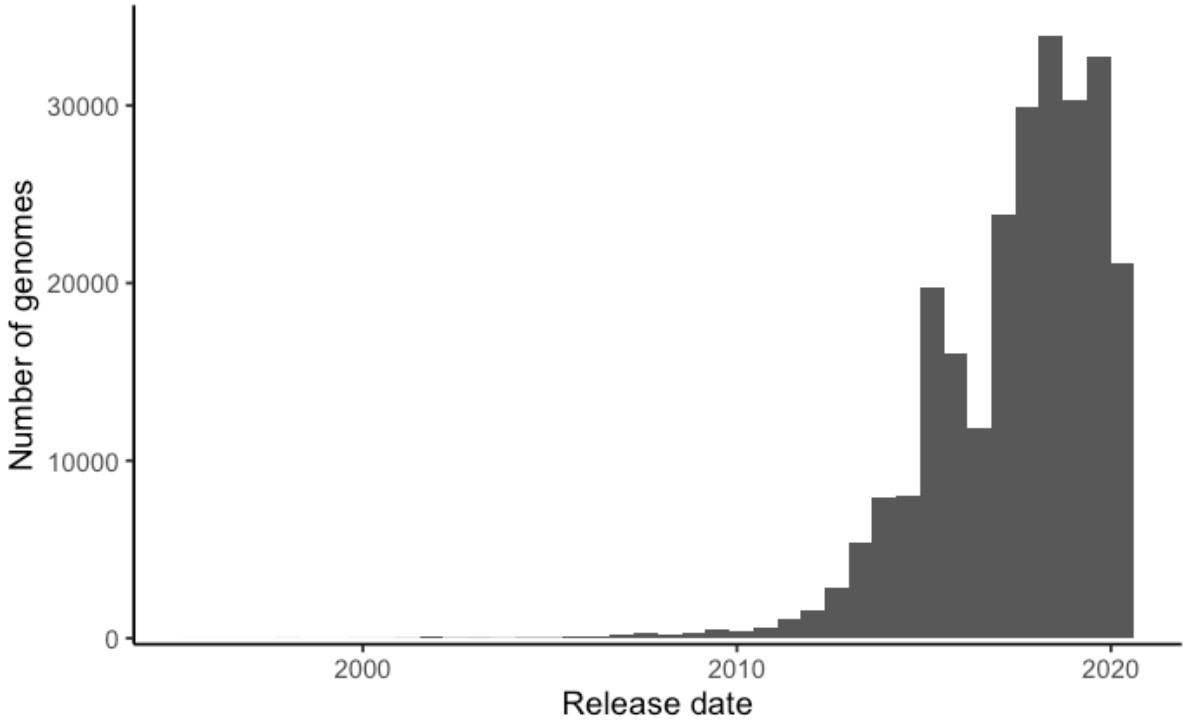

Figure 1.7: Number of bacterial and archaeal genomes released each year on NCBI. *Taken from the genbank_prokaryotes.txt file, downloaded on May 27th, 2020.*

The increasing number of genomes and technological advances enabled to address new questions regarding the evolution and lifestyle of these bacteria. First and foremost, we have been able to investigate the population structure and level of diversity in gene content between members of the same species [9,117,121]. Comparisons of pathogenic and non-pathogenic isolates have helped to identify novel virulence factors or to understand which mechanisms contribute to a pathogenic lifestyle, for instance, recombination rates, gene loss etc. [9,147]. On an epidemiological and clinical level, comparative genomes can unravel how bacterial populations spread geographically and to investigate outbreaks [266,267]. Publicly available genomes can be added to such analyses to put local investigations in a global context.

1.4.1 Methods for comparative genomics

The general approach for comparing genomes requires an initial subdivision of the population based on relatedness of the genomes, along with understanding the evolutionary relationships between these groups. It is important to subdivide the population in order to find biases in the data, and to be able to infer relatedness of isolates across studies. For comparing gene content, tools must first identify the genes of interest and be able to group the genes by homology. The diversity in gene presence or absence and variation within the genes that are present can then be compared across the population. The availability of easy-to-use scalable tools provides the foundation to conduct large scale comparative genomic analyses.

1.4.1.1 Defining the population structure

The genomes in a collection can be divided into groups by application of *in-silico* methods for typing of bacteria based on their genetic variation. The most basic *in-silico* approaches use similar marker genes to those used in laboratory settings to type bacteria. For instance, the 16S rRNA can be used for taxonomic identification [268]. *In-silico* serotyping can be used to define the antigen composition of sequenced isolates [269]. In *E. coli*, *in-silico* polymerase chain reaction (PCR) can be used to assign a phylogroup to a genome based on a set of marker genes [270,271]. An increased level is achieved by using the set of genes used in the MLST schemes, which uses the genetic variations of 7-10 housekeeping genes [272]. Finally, the highest resolution is afforded by grouping isolates based on the gene content across the whole genome. Whole genome multilocus sequencing typing (wgMLST) and core genome multilocus sequence typing (cgMLST) are new approaches which cover the full range of nucleotide differences in all core ORFs of an organism, expanding the original MLST definitions which include only a few genes [273–275]. cgMLST and wgMLST schemes transform variations between genes into simple character data, which reduces the computation power for comparing whole genomes.

Methods which are independent of gene content and are based on whole genome composition can also be used to define groups in the population. BAPS clustering is a Bayesian approach which stochastically partitions the population based on the molecular data [276]. More recently, methods that are based on the k-mer composition of genomes have been developed as they scale better with the increase in the number of genomes [277–279]. K-mers are extracted from a genome by counting all words of size “k" in the sequence. The similarity in k-mer composition can be measured between every pair of genomes and thus they can be clustered into groups. The distance in k-mers between every two clusters is used to elucidate the relationships between clusters [278].

Construction of a phylogenetic tree is used to define the evolutionary relationship between groups [280]. Tree topology and branch lengths are inferred from a multiple sequence alignment (MSA). The input MSA could include a set of genes, for instance, all the MLST genes or all core genes. Maximum likelihood and bayesian methods explicitly state a statistical model of sequence evolution [280]. They use a heuristic approach to sample different tree topologies and for each topology they calculate the branch lengths and parameters of the evolutionary model that produce the highest likelihood or posterior probabilities of the tree given the MSA [280–282]. These methods can be computationally demanding and they do not scale for very large sample sizes [280]. Given the increasing number of available genomes, approaches which are distance based or based on K-mer compositions can be used to estimate phylogenetic relatedness [278].

1.4.1.2 Methods for gene detection

Identifying a single gene of interest

The most basic form of comparing two genomes, is to identify the presence or absence of a particular gene in the query genomes. This is achieved by constructing a database of the gene(s) of interest, for instance, a database of AMR genes, and searching for those genes in the query genomes. The tools to search for the genes fall into two main categories: those that are based on searching for the genes in an assembly or otherwise those that align the genome reads against a database of genes [283]. From the first category, Basic Local Alignment Search Tool (BLAST) or DIAMOND are alignment tools which align an assembly against a gene database [284,285]. Tools for finding resistance genes (res-finder, Arg-annot), virulence genes (virulence-finder) and plasmid replicons (plasmid-finder) are all based on using BLAST+ with a curated database of genes [286–289]. The disadvantage of these tools is that genes may be missed if they were misassembled. To account for this, tools from the second category map the reads against a database of genes and thus are independant of the quality of the assembly [283,290]. Both these searches highly depend on the quality of the database of genes constructed. For instance, if the gene variant used to build the database is not representative of the genome being queried, the gene might not be found. Additionally, mutations in all positions are treated the same as a simple identity cutoff is used across the entire gene sequence.

An alternative approach is to use statistical representations of nucleotide or protein sequences based on an MSA of the gene being searched. These sorts of statistical representations can capture the diversity of a sequence and allow for variation in the search that are weighted by

their biological significance. These include position-specific scoring matrices and hidden markov models (HMMs) [291].

Defining the complete gene repertoire of a genome

Given a complete genome sequence, all the genes or coding sequences (CDSs) of the genome can be identified using genome annotation tools. The most widely used tool for predicting CDSs in genomes is Prodigal [292]. Prodigal uses a training set of well annotated genomes to learn the properties of a CDS in a query organism. These include the gene length, start codon usage, ribosomal binding site motif usage, GC bias and more [292]. These properties are used to choose the optimal set of genes. Pipelines for the complete annotation of prokaryotic genomes have been developed which use Prodigal for the CDS prediction [293,294]. These pipelines include an additional step to predict the function of the CDSs. They align the CDSs against databases of known sequences using BLAST+ and transfer the function from known homologous sequences [293,295]. Additionally, they use HMMER to scan the CDSs for domain HMMs from curated databases of HMM profiles like TIGRFAM and Pfam [296–298].

1.4.1.3 Grouping homologous sequences

Homologous genes are defined as genes that have shared common ancestry [299]. Grouping the identified genes by homology is a difficult task due to the different evolutionary trajectories that occur in prokaryotic evolution. Homologous sequences may be either orthologs or paralogs. Orthologs are homologs which share function and have evolved through speciation [299]. Paralogs are homologs that are the result of gene duplication events and are likely serving different functions [299]. In organisms with high rates of HGT, such as *K. pneumoniae* and *E. coli*, a large number of gene duplication events can lead to a very large number of homologs. Homology is thus inferred based on sequence identity. It is difficult to determine which identity threshold to use as a threshold for separating two sequences and define them as homologous or not [299]. In most cases this is solved by choosing an arbitrary value of percentage identity for two sequences to be homologs [299]. Following that, different approaches are used to distinguish between orthologs and paralogs. Some tools use a reciprocal best hits approach where two sequences are defined as orthologous only if they are each other's best BLAST hit [300–302]. Other methods use phylogenetic comparisons of a gene tree and a species tree to infer gene duplication and loss events and thus infer paralogy or orthology [303]. Other tools to separate homologous sequences based on their gene neighbourhood analysis (synteny) [304–306]. This approach can lead to over inflation of gene predictions in genomes with high levels of recombination. All of the above are further

complicated by the fact that it is difficult to benchmark any of the tools and predictions as the true grouping of genes cannot be truly known [307].

1.4.1.4 Pan-genome analysis

A pan-genome analysis requires defining groups of orthologous sequences in a collection of genomes to identify which genes are shared or unique across isolates or groups of isolates (Figure 1.2) [308]. Following the prediction of all the CDSs in the collection, the general approach is to infer sequence similarity between every two CDSs using tools like BLAST or DIAMOND and group them into orthologous sequences using the approaches described in section 1.4.1.3 [284,285]. For instance, it is possible to apply tools that were designed for general orthology inference on the collection of all CDSs extracted from the genomes of interest [301,304]. Additionally, there is a wide range of bespoke tools that were specifically designed for pan-genome analyses. The most widely used tool is Roary [305]. Roary uses BLAST+ to infer the sequence similarity between every two CDSs and groups the genes accordingly using markov clustering. Orthologs and paralogs are split based on local gene synteny. Other tools, like PanX and Pantaugral, use a phylogeny-based approach to split the genes into orthologs rather than gene synteny [309,310]. As a pan-genome analysis requires comparing all CDSs against each other, with the increasing size of genome collections being used, both the biological and computational complexities increase. For that reason, new tools are continuously being developed to address this problem.

1.5 Thesis outline

The increasing availability of a large number of complete genomes in public databases provides the opportunity to examine the gene pools of *K. pneumoniae* and *E. coli* on a scale larger than previously possible. In this thesis, novel insights on the distribution of genes and the patterns of gene sharing were investigated. The analyses presented required the development of novel approaches to answer the relevant questions and handle large datasets. The first half of the thesis (Chapters 2-3) is focused on a single class of genetic system, TA systems, primarily in *K. pneumoniae*. The second half of the thesis (Chapters 4-5) builds upon the insights found on the diversity of TA systems in *K. pneumoniae*, to answer similar questions on the distribution of all genes in a much larger collection of *E. coli* genomes. In Chapter 2, a tool to Search for LInked Genes (SLING) in large bacterial datasets is described. SLING is a command line tool that enables to search for gene arrays that are physically linked on the genome and to visualize their diversity across large collections. Examples are given for using SLING on 90 *E. coli* isolates and for two operons: TA systems and RND efflux pumps. In Chapter 3, the diversity and evolution of TA systems in a global collection of *K. pneumoniae*

genomes is thoroughly investigated using SLING. The analysis presents a classification of TA systems based on their distribution patterns in the *K. pneumoniae* population. Additionally, a diverse range of novel antitoxins were found and the fluid association between toxins and their antitoxins is described. Chapter 4 details the process of building a high-quality collection of over 10,000 *E. coli* isolates and defining the genes in the collection. Additionally, the properties of the collection are described including the population structure, all associated metadata as well as the efforts made to reduce the biases in the dataset. Chapter 5 uses the final curated collection of genes and genomes to classify the entire *E. coli* gene pool based on the distribution patterns of the genes in the dataset. Additionally, using the classification scheme, the level of gene sharing was measured between different *E. coli* lineages, exposing lineages which may be important in their contribution to gene flow in the *E. coli* population.

2 SLING: A tool to Search for LINked Genes in bacterial datasets

This chapter is a modified version of the published paper "SLING: a tool to search for linked genes in bacterial datasets" [311]. Alexander Harms, Cinzia Fino, Leopold Parts, Kenn Gerdes, Eva Heinz and Nicholas Robert Thomson contributed to the research of the original publication. All final language is my own.

2.1 Introduction

Operons or functionally linked gene arrays represent the most basic unit of transcriptional organization in prokaryotic genomes [312]. Genes involved in the same process or pathway are encoded in a single block, and transcribed under the same regulation [312]. Identifying homologues for a single gene is a difficult task that has been tackled using many methods, as was described in Section 1.4.1.2. The identification of two genes or more which are physically linked to each other further complicates the search. This is because the structure of operons and gene arrays with similar functions can vary substantially across isolates and species. The order of the genes is often changed, and individual genes may be lost or gained [313–315].

TA systems are an example of a simple two-gene operon and were presented in Section 1.3. Databases have been constructed which enable the search for TA systems using simple homology based search tools such as Blast+ [227,251,316–321]. The most well-curated and accessible database is the TA database, TADB [318,319]. However, a homology-based search does not always verify whether the identified genes represent intact CDSs or whether the toxin and the antitoxin are adjacent, meaning further downstream manipulations are required. Two tools have been published which allow for a direct search of the toxin and the antitoxin: RASTA and TAfinder, the TA search tool provided within TADB [318,322]. However, both of these tools are provided in an online interface which is not scalable when examining these systems on larger scales. Even more, RASTA, which was published over a decade ago, no longer in service. Furthermore, they do not allow the addition of custom sequences or domains in the search [318,323,324]. This limits the search, and the quality and relevance of the annotation is determined by the quality of the database. Users have to rely on updates to obtain the most up to date results.

Many other clinically important gene systems are encoded in operons; all secretion systems [323,325], CRISPR-cas systems [315,326], Resistance Nodulation Division (RND) efflux

pumps [327], and more follow this organization. For these more complicated operon structures, sophisticated methods have been developed for their annotation [318,322–324,328]. These tools are restricted to the specific operon which is being investigated as they rely on previously defined structures and sequences, or require reprogramming for identification of new genetic structures.

With the growing availability of large datasets for the surveillance of important pathogens [9,329,330], there is a need for a single flexible framework to annotate clinically relevant gene arrays across a range of isolates and examine their diversity. While a level of specificity will always be required to define the search of a specific operon, there is room to develop generic methods which could search for a range of operons with only a few input requirements from the user.

2.2 Aims

The aim of this chapter was to develop a tool to search for and group operons in large bacterial datasets. In many operons or gene arrays, there is a single conserved gene which is always present together with its neighbours in a rule-defined proximity and orientation. This property provides the potential to capture the diversity of the gene array based on the diversity of the single conserved gene and its neighbours. The precise aims of this chapter were:

- Define the SLING pipeline, a tool to Search for LINked Genes
- Construct the required settings to search for TA systems, and apply these on a collection of *E. coli* isolates.
- Construct the required settings to search for RND Efflux Pumps and apply these on a collection of *E. coli* isolates.

2.3 Methods

2.3.1 SLING specifications

SLING was implemented in Python (2.7) and is available to download from https://github.com/ghoresh11/sling. The steps of the SLING pipeline are detailed in Section 2.4.1 and in Figure 2.1.

Genome preparation Complete genomes or assembled contigs in FASTA format were six-frame translated using Biopython v1.68 [331]. By default, translation is performed using the standard codon table and the permitted start codons are [ATG, TTG, GTG]. SLING will search

for the longest CDS beginning with ATG, if it is not found it will search for the longest CDS beginning with TTG and finally GTG. Annotation files of the provided genomes in GFF format can also be provided.

Searching HMMER (v3.1b2) [296] was used to search all CDSs for the profiles of the primary gene provided by the user. The cut off used for a CDS to be considered a 'hit' for downstream analysis is a HMMER bit score of the overall sequence/profile comparison of at least 20. The cutoff was chosen based on the scores of toxin HMM profiles in known toxin sequences downloaded from TADB [318,319].

Filtering 'Partner' genes were searched in proximity to the hits according to structure requirements provided by the user. The structure requirements include the orientation of the partner gene relative to the conserved gene (upstream, downstream, or both for a three-component array), the minimum and maximum length of the conserved gene, the minimum and maximum lengths of the partner genes (upstream and downstream if applicable), and the limitations on the location of the partner gene relative to the conserved gene (maximum overlap and distance). If no partner is found under the given requirements, the hit is discarded. For the built-in HMM collections presented in this thesis, these requirements are provided by SLING; however, the default values can easily be overridden. Partner genes which have eight or more consecutive unknown nucleotides (Xs or Ns) are removed at this stage and not considered by SLING.

Profile-specific length requirements. The user can provide SLING with a file containing the expected length of proteins of each of the profiles in the HMM collection, and a limit on the maximum permitted difference between a hit's length and its expected length. This is useful when scanning for multiple profiles of conserved proteins that have versatile expected lengths.

Grouping Sequence similarity networks (SSN) are constructed for all the hits and the partners identified using protein-protein BLAST+ (v2.7) [285]. When using an orientation requirement of "either", SLING will treat upstream and downstream partners the same to form a single SSN. When using "both", SLING will generate an SSN for the upstream partners and the downstream partners separately.

Each node in an SSN is either a hit or partner sequence. An edge is drawn between two hit nodes or two partner nodes only if they meet the minimum requirements of sequence similarity as provided by the user for the BLAST output. The default requirements applied for the results

in this paper are an e-value of 0.01 and a percent identity of 30. All sequences found in the same connected component in the SSN are considered to be in the same hit/partner group.

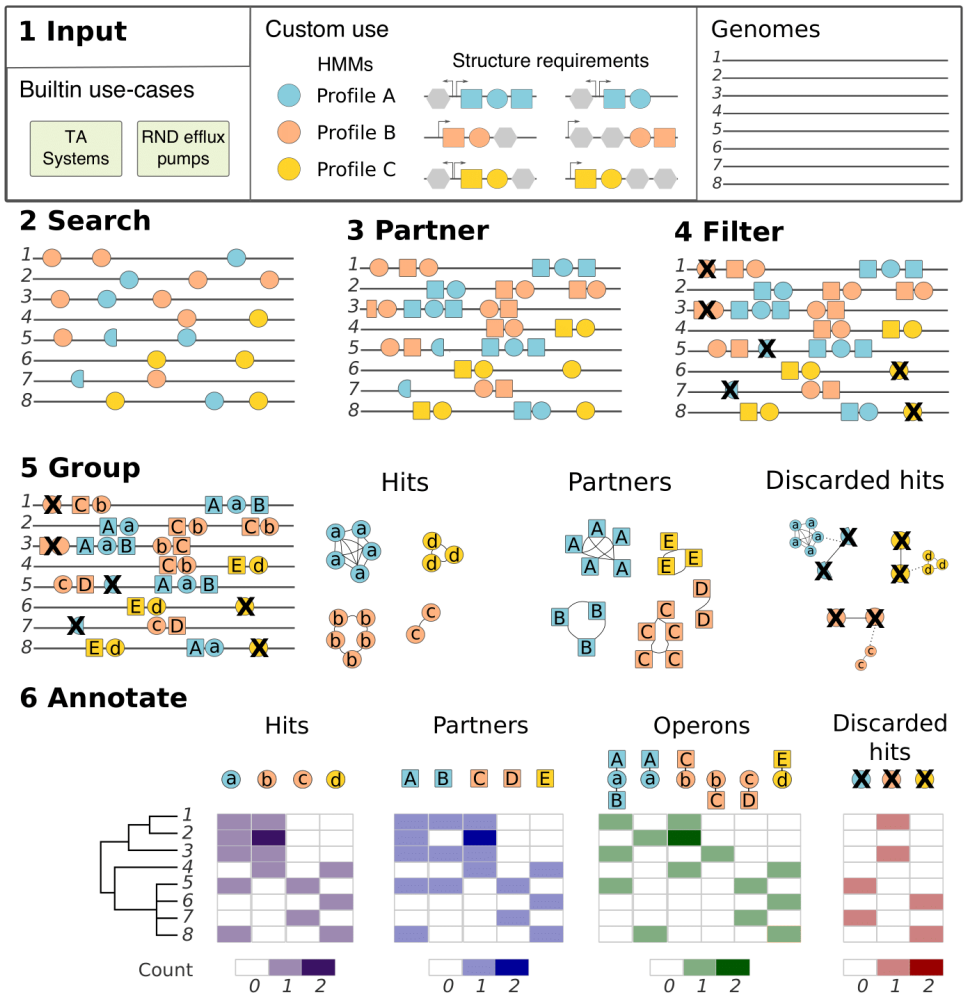

Figure 2.1: Overview of the SLING pipeline. *(1) SLING input. The user may use one of the built-in cases or otherwise provide SLING with a collection of HMM profiles and structural requirements. The structural requirements presented provide a simple example of gene arrays with multiple possible structures (top left). Grey octagons represent variable genes. Circles represent conserved genes each with a matching HMM profile represented by a unique colour which are used in the SLING search. Squares represent the partner genes consistently found in a rule-defined proximity to the conserved gene. (2) HMM profile hits are found in the input genomes. (3) Partner genes are located. (4) Partner genes are filtered based on the given structural requirements. (5) Hits, partners and discarded hits are grouped (alphabetic labelling) using sequence similarity networks. Discarded hits are mapped back to the accepted hits. (6) SLING outputs can be loaded into ITOL for visualisation of results. The phylogenetic tree must be provided for visualisation.*

Reporting discarded HMM matches The discarded hit sequences are grouped in an SSN as described above. Each connected component in this network is then mapped back to the clusters in the hits network and the discarded hit clusters are labelled according to their equivalent hit cluster.

2.3.2 Strains and phylogenetic analysis

The core gene phylogeny of 91 EPEC strains taken from [115] (See Section 1.1.2.3) was inferred from a core gene alignment generated using Roary [305], and a maximum likelihood tree from the informative single nucleotide polymorphisms (SNPs), chosen using SNP-sites [332] (v2.3.2), was constructed using RAxML (v8.2.8) [282] with 100 bootstrap replicates.

2.4 Results

2.4.1 SLING overview

SLING is a command line tool which requires a collection of assembled genomes (contigs or complete), HMMs representing a conserved gene within the gene array of interest and optional structural requirements as input (Figure 2.1). Each HMM profile is used to search the genomes for the presence or absence of the primary gene. If the gene is detected, referred to as a 'hit', SLING attempts to identify the partner protein CDSs proximal to it. The results are filtered to match the provided structural requirements. These include the distance between the partner and hit, their permitted lengths and the orientation of the 'partner' gene relative to the conserved gene. If the structural requirements are unknown, SLING will search for the closest neighbouring genes with no limitations. Hits, partners and discarded hits are grouped using SSNs. Finally, SLING reports the number of occurrences of each hit group, partner group, complete array group and discarded hit group found in each genome. These can easily be loaded into statistical analysis tools or into ITOL [333], an online tool for display and management of phylogenetic trees, creating an immediate interface for the user to examine the distribution across large datasets. SLING is available to download from https://github.com/ghoresh11/sling. Full details are provided in Section 2.3.1 and in the package wiki (https://github.com/ghoresh11/sling/wiki).

2.4.2 TA systems search

SLING can be used to search for simple two-component operons, such as TA systems. As SLING is based on a CDS search, the focus is on type II TA systems where cognate antitoxin is a protein which inhibits the toxin through direct interactions [238] (Figure 1.6). For a

complete introduction on TA systems, refer to the Section 1.3 of the Introduction. Type II systems are well studied and their structure is generally known; the antitoxin and toxin genes are transcriptionally coupled with well defined rules describing the gene orientations and distance separating them [238,316]. Moreover, TADB, which has an extensive database of type II TA systems, was available as a resource to benchmark the approach [318,319] . Following the same set of rules, type IV systems were also included in which the antitoxin is also a protein which inhibits the toxin's activity via the toxin's target [334]. Only a few type IV systems have been described so far, and appear to be rare compared to the abundant type II TA systems [334].

2.4.2.1 Construction of profile HMM library and structural requirements

To generate a collection of toxin HMM profiles, used as the primary gene in SLING, type II and type IV toxin sequences were retrieved from the web based resource for TA loci, TADB [319] and were supplemented by additional toxin sequences based on a literature search. All the toxin sequences were scanned against the Pfam protein domain database (v30.0) with HMMER (v3.1b2) to identify known toxin domains, obtaining an initial set of 153 putative HMM profiles [296,298]. These HMM profiles were manually curated to remove antitoxin domains and domains of non-protein-based TA systems which were not the subject of this investigation. Additionally, HMM profiles which had fewer than five hits were removed for further analysis unless they were a domain of a well described toxin.

A test dataset of 33 *K. pneumoniae* genomes and plasmids taken from [335] was scanned with the remaining HMM profiles. This dataset was used in order to characterise the Pfam profiles on a small collection of genomes. For each profile, the total number of HMMER hits were counted across the 33 genomes and their average length was compared to the length of the toxins containing the same profile on TADB (Figure 2.2A,B). This enabled the identification and removal of Pfam profiles which had many hits of the expected length of a toxin that do not always represent a true toxin. Keeping such profiles in the TA search would lead to high false discovery rate. For instance, the Acetyltransf domains often had a high number of hits within the expected length of a toxin and were removed (Figure 2.2B,C). Other profiles, like DUF294 and NTP_transf_2 did not have many hits, however, they did show high variability in their length relative to the lengths of the toxins containing them on TADB. For these toxins, their profiles were kept in the search and an option to apply a profile-specific-length limitation within SLING was added. Thus, only hits which were up to 100 aa longer or shorter than the average toxin length were accepted for downstream steps (Figure 2.2D). Finally, most profiles showed both a low hit count as well as fell within the range of expected lengths (Figure 2.2B,E). The final collection, following this curation step, consisted of 54 toxin profiles.

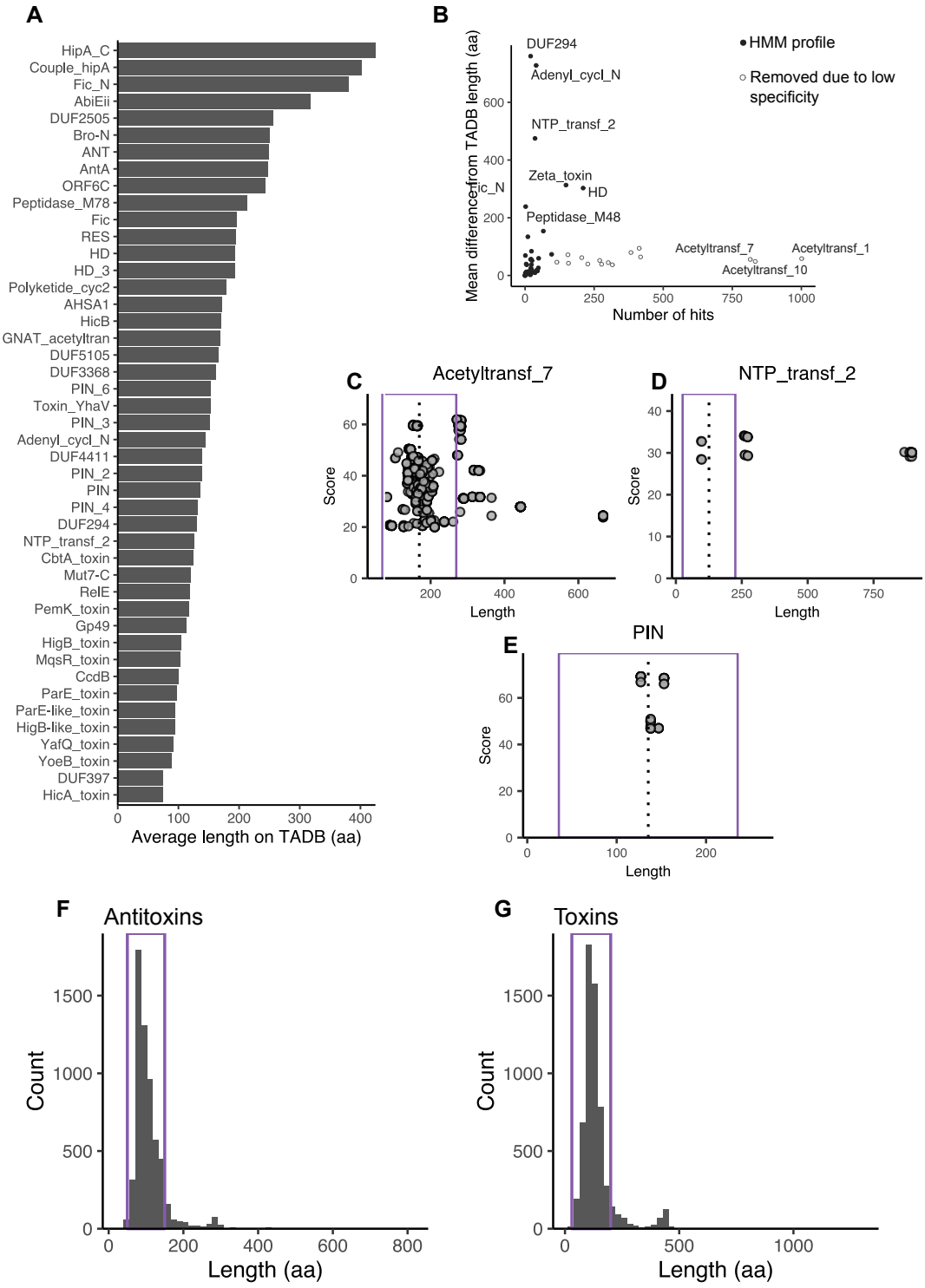

Figure 2.2: Defining the HMM collection and structural requirements for toxins. A *Mean length of toxin sequences in TADB* ***[318,319]*** *containing each of the HMM profiles.* ***B*** *Number of hits in 33* Klebsiella *genomes relative to the mean difference of those hits in protein length relative to the profiles' mean length as found on TADB (presented in* ***A)****. Empty dots are profiles which were removed due to low specificity as there were many hits which differed significantly in length relative to the length of the protein in TADB.* ***C-E*** *Length of all hits in 33*

Klebsiella *genomes relative to their HMMer bit-score. Dotted line represents the mean length of the profile in TADB (as presented in **A**). Purple rectangle represents the length cut-off defined in SLING for an ORF to be considered a valid toxin. **C** Example of low specificity HMM profile which has been removed. **D** Example of HMM profile with large length range, but with high specificity for ORFs within the expected length range. **E** Known toxin domain with small length-range and number of hits. **F, G** Length distribution of all antitoxins (**F**) and toxins (**G**) downloaded from TADB. Purple rectangles represent the length cut-offs defined in SLING.*

The length distributions of the toxin and antitoxin sequences downloaded from TADB were plotted to define the length requirements. Over 90% of antitoxins were between 50 and 150 aa long; therefore, these were used as the relevant cut-offs (Figure 2.2F). The permitted length of proteins containing toxin profiles which were present in TADB was determined based on their mean length in TADB (detailed above). Some toxin profiles were taken from a literature search and thus were not present in TADB and an average length was unavailable. For these, a minimum length cut-off of 30 aa and maximum length cut-off of 200 aa were chosen as these covered over 90% of toxin sequences in TADB (Figure 2.2G).

Table 2.1 Search parameters used in SLING

	Default	TA systems	RND efflux pumps
Order	either	either	upstream
Minimum hit length (aa)	1	30	700
Maximum hit length (aa)	10000000	200	1500
Minimum downstream length (aa)	1	50	NA
Maximum downstream length (aa)	10000000	150	NA
Minimum upstream length (aa)	1	50	100
Maximum upstream length (aa)	10000000	150	1000
Maximum distance between hit and partner (bp)	10000000	50	20
Maximum overlap between hit and partner (bp)	300	20	500
Maximum difference from average length (if given) (aa)	10000000	100	200

Finally, a distance of up to 50 bp and an overlap of at most 20 bp were permitted between the toxin and antitoxin genes. The orientation requirement was set based on the knowledge that the partner gene, i.e. the antitoxin, can be either upstream or downstream of the toxin gene (Table 2.1) [316].

2.4.2.2 The process for setting up a TA search are applicable to other operons

A similar process can be applied to construct the HMM profile libraries of other genes and to define the structural parameters. Another example will be presented in Section 2.4.3.1 and the general approach is summarised in Figure 2.3. HMM profiles can also be generated directly from an MSA of a collection of genes using HMMER [296]. Finally, if the structural requirements are unknown, SLING provides default parameters for a flexible search which will identify the closest partner genes proximate to the primary gene (Table 2.1).

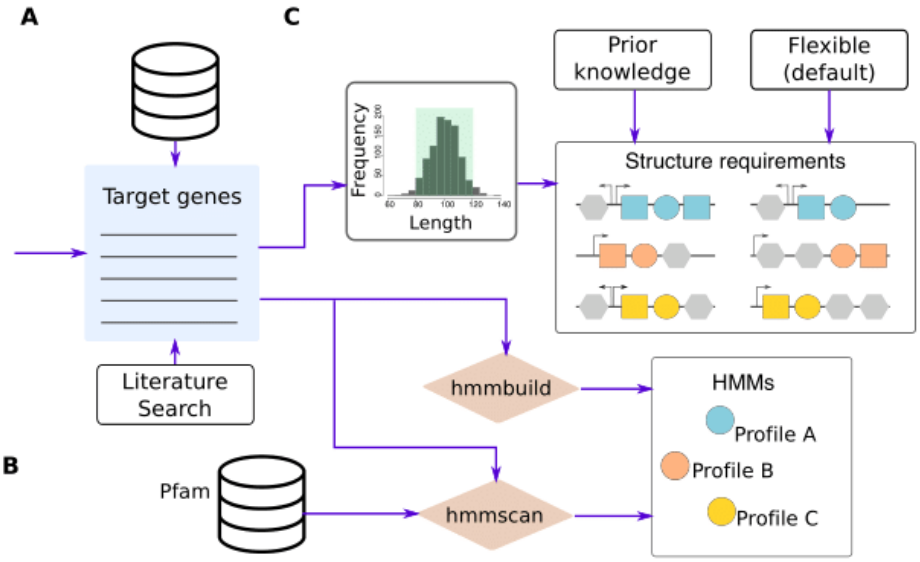

Figure 2.3: General construction of HMM profiles and structural requirements for SLING input. A *A collection of known target genes is required, taken from existing databases (toxins; TADB, RND pumps; Uniprot), a literature search or other sources.* ***B*** *HMM profiles can be generated directly from an MSA of the target sequences using HMMER [296]* hmmbuild *or can be scanned by HMMER* hmmscan *against existing HMM profile databases, for instance, Pfam [298].* ***C*** *Structural requirements can be inferred from the target gene sequences, known from prior knowledge or otherwise, flexible using SLING's default parameters.*

2.4.3.3 Benchmark on *E. coli* K-12

SLING identifies new and known TA systems in *E. coli* K-12.

SLING was used to search *E. coli* K-12 strain MG1655 (NC_000913.3) for TA systems. SLING identified 23 TA systems in total (Figure 2.4B). These results were compared to the *E. coli* K-12 strain MG1655 TA systems in TADB and those predicted by TAfinder using the same parameters used in SLING [318,319]. Nine of the 23 systems were identified by all three methods. TADB missed five TA predictions which were identified by the other two methods, whereas TAfinder missed one. A single system, identified by TADB, was missed by both SLING and TAfinder, the rnlAB system. The RnlA toxin has a length of 397 aa, beyond the maximum length threshold of 200 aa for a toxin applied in our implementation.

SLING identified eight TA systems which were not predicted by TADB or TAfinder. Of these, four have been predicted in the past to be TA systems; the YkfI-YafW system [334,336], the GnsAB TA system [337], the RatAB system [338] and the YdaST system [339]. Four more predictions have not been previously described as TA systems and are candidates for further investigation. One contains an HD domain, two contain a GNAT domain and the last contains a YdaT toxin domain, consistent with their proposed function.

TADB and TAfinder identified TA systems that were not identified by SLING. Thirteen of the TADB results belonged to TA system classes that were not investigated in this study. An additional two toxins were predicted which, using HMMER, did not contain any described toxin profile used by SLING. Finally, TAfinder predicted three TA systems which we attempted to retrieve from the reference genome but were unable to identify complete CDSs at the relevant locus.

2.4.2.4 Application on EPEC collection

To search for TA systems in a diverse set of related bacteria SLING was applied with the settings described for TA search on a collection of 70 EPEC isolate genomes taken from [115], supplemented by an additional 21 commonly studied *E. coli* reference strains (taken from [115]). The EPEC isolates were collected from children presenting with diarrhoea from seven centres in Africa and Asia [115].

SLING identified a total of 94 different TA operons in the complete *E. coli* collection built of 44 toxin (hit) clusters and 80 antitoxin (partner) clusters. SLING generated an output of the absence and presence of these systems across the dataset that can be loaded into a statistical learning tool, enabling to look for association with the metadata and view in ITOL. Below are examples of three toxins which are presented to illustrate the type of visualisation, analysis and interpretation that can be accomplished using SLING (Figure 2.4C).

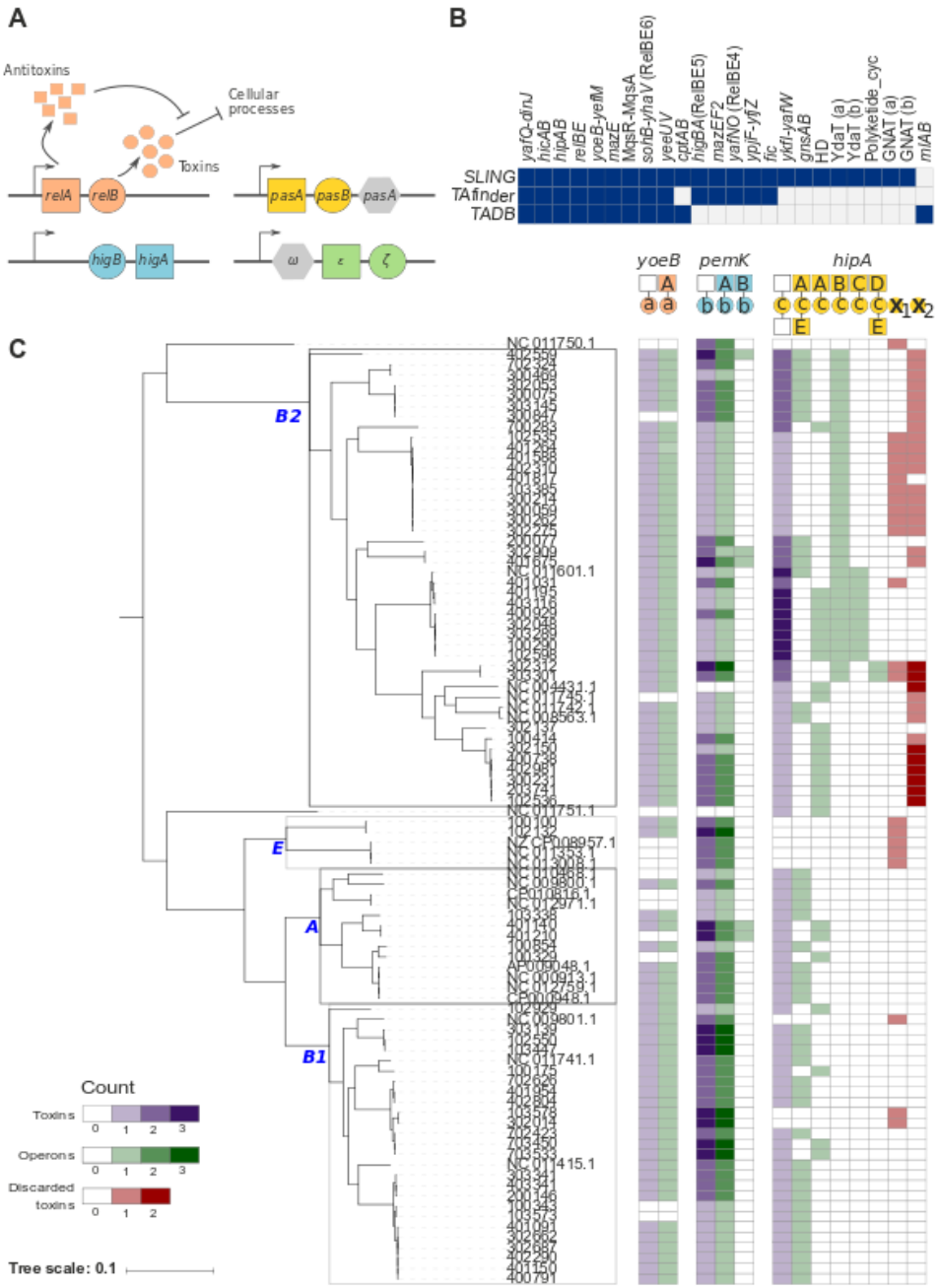

Figure 2.4: Identification of TA systems using SLING. A *Possible operon structures of TA operons. Each toxin has a unique HMM profile, represented by a different colour.* ***B*** *Identification of TA systems in* E. coli *K-12 using SLING, TADB and TAfinder. Prediction of a TA operon by a method is represented in a dark blue square. Novel TAs predicted using SLING named by the Pfam profile by which they were identified.* ***C*** *Description of the diversity of three toxins and their cognate antitoxins in the* E. coli *collection. Darker squares represent presence of a toxin or operon in an isolate.*

YoeB toxin presents low antitoxin repertoire, with low evidence of gene loss/gain. The YoeB profile containing toxin was always identified as partnered to the same antitoxin. This TA pair was ubiquitous, present across all phylogroups. In addition, there was no evidence of duplication events, with a single copy of the operon identified in each isolate. *yoeB* was never found as an orphan toxin, however there were examples of loss or gain of the whole operon in nine locations in the phylogeny, i.e. the antitoxin was never lost on its own. This observation strengthens the hypothesis that this protein serves as a toxin in a TA system.

PemK toxin presents medium antitoxin repertoire, with high evidence of gene loss/gain. The second toxin (Figure 2.4C), containing a PemK profile, showed diversity in its antitoxin repertoire: it was found with two different antitoxins: A and B. Most copies of this toxin were observed with one of the antitoxins (A; 97%), which was present across all the phylogroups. For this operon, there was a strong indication of gain events followed by fixation and vertical propagation; a subclade with a copy number of *n* was often found within a clade with copy number *n-1*. This phenomenon occurred independently multiple times in the phylogeny. The pervasiveness of this operon can either allude to its importance, or otherwise, suggests it is successful at spreading in the population and persisting. The second operon (B) was rare and found only in five isolates in a single copy. It was most likely acquired in three independent events. Finally, like *yoeB* toxin, this toxin was always found partnered to an antitoxin.

HipA toxin presents a high antitoxin repertoire, with low evidence of gain/loss of the same genes. The final toxin (Figure 2.4C), containing a HipA profile, presents a higher diversity in its antitoxin repertoire with five candidate antitoxins. Four of these antitoxins (A-D) are upstream to the toxin, whereas the last antitoxin (E) was found downstream to the toxin and was always present with one of the upstream antitoxins.

Looking at their phylogenetic distribution, although many of the isolates have more than one copy of the *hipA* toxin, it was apparent that within one genome each individual toxin gene was partnered with a different antitoxin. The majority of toxin genes were linked to antitoxin A (62%), which together were present across all phylogroups (Figure 2.4C). The three other antitoxins (B, C and D) are lineage specific and were only present in phylogroup B2. Interestingly, all isolates with antitoxins C or D also had antitoxin B.

Although *hipA* is a well described toxin, we observed multiple cases in which SLING filtered the predicted toxin gene out due to deviations from the expected operon structure of a TA system (Figure 2.4C). These genes were marked as discarded by SLING as a result of this. However, analysis of these discarded toxins showed that they formed two separate sequence

clusters: X_1 and X_2. All the X_1 toxins coincided with isolates which were missing the A antitoxin. As for X_2, all the discarded toxins were within phylogroup B2, coinciding with isolates which were missing antitoxins B and C.

2.4.3 RND efflux pumps search

Efflux pumps play an important role in multidrug resistance as they confer a mechanism for the efflux of antibiotics [340]. One example of this are the RND family of membrane transporters found in Gram-negative bacteria [327,341]. RND family pumps consist of three components: an outer membrane protein (OMP), a periplasmic fusion protein (MFP) and an RND pump (Figure 2.5A). In most cases, the MFP and RND components are found in an operon, whereas the OMP is located in a different location [327]. RND efflux pump operons, unlike TA systems, are complex operons which often include a large range of genes often found in different orders and orientations [327]. However, these operons always contain an RND efflux pump protein which is highly conserved and, in most instances, the MFP is located upstream of it and transcriptionally coupled to it [327]. This property makes these operons relevant for a search using SLING by setting the RND protein as the primary gene and applying flexible structure requirements on the partner gene.

2.4.3.1 Construction of profile HMM library and structural requirements

3,325 RND efflux pump sequences were downloaded (on 07.11.17) from Uniprot [342] by searching for the name of 26 known RND pump genes (Figure 2.6A) [343]. The sequences originated from 295 different genera. Sequences were clustered using cd-hit (v4.7) to remove redundant sequences which share 90% identity [344]. The remaining 1,242 sequences were searched using HMMER (v3.1b2) against the Pfam database (v30.0) to identify known RND pump domains [296,298] (Figure 2.3B). A total of 29 Pfam profiles were identified in these sequences, of which a single profile, ACR_tran (PF00873), was present in over 99% of the sequences and thus was chosen to represent all RND pumps.

The length distribution of the above mentioned RND pump proteins were plotted (Figure 2.6B). A minimum length of 700 aa long and maximum length of 1500 aa long were chosen for the RND pump protein, covering over 94% of the downloaded sequences. For the partner gene, 23,133 MFP sequences were downloaded (on 07.11.17) from Uniprot [342] by a keyword search. The length distribution of these proteins was plotted and a minimum length of 100 aa and maximum length of 1000 aa were chosen as flexible requirements for different partner genes as these thresholds cover the length of over 99% of membrane fusion proteins

downloaded [342] (Figure 2.6C). Finally, a maximum of 500 bp distance between the partner and the RND pump, and at most 20 bp overlap were allowed (Table 2.1).

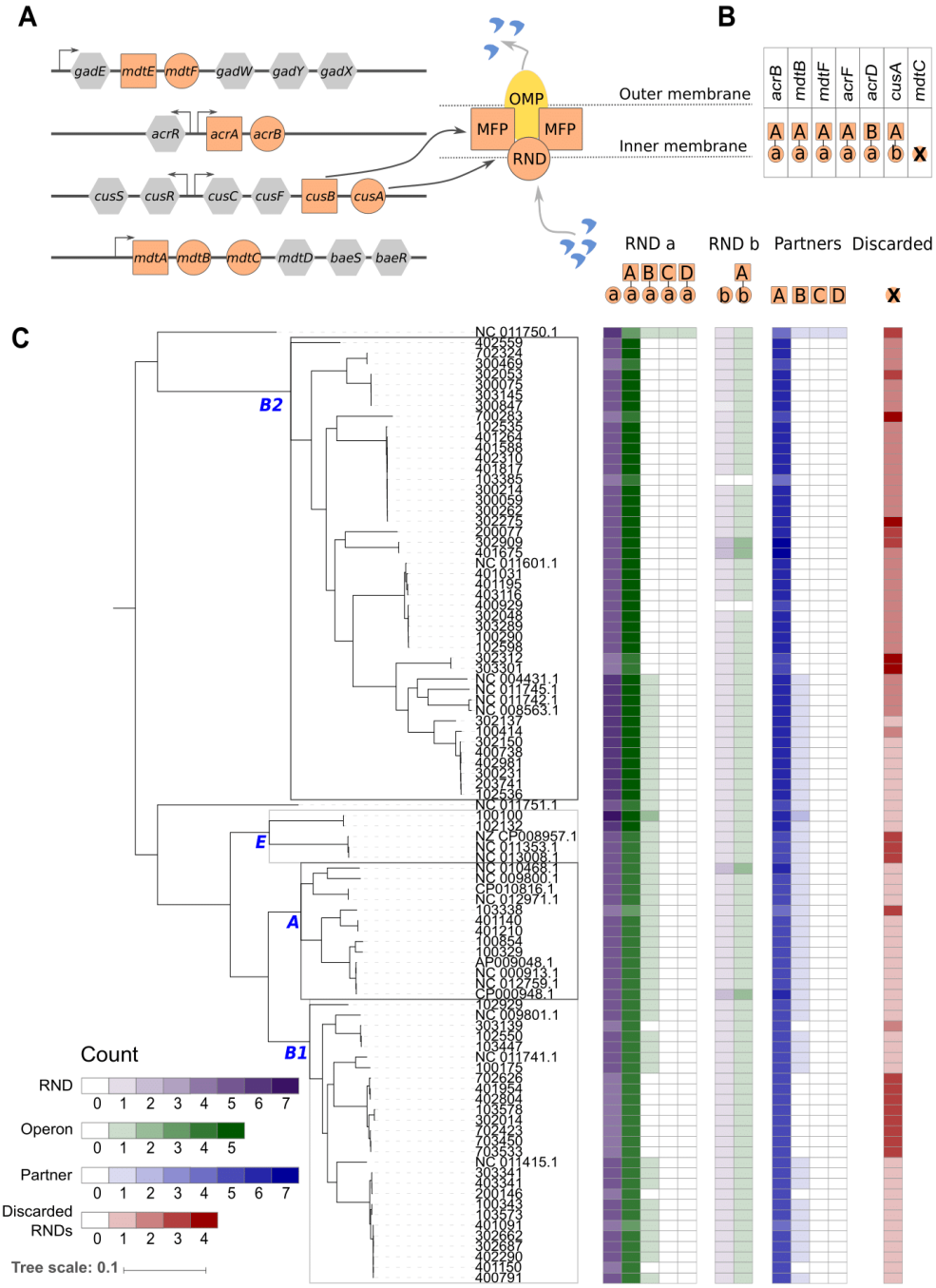

Figure 2.5: Identification of RND efflux pumps using SLING. A *Four example operon structures of RND efflux pumps present in* E. coli *K-12. All RND pump proteins share a single conserved HMM profile, represented by a single colour (ACR_tran;PF00873).* ***B*** *The corresponding annotation of RND efflux pumps in* E. coli *K-12 relative to the SLING output.* ***C*** *Annotation of RND efflux pumps in the* E. coli *collection. Darker squares represent presence of an RND pump protein or an operon in an isolate.*

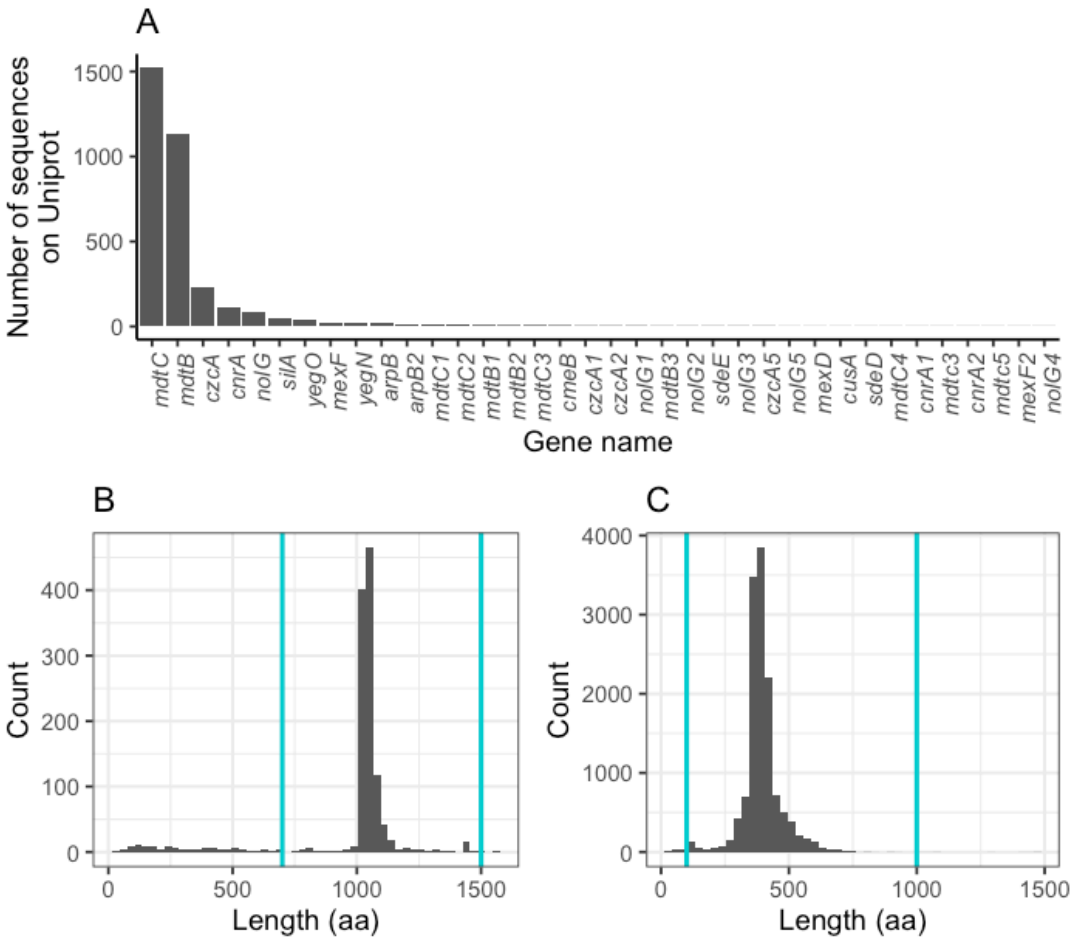

Figure 2.6: Defining the HMM collection and structural requirements for RND efflux pumps. A *Number of sequences retrieved from Uniprot using a name search of known RND efflux pumps genes.* ***B,C*** *Length distribution of RND efflux pump proteins (**B**) and MFPs (**C**) downloaded from Uniprot. Turquoise lines represent the cut-offs chosen as the length structural requirements for search using SLING.*

2.4.3.2 Benchmark on *E. coli* K-12

Seven RND efflux pumps are reported in the literature for *E. coli* K-12 strain W3110 (AP009048.1) [327]. Of these, SLING identified six RND pumps which fit the structure requirements applied in our analysis: *acrB*, *cusA*, *mdtB*, *acrF*, *acrD* and *mdtF* (Figure 2.5B). Since *mdtC* pump is found downstream to another RND pump, *mdtB*, (Figure 2.5A) this pump was discarded by SLING as the upstream gene was not in the correct length.

2.4.3.4 Application on EPEC collection

Five unique RND pump operons were identified in a SLING search on the collection of 90 EPEC and reference *E. coli* strains (Figure 2.5C). These operons consisted of two unique RND protein (hit) clusters (a and b) and four partner protein clusters (A-D).

The A partner protein is indeed an MFP and includes all the known MFPs found in *E. coli* K-12 (Figure 2.5B). It was highly prevalent and was observed in two different operons, with the two RND pump proteins (a and b). The “A-a” operon was ubiquitous, with at least four copies per strain. Reducing the identity threshold applied to group the proteins would have likely separated this operon into its corresponding operons in K-12. The “A-b” operon, on the other hand, was found in a single copy in most isolates. The “b” pump corresponded to the *cusA* RND pump in *E. coli* K-12, whereas the “a” pump represented all the other known RND pumps in *E. coli* K-12 (Figure 2.5B).

The B partner protein is a histidine kinase. This protein is identical in sequence to the *narQ* gene, found upstream to the *acrD* RND pump in *E. coli* K-12 [327]. This operon was missing in specific clades within the B1 and B2 phylogroups. These clades were correlated with the discarded hits, suggesting two events occurred that led to deviation from the expected operon structure in these clades.

Finally, the C and D partner proteins were only observed once and in a single isolate (ExPEC reference strain, *E. coli* IAI39). Both proteins were short with “C” partner protein 138 aa long and the “D” partner protein 310 aa long. BLAST results of protein “C” against the non-redundant protein sequence database suggest it is a histidine kinase similar to partner protein “B” (*narQ*). Protein “D”, on the other hand, is a truncated RND pump protein.

2.5 Discussion

SLING is an open source tool to examine the diversity of operons or gene arrays in bacterial datasets by using one of the conserved genes within the array to identify the linked genes which appear in a rule-defined proximity (Figure 2.7A). By examining the diversity of the neighbouring genes, we can elucidate incidences where there are deviations in the operon structure between isolates as well as deviations from what is expected to be the canonical operon structure of a specific system (Figure 2.7B). Examples of this were presented for the diversity of toxins as well as RND efflux pump proteins and their partner genes (i.e. antitoxins and MFPs) in a collection of *E. coli* isolates. While some genes presented a high diversity in their possible neighbours, others presented low diversity. Likewise, by examining the diversity of the neighbouring genes, SLING helped to further sub-categorise the gene combinations according to varying indications of these arrays being lost or gained.

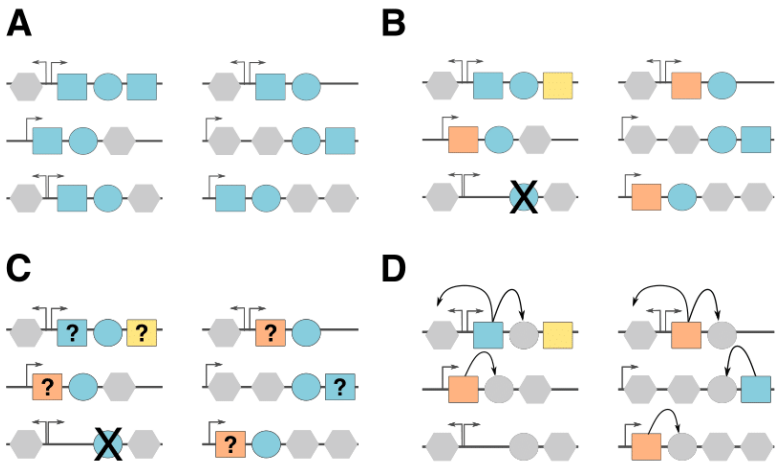

Figure 2.7: Utility of SLING. A *Search for gene pairs and triplets based on a single conserved gene (circle) and set of rules on the order and orientation of the neighbouring genes (squares)* ***B*** *Test the defined rules by examining the diversity of the neighbouring genes and identifying gene arrays which deviate from the expected structure* ***C*** *Directly identify new genes (squares)* ***D*** *Iteratively identify new genes by using the novel neighbour genes (squares) as the input HMM profiles.*

Two settings for TA systems and RND efflux pumps were described and these are built into the SLING interface for quick application using simple command line prompts, which are detailed on the tool's wiki page (https://github.com/ghoresh11/sling/wiki). Beyond these, SLING's advantage is in its flexibility; users can easily provide new profiles into its search, enabling identification of new and not well studied systems without relying on the developer to update the code or database. Thus, the utility of SLING is not limited to these operons and can be applied to other important operons or gene pairs such as CRISPR-cas systems, restriction-modification systems, secretion systems, and more. Users may construct HMM libraries and structural requirements in their area of expertise which can be shared with the community by uploading them to the public repository, enabling the extension of the built-in SLING use cases.

Additional advantage of SLING is that its protein search is based on an HMM profile search, rather than a sequence-based search, which allows SLING to capture more diverse members of a protein and not rely on a single sequence, likely taken from a lab strain which may not be representative in a collection of clinical or natural isolates. This advantage is also a limitation, it may be difficult to construct an HMM profile for an unknown gene when not many representative sequences are available. SLING also searches for the genes using a six-frame translation of the input genomes in addition to searching the CDSs predicted by annotation

tools. This allows the identification of short CDSs which may have otherwise been omitted by the annotation tools.

When searching for an unknown set of linked genes, SLING can also be used as a discovery tool. By applying default flexible structural requirements to find a partner gene, SLING can identify any set of genes which are linked to the primary gene. SLING can also be used to search for novel genes either directly, by looking at the partner genes identified (Figure 2.7C), or indirectly, but constructing HMM profiles of the newly identified partner genes and iteratively using these as the conserved gene (Figure 2.7D). These ideas are explored in Chapter 3 of this thesis, where SLING was used to examine the diversity of TA systems across a global collection of *K. pneumoniae* isolates.

3 The diversity of type II and type IV toxin-antitoxin systems in the global *K. pneumoniae* population

This chapter is a modified version of the paper "Type II and type IV toxin-antitoxin systems show different evolutionary patterns in the global K. pneumoniae *population" [345]. Cinzia Fino, Matthew Dorman and Alexander Harms conducted the phenotypic experiments for testing the activity of toxins and antitoxins which were selected by me. Leopold Parts, Kenn Gerdes, Eva Heinz and Nicholas Robert Thomson contributed to the research of the original publication. All final language is my own.*

3.1 Introduction

TA systems are bicistronic operons which encode for a toxin, which inhibits cellular processes, and an antitoxin which counteracts the toxins' activity [346], and were introduced in Section 1.3 of this thesis. While TA systems have been well studied in a limited number of laboratory and clinical isolates of *E. coli* [347–350] and *Salmonella enterica sv.* Typhimurium [253], there have been few studies in any bacterium that have considered investigating these systems using large clinically relevant collections. In this chapter, SLING, which was presented in Chapter 2 as a tool to search for operons in large datasets, is used to examine the diversity of TA systems across a collection of *K. pneumoniae* genomes.

Since their first description as plasmid addiction systems, it has become clear that TA systems are ubiquitous across a broad range of prokaryotic plasmids and chromosomes [251,316,335,350–353]. The first study examining the distribution of TA systems on a large scale was conducted in 2005, when Pandey and Gerdes used BLAST to search for TA loci across 126 prokaryotic genomes. It was then revealed that TA systems were highly abundant in the chromosomes of free-living Gram-negative and Gram-positive bacteria [316]. In 2009, Makarova et al. used a guilt-by-association approach to identify novel type II TA combinations across the non-redundant protein and COG databases [251,300]. The distribution of the predicted TAs was examined across large evolutionary scales, and it was revealed that specific TAs were significantly over- or under-represented in various taxa, suggesting different dynamics to their propagation depending on the genetic and ecological backgrounds of their host. Additionally, the distribution of TAs was examined in more detail in a set of 41 closely related prokaryotic genomes. An exceptionally high level of variability in the TA system repertoire was observed, even at these close evolutionary ranges. These results paved the path for future studies, as it was evident that these were highly diverse genetic systems which

have yet to be explored. Since, studies on the distribution of these systems were mostly focused on small high-quality genome collections of reference laboratory strains, which do not necessarily represent the diversity in clinical samples [335,350]. Nonetheless, a study on the distribution of type II TA systems in *E. coli* revealed that these systems were differentially distributed across the *E. coli* phylogroups [350]. A similar study in *K. pneumoniae* revealed that type II TA systems are differentially distributed across *K. pneumoniae* isolates from different sources and across plasmids and chromosomes [335].

The large range of TA systems and their ubiquitous nature across species, plasmids and chromosomes suggest that these elements have an essential role in prokaryotic cell biology, beyond their role in plasmid maintenance. Indeed, they have been implicated in other important cellular processes, many of which contribute to resistance and pathogenicity (See Section 1.3.3). These include the formation of antibiotic-induced persistence [348], defence against bacteriophages, biofilm formation [346,354,355], and through transcriptional read-through, influence the expression of adjoining genes [255]. Therefore, a more systematic approach which examines these systems in a collection of clinically relevant genomes can reveal whether their presence is associated with clinically important genes.

3.2 Aims

The aim of this chapter was to use SLING to systematically analyse the diversity of TA systems in a collection of 259 *K. pneumoniae* isolates. The precise aims of this chapter were:

- Describe the distribution of toxins and their antitoxins in a global and clinically relevant collection of *K. pneumoniae* isolates using SLING.
- Test the activity of predicted toxin and antitoxin pairings
- Examine the connection between the presence of these systems and the presence of clinically important genes including AMR genes and virulence genes.

3.3 Methods

3.3.1 Strains and phylogenetic analysis

Assemblies of 259 *K. pneumoniae* species complex strains taken from [9] were assembled using VELVET (v1.2.07) [356] and annotated using Prokka (v1.5) [293] [357]. The core gene phylogeny was inferred from a core gene alignment generated using Roary [305], and a maximum likelihood tree from the informative SNPs, chosen using SNP-sites [332] (v2.3.2), was constructed using RAxML (v8.2.8) [282] with 100 bootstrap replicates.

3.3.2 Toxin-antitoxin prediction

SLING (v1.1) [311] was used to search for toxins and their cognate antitoxins using the built-in toxin domain database provided in SLING. Please refer to Chapter 2 of this thesis for a complete description of SLING's search strategy. The default structural parameters for a TA search in SLING were applied in the filtering step (minimum toxin length: 30 aa, maximum toxin length: 200 aa, minimum antitoxin length: 50 aa, maximum antitoxin length: 150 aa, maximum overlap between toxin and antitoxin: 20 bp, maximum distance between toxin and antitoxin: 50 bp, order: antitoxin either upstream or downstream to toxin, maximum difference: 100 aa). A cut-off of 75% aa sequence identity was used during the grouping step.

The local sequence identity and alignment coverage per toxin and antitoxin group were extracted from the BLAST+ results from the SLING output. All the antitoxin and toxin sequences from each group were aligned using MUSCLE (v3.8.31) [358]. The global sequence identity was calculated as the pairwise sequence identity between every two sequences in the MSA.

3.3.3 Statistical analysis

Statistical analyses were performed in R (v3.3.1). Toxin and antitoxin accumulation curves were generated using the *specaccum* function in the vegan [359] library with 100 random permutations. PCA was performed using the *prcomp* function. Association between toxins and lineage or the presence of AMR genes, virulence genes or plasmid replicons were performed using Fisher's exact test and corrected for multiple testing using the False Discovery Rate (FDR) with the *p.adjust* function. Differences between groups (*K. pneumoniae* complex species, toxin categories) were assessed using the Wilcoxon test and corrected using FDR. Plotting was done using ggplot2 [360].

3.3.4 Toxin group classification

Toxin groups which were observed in over 80% of isolates of all species were assigned as "ubiquitous". Toxin groups which had at least 4 copies and were found to be significantly associated with *K. pneumoniae* complex species (Fisher's exact test, FDR corrected, $p<0.01$) were assigned "species associated". Toxin groups which were not ubiquitous or species associated were assigned "sporadic" if they had 26 copies or more or otherwise, if they were found to be significantly associated with the presence of AMR genes, virulence genes or plasmid replicons (Fisher's exact test, FDR corrected, $p<0.01$). The remaining toxin groups were assigned "rare".

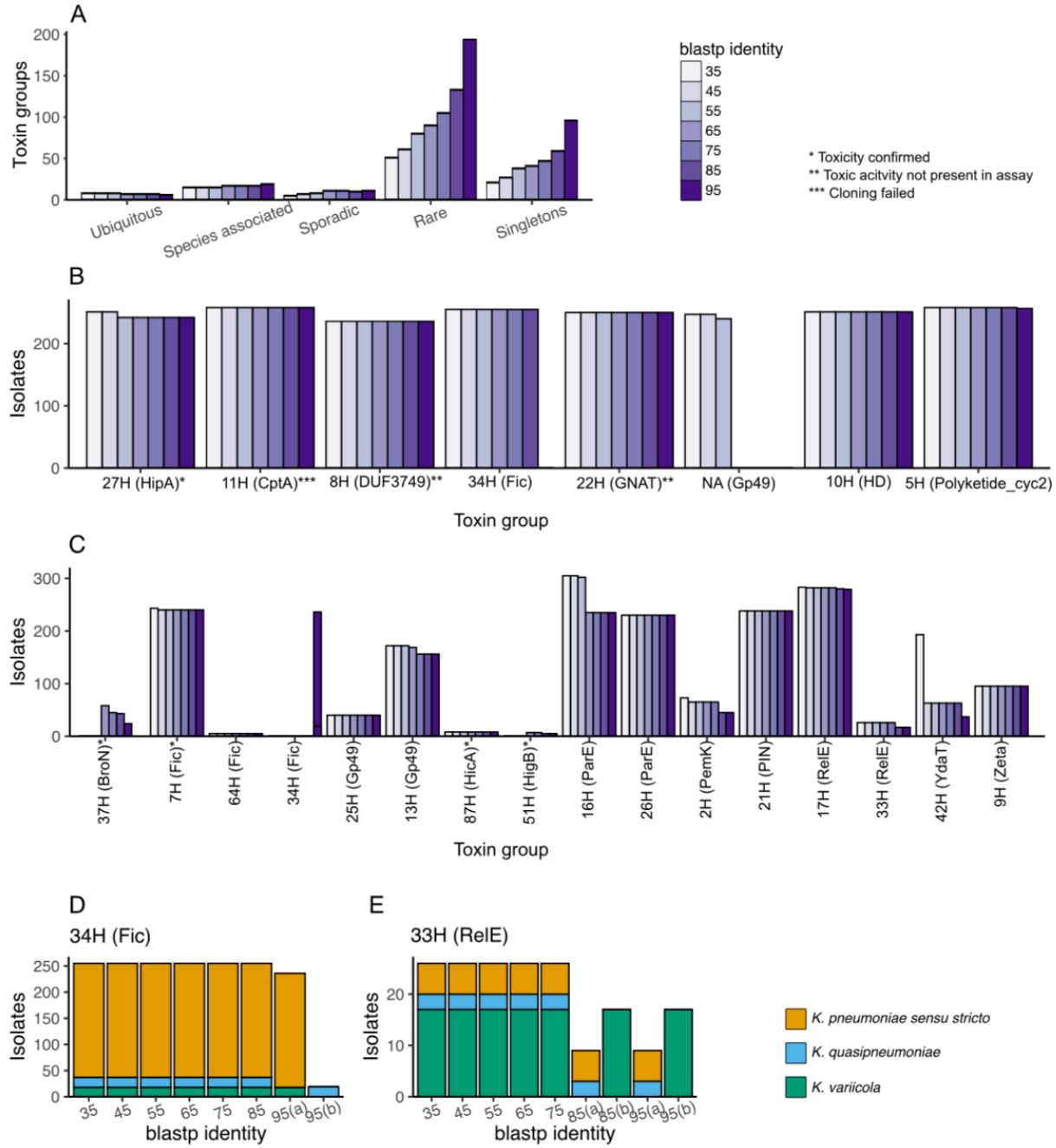

Figure 3.1: Effect of modifying the blastp identity threshold in SLING on the toxin group clustering. A *number of toxin groups from each toxin class for each identity threshold applied. Singletons are toxin groups with only one member.* ***B,D*** *Ubiquitous (**B**) and species associated (**C**) toxin groups under each identity threshold applied. When a bar is missing, the toxin group was not classified as ubiquitous or species associated under the given threshold.* ***D,E*** *Examples of clusterings across thresholds for a ubiquitous toxin group Fic (**D**) and a species associated toxin group RelE_1 (**E**).*

Changing the sequence similarity thresholds for grouping toxins increased the number of toxin groups, however the number of ubiquitous, species-associated and sporadic toxin groups

stayed constant. There was an increase in the number of rare toxin groups which is driven by an increase in the number of singleton toxin sequences (Figure 3.1A). The ubiquitous toxin groups and species-associated toxin groups were robust and stable across all identity thresholds (Figure 3.1B,C). The chosen BLAST identity cut-off of 75% allowed separation of sequences which share similar domains, for instance, DNA binding domains, yet kept homologous sequences together and did not separate sequences by species due to drift (Figure 3.1D,E).

3.3.5 Definition of novel vs known antitoxins

All *in-silico* predicted and experimentally validated type II and IV antitoxin sequences were downloaded from the toxin-antitoxin database TADB (v2, downloaded on 27.08.17) [318,319] and pairwise comparisons between all antitoxin sequences identified by SLING were performed using protein-protein BLAST+ (v2.7) [285]. A SLING antitoxin group was marked as “known” if one or more of the antitoxins in that group shared at least 75% identity and an e-value of 0.01 or lower with an antitoxin from TADB (consistent with the definition of an antitoxin group). Interpro-scan (v5) was used to assign function to the sequences of the novel antitoxins [361]. Sequences which were assigned as antitoxins by Interpro-scan were also marked as “known”. Otherwise, the group was marked as “novel”.

3.3.6 Orphan antitoxins

Antitoxin sequences from an antitoxin cluster were grouped using cd-hit (v4.7) [344] with an identity threshold of 90% and word size of 5 to remove redundant sequences. An antitoxin protein database of the cd-hit representative antitoxins was constructed using BLAST+ (v2.7) [285]. The six frame-translated *K. pneumoniae* genomes from the SLING output [311] were aligned against the antitoxin database using blastn [285]. A CDS was considered an “orphan antitoxin” if a) it was between 50 and 150 aa long, b) it shared 75% sequence identity or more to an antitoxin in the collection and c) the alignment was 50 aa or longer. These settings were chosen to be consistent with the definitions of an antitoxin in the original SLING analysis. The sequences 1,000bp upstream and downstream to the orphan antitoxins were clustered with the respective 1,000bp sequences surrounding the original antitoxin in the viable toxin-antitoxin pair using cd-hit-est with 80% identity threshold and word size of 5. If orphan antitoxin context sequences were in the same cd-hit cluster as the sequences of the original antitoxin, they were marked as “same” and “different” otherwise.

3.3.7 Identification of AMR genes, virulence genes and plasmid replicons

A collection AMR genes were obtained from the modified version of ARG-ANNOT available on the SRST2 website (https://github.com/katholt/srst2/tree/master/data, downloaded on 02.10.16) [288,290]. A dataset of virulence factors was obtained from the *Klebsiella*-specific BIGSDB (http://bigsdb.pasteur.fr/klebsiella/klebsiella.html, downloaded on 22/02/16). The PlasmidFinder database (v1.3) of plasmid replicons was downloaded using ARIBA (v2.12) [283,287]. Presence or absence of a gene in a genome was determined using ARIBA (v2.12) with default settings [283]. Nucleotide-nucleotide BLAST+ (v2.7) of the assemblies against the target gene databases was used to identify contigs which contained a gene of interest (AMR, virulence or plasmid) [285]. A match was determined if any of the associated genes had a BLAST bit score of 200 or more.

3.3.8 Phenotypic testing[5]

Bacterial strains, plasmids, and oligonucleotides used in this study are listed in Appendix A. The sequences of synthesised genes, including mutated ribosomal binding sites and restriction sites where appropriate, are listed in Tables 3.1 and 3.2.

Strains were cultured routinely on lysogeny broth (LB) media. Where appropriate, bacteria harbouring plasmids were cultured on LB media supplemented with 100 µg/ml ampicillin or 30 µg/ml chloramphenicol.

Toxin and antitoxin sequences predicted from computational analysis were synthesised, cloned, and sequence-verified using the GeneArt DNA synthesis service (ThermoFisher Scientific, DE). Toxin sequences were cloned into pNDM220 under P*lac* control [362], and antitoxin sequences into pBAD33 under P*ara* control [363] (Tables 3.1 and 3.2). LB agar plates were supplemented with 1 mM of isopropyl β-D-thiogalactopyranoside (IPTG) for the induction of P*lac* and 0.2% w/v of L-arabinose for the induction of P*araB*. Overnight cultures were washed once and then serially diluted (10^{-1} to 10^{-6}) in sterile phosphate-buffered saline (PBS). 10 µl of the original and diluted cultures (10^{-1} to 10^{-6}) were spotted on LB agar plates containing the induction supplements.

[5] This work was conducted and written by Cinzia Fino, Matthew Dorman and Alexander Harms.

Table 3.1 Phenotypic testing of identified toxins.

Toxin ID	Constuct ID	Pfam domain	Status	Category	TA type	5' Restriction Site	3' Restriction Site
doc	pMJD119	doc	Control - toxic	Control	II	KpnI	KpnI
27H	pMJD127	HipA	Toxic	Ubiquitous	II	KpnI	KpnI
61H	pMJD130	CcdB	Toxic	Sporadic	II	KpnI	KpnI
51H (39P)	pMJD128	HigB	Non-Toxic	Species associated	II	KpnI	KpnI
51H (147P)	pMJD131	HigB	Toxic	Species associated	II	KpnI	KpnI
8H	pMJD121	DUF3749	Non-Toxic	Ubiquitous	II	KpnI	KpnI
87H	pMJD132	HicA	Toxic	Species associated	II	KpnI	KpnI
24H	pMJD134	Gp49	Toxic	Sporadic	II	KpnI	KpnI
72H	pMJD129	HD	Non-Toxic	Sporadic	II	KpnI	KpnI
12H	pMJD122	RES	Non-Toxic	Sporadic	II	KpnI	KpnI
44H	pMJD138/9	ParE	Toxic	Sporadic	II	KpnI	KpnI
14H	pMJD133	Gp49	Toxic	Sporadic	II	KpnI	KpnI
31H	pMJD125	HicA	Toxic	Rare	II	KpnI	KpnI
54H	pNDM_54H	BroN	Non-Toxic	Rare	II	KpnI	KpnI
22H	pMJD124	GNAT	Non-Toxic	Ubiquitous	II	KpnI	KpnI
7H	pMJD120	Fic	Toxic	Species associated	II	KpnI	KpnI
11H	Failed	CptA	Cloning failed	Ubiquitous	IV	KpnI	KpnI
37H	pMJD126	BroN/ANT	Toxic	Species associated	II	KpnI	KpnI
18H	pMJD123	CcdB	Non-Toxic	Sporadic	II	KpnI	KpnI

Table 3.2 Combinations of toxin-antitoxins tested for antitoxin inhibition.

Antitoxin ID	Paired toxins	Operon structure	Status	Novelty	Predicted function	5' Restriction Site	3' Restriction Site
PhD	doc		Control - inhibited	Control	Control	KpnI	HindIII
52P (31H)	31H (HicA)	52P-31H	Inhibited	Novel	Domain of unknown function (DUF1902)	KpnI	HindIII
52P (54H)	31H (HicA)	54H-52P	Inhibited	Novel	Domain of unknown function (DUF1902)	KpnI	HindIII
3P	7H (fic)	3P-7H	Inhibited	Known	Known	KpnI	HindIII
168P	7H (fic)	3P-7H-168P	No inhibition	Novel	Unassigned	KpnI	HindIII
24P	27H (hipA)	24P-27H	Inhibited	Known	Known	KpnI	HindIII
27P	14H (Gp49)	14H-27P	Inhibited	Novel	DNA binding	KpnI	HindIII
23P	44H (ParE)	44P-44H-23P	No inhibition	Novel	Unassigned	KpnI	HindIII
44P	44H (ParE)	44P-44H	Inhibited	Novel	Unassigned	KpnI	HindIII
45P	87H (hicA)	87H-45P	Toxic	Known	Known	KpnI	HindIII
48P	61H (CcdB)	48P-61H	Inhibited	Known	Known	KpnI	HindIII
62P	37H (BroN)	62P-37H	Toxic	Novel	consensus disorder prediction	KpnI	HindIII
26P	37H (BroN)	37H-26P	No inhibition	Novel	Domain of unknown function (DUF4222)	KpnI	HindIII
67P	24H (Gp49)	24H-67P	Partial inhibition	Known	Known	KpnI	HindIII
147P	51H (HigB)	51H-147P	Inhibited	Novel	DNA binding	KpnI	HindIII
39P	51H (HigB)	51H-39P	Inhibited	Novel	DNA binding	KpnI	HindIII

Lyophilised plasmids were rehydrated in nuclease-free water. In order to ensure that *in vitro* validation experiments were performed using a single clone of each synthesised construct, each plasmid was propagated and prepared from a cloning strain of *E. coli*. Briefly, *E. coli* was

cultured aerobically in 100 ml LB broth to an OD_{600} of approximately 0.5 (200 rpm, 37 ˚C). Cells were harvested by centrifugation and resuspended in ice-cold 10 mM calcium chloride ($CaCl_2$) solution. Cells were washed three times in $CaCl_2$ solution, collected by centrifugation, resuspended in 10 mM $CaCl_2$ containing 25% v/v glycerol, and frozen at -80 ˚C. One microlitre of each plasmid solution was used to transform these chemically competent *E. coli* by heat shock (plasmid incubated with bacteria on ice for 30 min, heat shock at 42 ˚C for 30 sec, 5 min immediate recovery on ice). Transformed cells were recovered for one hour at 37 ˚C (200 rpm), and transformants were selected for on solid LB media supplemented with appropriate antibiotics. One colony was picked and single-colony purified; the purified clone was then cultured overnight in 5 ml LB supplemented with antibiotics. Plasmids were extracted from 2 ml of each culture using the QIAprep Spin Miniprep kit (Qiagen, #27104) and the remaining culture was mixed with glycerol (25% v/v final concentration) and stored at -80 ˚C.

3.4 Results

3.4.1 Type II and type IV TA systems are highly abundant in the *K. pneumoniae* species complex

259 *K. pneumoniae* species complex genomes representing the global diversity were included in this study [9] (See Section 3.3.1). These include 222 *K. pneumoniae sensu stricto*, 18 *K. quasipneumoniae* and 19 *K. variicola* isolates (Figure 3.2A), including isolates taken from community and hospital acquired infections, those causing invasive and non-invasive disease and those isolated from both animals and plants [9].

SLING was used to search for TA pairs within our genomic dataset [311]. For clarity, a group of toxins or antitoxins which have been clustered together based on their amino-acid sequence identity are referred to as “toxin group” and “antitoxin group”, respectively. The toxin groups were named by the profile by which they were found.

Using a collection of 55 (52 type II, 3 type IV) Pfam toxin profiles as the input for SLING [311], a total of 140 toxin groups (130 type II, 10 type IV) and 233 antitoxin groups (211 type II, 23 type IV), forming 244 different toxin-antitoxin structures in the genomes included in this study were identified (Appendix B and E). Altogether, TA systems were highly prevalent in all members of the *K. pneumoniae* species complex, with a median of 19 loci per isolate genome (range 11-29, Figure 3.2B). A PCA showed a clear separation into the three species based on toxin repertoire (Figure 3.2C). Furthermore, *K. variicola* had a higher median of 22 TA systems per isolate compared to 18 and 19 in the other two species (Figure 3.2D; pairwise Wilcoxon

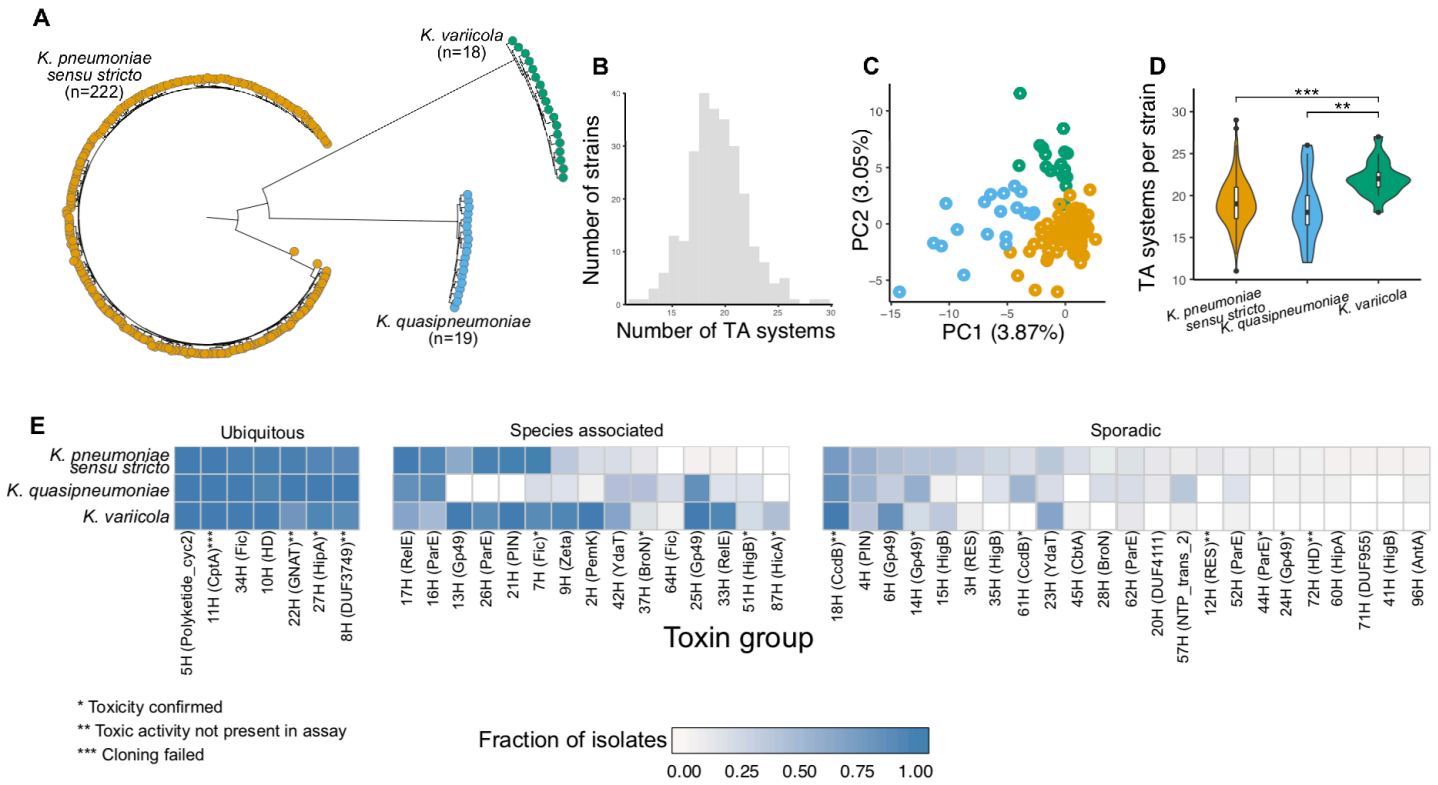

Figure 3.2: Diversity of toxins in* K. pneumoniae *species complex. ***A*** *Core gene phylogeny of the 259 selected* K. pneumoniae *species complex genomes.* ***B*** *Number of predicted TA systems per isolate.* ***C*** *First two principal components of PCA analysis of toxin repertoire coloured by* K. pneumoniae *complex species (yellow:* K. pneumoniae *sensu stricto, blue:* K. quasipneumoniae*, green:* K. variicola*).* ***D*** *Number of predicted TA systems per isolate, stratified by* K. pneumoniae *complex species. Pairwise Wilcoxon rank sum test, *p<0.01, **p<0.001, FDR corrected.* ***E*** *Fraction of isolates from each* K. pneumoniae *complex species possessing each of the toxin groups. Toxin groups are categorised by their distribution patterns (detailed in Appendix B). The toxin Pfam profile used to identify the toxin group is in brackets.*

rank sum test p<0.01, FDR corrected). These figures are slightly higher to those observed in previous studies on TAs in *K. pneumoniae* and *E. coli [335,350]*.

Based on sequence similarity, the number of defined toxin groups per toxin Pfam profile ranged from 1-13 (Figure 3.3). The mean sequence variation within any one toxin group ranged from 68.95-100% local identity at the amino-acid level covering 59.33-100% of the full length of the protein (46.37-100% amino-acid identity over the complete protein) (Appendix B). This highlights the diversity of candidate toxins linked to functionally tested domains that were identified. For instance, the sequences of toxin group 31H containing the HicA domain were aligned to the toxins containing the HicA domain taken from TADB [318,319] (Figure 3.4). While some key residues are conserved throughout, there are considerable variations between the sequences taken from TADB to each other as well as to our predicted toxin.

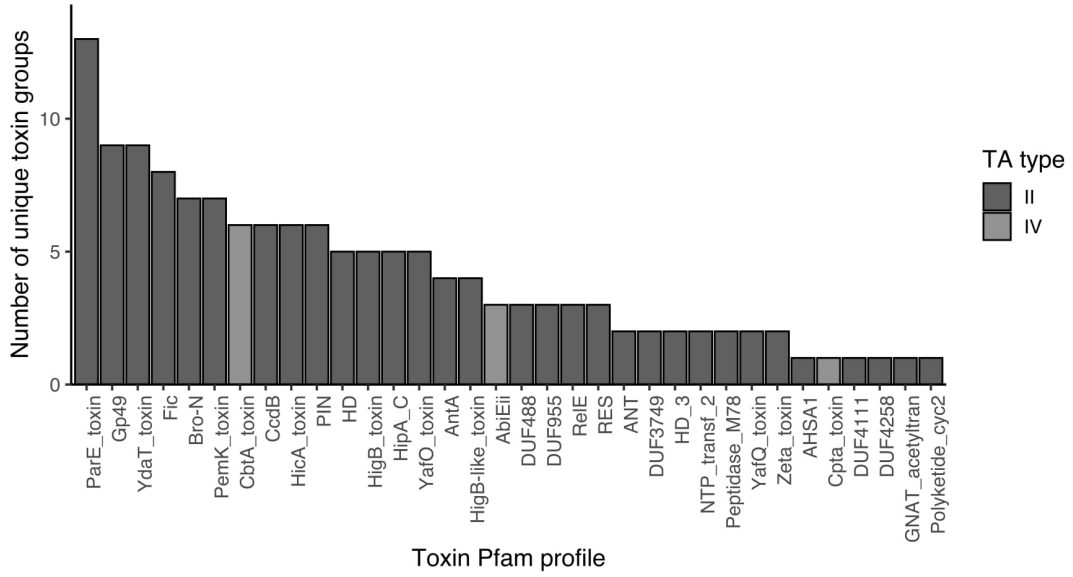

Figure 3.3: Number of unique toxin groups for each of the toxin Pfam profiles used in the search. *Bars are coloured based on the type of TA system the toxin profile is associated with.*

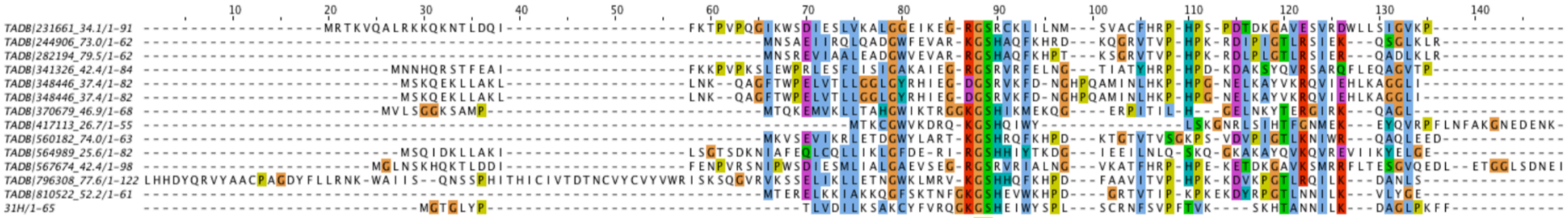

Figure 3.4: Example of diversity of toxins containing a HicA toxin Pfam profile domain. *MSA of all the toxin sequences from TADB containing the HicA toxin Pfam profile, and a representative of toxin group 31H, containing the HicA domain in the* K. pneumoniae *dataset. Alignment was produced using mafft (v7.205) [364]. Image was produced using JalView (v.210) [365].*

3.4.2 Redefining toxins based on their distribution patterns

The 140 identified toxin groups were categorised into four categories based on their distribution patterns in the dataset (See Section 3.3.4) (Figure 3.2E, Appendix B). Seven toxin groups were ubiquitous (one type IV), present in over 80% of the isolates included in this study and from all three species. Fifteen toxin groups, all type II toxins, differed in prevalence between the three species (Fisher's exact test $p<0.01$, FDR corrected, Figure 3.2E). Twenty-three toxin groups (one type IV) (17%) were distributed sporadically with no species association, including a number which were associated with clinically relevant genes. Finally, the remaining 95 toxin groups (eight type IV) (68%) were rare and found in fewer than 10% of the isolates (Appendix B).

Within the ubiquitous toxin groups, we observed significantly higher nucleotide identity for toxins within the same species compared to toxins from other species (median 99.4% compared to 93.51%, Wilcoxon rank sum test, $p<0.001$, Figure 3.5). The median nucleotide identity for sporadic toxin groups for toxins within a species was 97.06% compared to 96.57% between species. This elucidates the evolution of the ubiquitous toxin groups due to genetic drift within a specific member of the species complex, compared to the likely mobile, sporadic toxin groups where this effect was not observed.

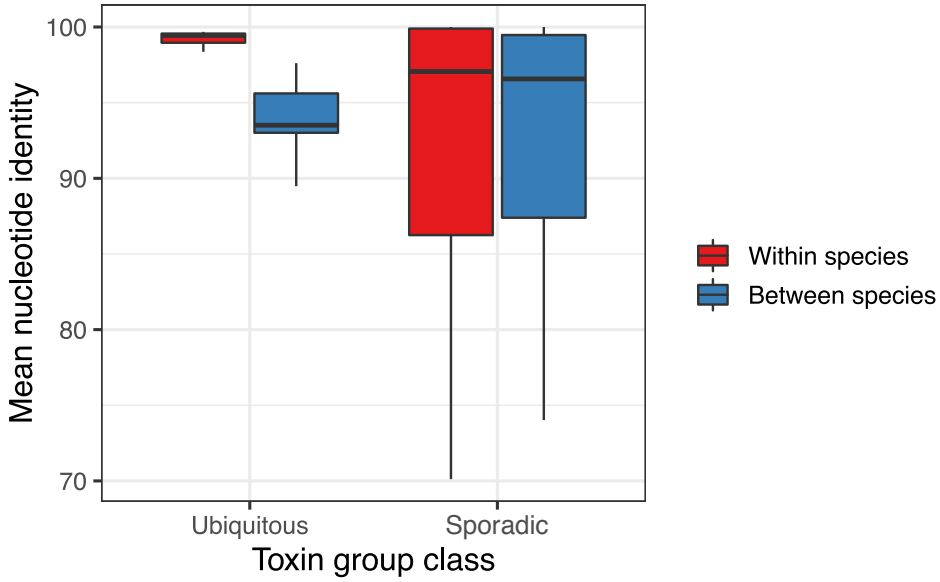

Figure 3.5: Nucleotide identity of toxins within and between species. *Mean nucleotide identity between toxins originating from the same* K. pneumoniae *species and from different* K. pneumoniae *species for all the ubiquitous and sporadic toxin groups.*

The seven ubiquitous toxin groups are known to inhibit translation via mechanisms that do not include RNA cleavage: toxin group 5H (polyketide_cyc) is a homolog of the RatA toxin in *E. coli* which inhibits translation by binding to the 50S ribosomal subunit [338]. Similarly, toxin group 34H (Fic) is a Doc toxin which inhibits translation by phosphorylating and concomitantly inactivating elongation factor TU (EF-Tu) [245]. Toxin groups 22H and 8H with the GNAT and DUF3749 domains are acetyltransferases known to inhibit translation by acetylating aminoacyl-tRNA [366,367]. Group 27H contains a HipA domain which is well described for its association with the high persister phenotype [348,368] and inhibits translation by phosphorylating and concomitantly inactivating glutamyl-tRNA synthetase [369]. Toxin group 11H with the CptA domain belongs to type IV TA system which inhibits cytoskeleton assembly [370]. Finally, group 10H with the HD domain is a phosphohydrolase which is a putative toxin domain from TADB but its exact function is unknown [311,318,319].

The species associated toxin groups presented different distribution patterns across the three *K. pneumoniae* complex species included in this study. *K. pneumoniae sensu stricto* possessed three toxin groups in lower prevalence compared to the other two species (51H (HigB), 64H(Fic) and 25H (Gp49)) (Figure 3.2E). *K. variicola* possessed five toxin groups in higher prevalence compared to *K. pneumoniae sensu stricto* and *K. quasipneumoniae* (42H (YdaT), 9H (Zeta), 2H (PemK), 33H (RelE) and 87H (HicA) domains). Toxin group 87H (HicA) was specific to *K. variicola* and was not observed in the other two species in the dataset. On the other hand, toxin groups 16H (ParE) and 17H (RelE) domains were less common in *K. variicola.* Finally, *K. quasipneumoniae* lacked three toxin groups (21H (PIN), 26H (ParE) and 13H (Gp49)), and rarely possessed toxin group 7H (Fic). On the other hand, toxin group 37H (BroN) was observed in higher prevalence in *K. quasipneumoniae* relative to the other two species. Of these *K. quasipneumoniae* isolates, 11% possessed three copies of this toxin group and 16% possessed two copies (Figure 3.6).

3.4.3 Prediction of novel antitoxins

Accumulation curves of the unique toxin and antitoxin groups identified using SLING suggested that sampling additional *K. pneumoniae* species complex genomes would lead to further identification of new candidate antitoxins (Figure 3.7A). To assess whether the identified antitoxins were known or novel, their sequences were aligned against all type II and type IV antitoxin sequences retrieved from the TADB database [318,319] (See Section 3.3.5). 195 (173 type II, 22 type IV) of the 233 (211 type II, 23 type IV) antitoxins detected in this study were not identified in TADB and were seen to be novel candidate antitoxins linked to a known toxin (Appendix C). For completeness, a predicted function was assigned to the 195 novel

antitoxin groups using interpro-scan (Appendix C) [361]. 19 additional antitoxin groups were matched to known antitoxins by interpro-scan which were not in TADB (antitoxins of toxin profiles YdaT (8), CbtA (4), CcdB (2), Fic (1), PemK (1), PIN (1), HigB (1) and HicA (1)), leading to a final count of 176 novel antitoxins (76%).

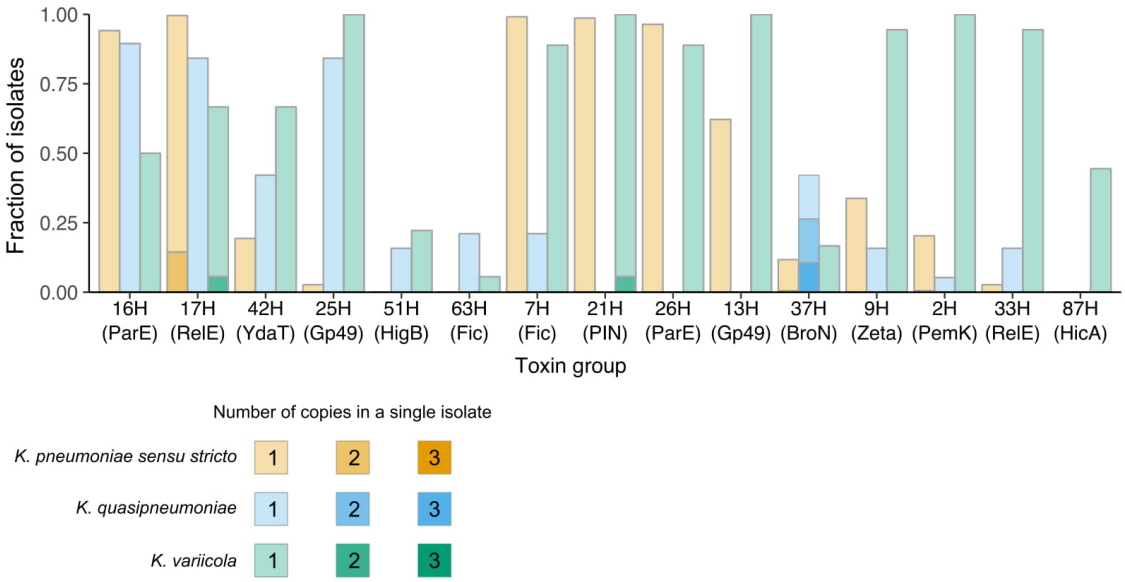

Figure 3.6: Copy number of species-associated toxins. *Fraction of isolates of each of the* K. pneumoniae *complex species possessing each of the species associated toxin groups. Darker shades indicate multiple copies of the toxin group present in an isolate of a species*.

72% of novel antitoxins (127/176) could not be assigned a putative function (Appendix C). Five groups contained one of the toxin profiles used in the toxin search and are the result of disrupted toxins. Twelve groups were predicted to be DNA binding or transcriptional regulators which are plausible functions for antitoxins due to the auto-regulation of the TA operon through conditional cooperativity [346,371]. Another 12 groups were assigned to be intrinsically disordered proteins [372]. The remaining groups contained profiles indicating other functions such as domains of unknown function, ABC transporters, prophages and other functional categories (Appendix C).

For each of the toxin groups, the arrangement of the linked antitoxin was examined: upstream of the toxin (denoted AT-T) or downstream of it (denoted T-AT) (Figure 3.7B). 72% of the known antitoxins were located upstream of the toxin compared to 50% of the novel antitoxins, i.e. novel antitoxin candidates were more commonly located downstream of the toxin relative to the known antitoxins (p = 0.007, Chi squared test).

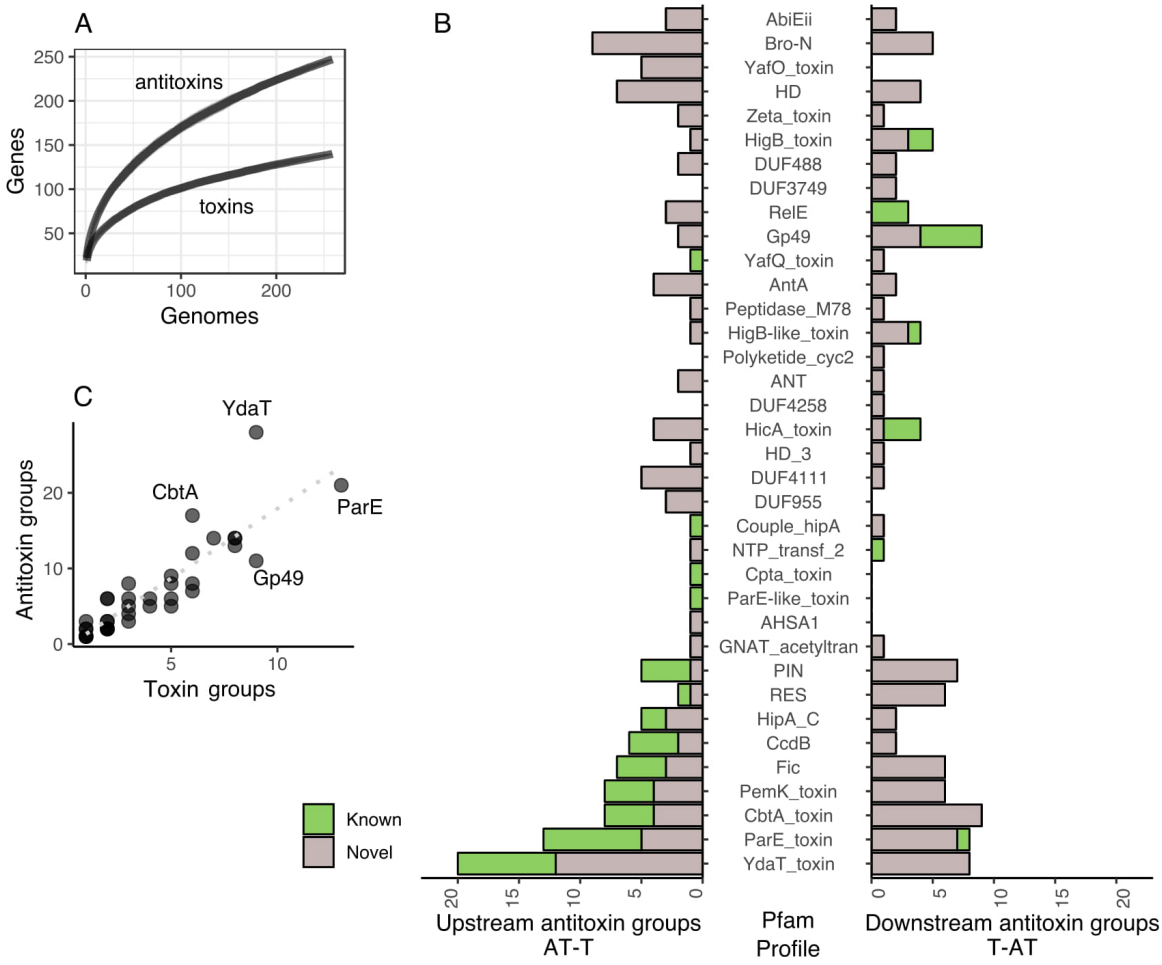

Figure 3.7: Identification of novel antitoxins in the* K. pneumoniae *genomes. A *Accumulation curves of unique toxin and antitoxin groups found in an increasing collection of K. pneumoniae genomes.* ***B*** *Number of antitoxin groups found only upstream (AT-T) and downstream (T-AT) relative to each toxin Pfam profile, coloured by known or novel.* ***C*** *Number of toxin groups of each toxin Pfam profile, relative to the number of antitoxin groups found in their proximity.*

3.4.4 Fluid association and distribution of toxin-antitoxin pairings

Looking at the association between specific toxins and antitoxins we found that with a greater number and diversity of defined toxin groups belonging to the Pfam profile used to search for the toxins, there were concomitantly more antitoxin groups linked to those toxins (0.88 Pearson correlation, 3.7C). The exceptions included the YdaT domain which was found with 28 candidate antitoxin groups and linked to only 9 toxin groups. This both suggests there is coevolution of TA pairs along with instances where a range of different antitoxins can inhibit the same toxin.

We found that a single toxin group can be found with up to a maximum of 12 discrete antitoxins, highlighting the “mix and match” nature of toxin-antitoxin associations [317]. It is important to note that the antitoxin groups are substantially different from each other as a cut off of 75% local amino-acid sequence identity was applied for two antitoxins to be in the same group. Furthermore, the mean sequence variation within any one antitoxin group ranged from 74.64-100% local identity at the amino-acid level covering 61-100% of the alignment length (59.88-100% aa identity over the complete protein), highlighting further the diversity in the candidate antitoxins identified (Appendix C).

In addition to a range of different antitoxins paired to the same toxin, toxins showed a range of operon structures (Figure 3.8A); some toxin groups were linked to a single antitoxin in a conserved position either upstream or downstream of the toxin. Other toxin groups were found in multiple arrangements with the antitoxin sequence and/or location of the antitoxin relative to the toxin changing (Figure 3.8B-H). For the ubiquitous toxin groups, three groups were found in a single arrangement (groups 11H (CptA, a type IV toxin), 5H (polyketide_cyc) and 8H (DUF3749)) (Figure 3.8B). Three other toxin groups (groups 22H (GNAT), 34H(Fic) and 27H(HipA)) were observed in two or three structures often with one structure dominating (>90% of isolates) and the others being rare occurrences of the other structures (<3% of isolates, Figure 3.8C-D). Although the HD toxin group was classified as ubiquitous, one TA arrangement, observed in 80% of isolates, was specific to *K. pneumoniae sensu stricto,* missing in *K. variicola* and replaced by a structure specific to *K. variicola* (Figure 3.8D).

The species-associated toxin group 7H (Fic), was observed in one arrangement which was specific to *K. variicola* (Figure 3.8E). Toxin group 51H (HigB) was associated with two unique antitoxins with one being specific to *K. quasipneumoniae* (Figure 3.8F). Alternatively, other toxin groups possessed multiple operon structures with no clear species association, for instance, toxin group 42H (YdaT) was observed with seven antitoxin groups in eight different arrangements (Figure 3.8G). Other than in a single case (18H (CcdB)), the sporadically distributed toxins were not seen in species-specific arrangements emphasising they are unlikely to be vertically inherited (Figure 3.8H).

Most of the antitoxins identified were toxin group specific. However, antitoxin group 52P was observed with toxin group 31H (HicA) in seven isolates and with toxin group 54H (BroN) in a single isolate. Interestingly, it was always observed upstream to the 31H (HicA) toxin and downstream of 54H (BroN) toxin. The antitoxin proximate to 31H (HicA) shared 83.2% amino acid sequence identity with the antitoxin proximate to 54H (BroN) antitoxin. This antitoxin was

not found in TADB but encodes for a domain of unknown function DUF1902 (PF08972) which is in the same Pfam clan as many other antitoxins (Met_repress, CL0057).

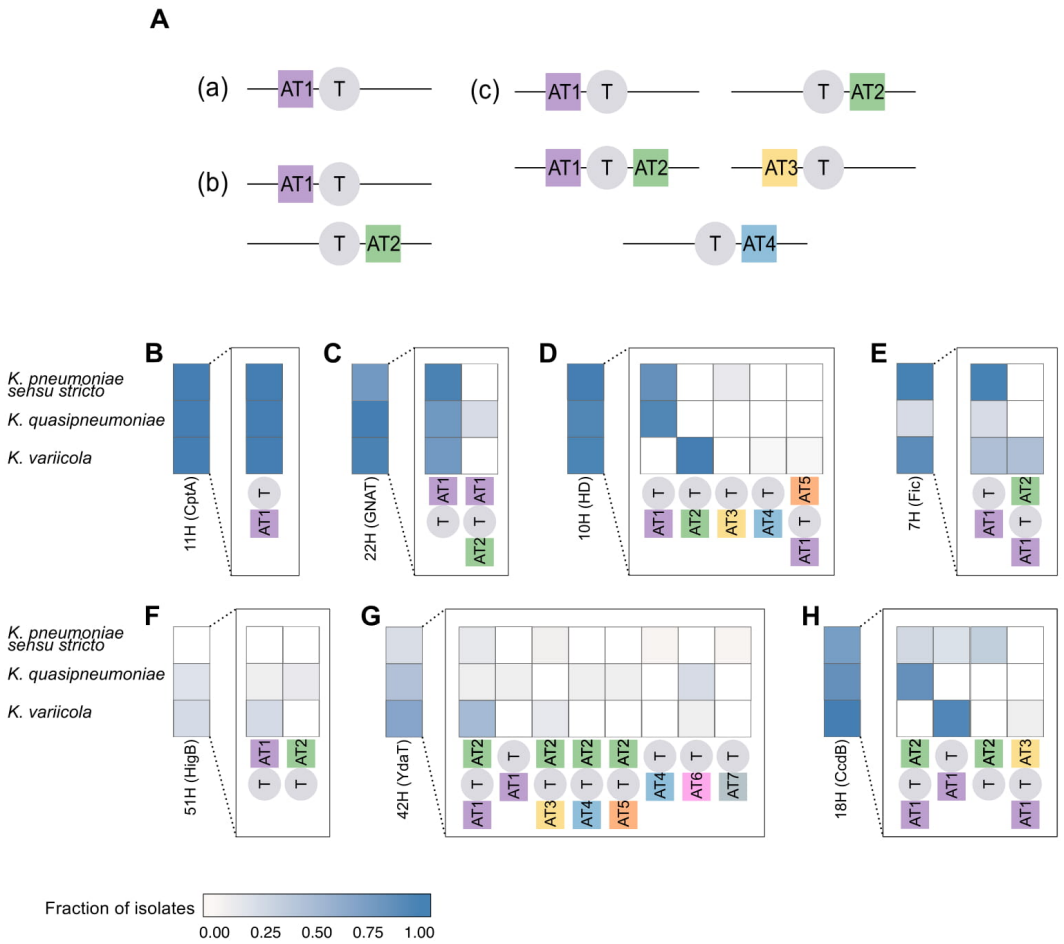

Figure 3.8: Diversity in the observed operon structures for the different toxin categories. A *Examples of range of antitoxins and possible operon structures for a toxin (a) toxin group found in a single structure with a single antitoxin group (b) toxin group found in two different structures with two different antitoxin groups (c) toxin group found in five different structures with four different antitoxin groups.* ***B-H*** *Fraction of isolates from each* K. pneumoniae *complex species possessing each of the operon structures of seven example toxin groups: (**B-D**) ubiquitous, (**E-F**) species associated, (**H**) sporadically distributed.*

3.4.5 Phenotypic testing *in silico* predictions of toxins and confirmation of novel antitoxins[6]

[6] This work was conducted by Cinzia Fino, Matthew Dorman and Alexander Harms.

Due to the apparent diversity of TA systems within and between species and the novel combinations of toxin and antitoxins found in this study, 17 candidate toxins, representing the diversity of toxins within a given group and from a range of genomic backgrounds, were tested for their ability to inhibit bacterial growth in an *Escherichia coli* model system (See Section 3.3.8). Selected were: four ubiquitous, four species associated, seven sporadically distributed and two rare candidate toxins (Table 3.1, Figure 3.9).

The toxicity of all the species associated toxins that were tested was confirmed (groups 51H (HigB), 7H (Fic), 87H (HicA) and 37H (BroN)) (Figure 3.9). Of the remaining toxins, toxicity was observed for the 27H (HipA) toxin group which is ubiquitous across the species complex as well as four of the seven sporadically distributed toxins tested from groups 14H (Gp49), 24H (Gp49), 61H (CcdB), 44H (ParE), and a rare toxin from the 31H (HicA) group. The ubiquitous type IV toxin we tested, 11H ((CptA)), could not be successfully synthesised or cloned, likely due to its toxic activity. The rest of the toxins tested showed no toxic activity under the conditions tested in our assay (summarised in Table 3.1).

Subsequently, 14 candidate antitoxins were tested for their ability to counteract the toxicity of their cognate toxin in the *E. coli* model system (including 10 novel antitoxins; this study; Figure 3.10; Table 3.2). Eight of the fourteen antitoxins (57%) led to complete inhibition of the toxic activity, five of which were novel antitoxins. Three of the confirmed novel antitoxins were predicted to contain DNA binding domains by interpro-scan (39P, 27P, 147P). One antitoxin contained a domain of unknown function (52P) and the final antitoxin did not match any existing entry in Interpro (44P). Three of the confirmed antitoxins in the T-AT format were located downstream of the toxin (groups 27P (Gp49), 147P (HigB) and 39P (HigB)). An additional known antitoxin only partially inhibited toxicity (67P).

For completeness, for operons that had the structure AT1-T-AT2, both AT1 and AT2 were tested. In both cases, AT1 only was confirmed to inhibit the toxin's activity while we did not observe toxin inhibition activity with AT2.

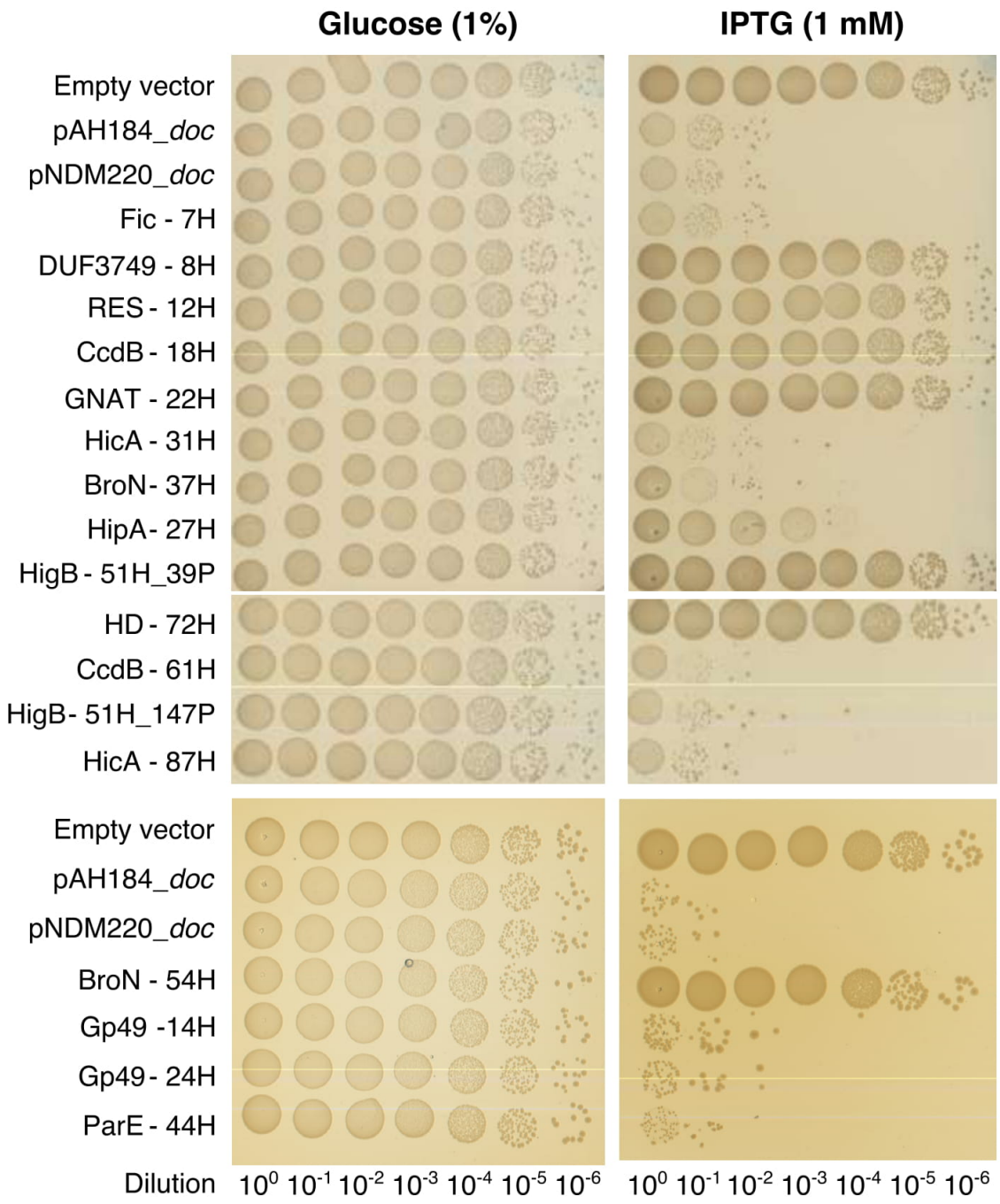

Figure 3.9: Phenotypic testing of selected toxins. *This work was conducted by Cinzia Fino, Matthew Dorman and Alexander Harms. LB agar plates were supplemented with 1mM IPTG for the induction of toxin Plac promoters. Overnight cultures were serially diluted (10^{-1} to 10^{-6}) in PBS containing the inducing supplements.*

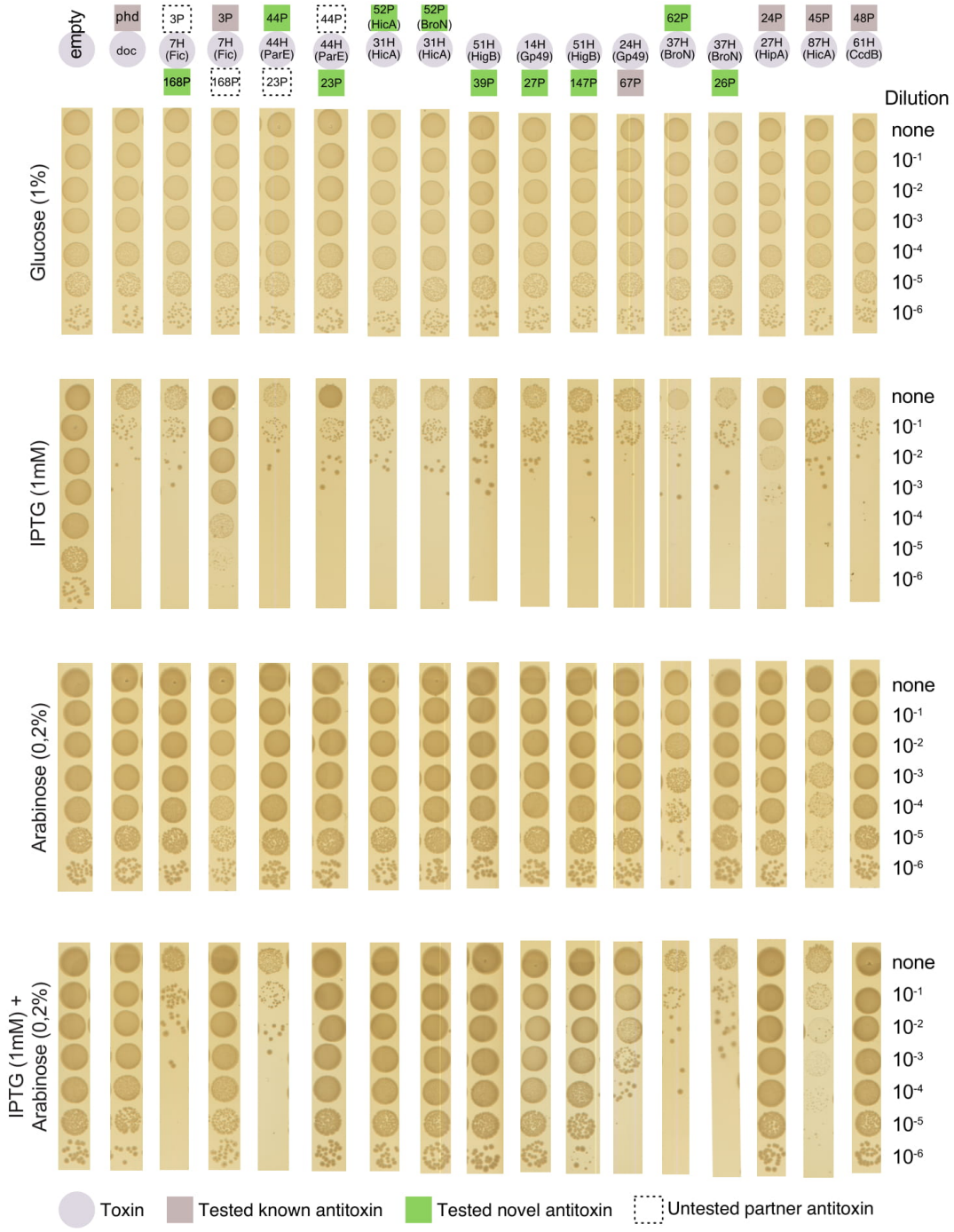

Figure 3.10: Phenotypic testing of predicted toxin-antitoxin combinations. *This work was conducted by Cinzia Fino, Matthew Dorman and Alexander Harms. Toxins in circles, antitoxins in squares. Tested novel antitoxins in green and tested known antitoxins in gray. For operon structures AT1-T-AT2, the untested partner antitoxin is in a dashed square. LB agar plates were supplemented with 1 mM IPTG for the induction of toxin Plac promoters' and 0.2% w/v of L-arabinose for the induction of antitoxin Para promoters'. Overnight cultures were serially diluted (10^{-1} to 10^{-6}) in PBS.*

Finally, these data revealed some more unexpected findings. In two cases the predicted antitoxins were themselves found to be toxic in our experimental system (45P, 62P) (Figure 3.10). One of these antitoxins is a well-described antitoxin with a HicB domain (62P). In addition, we confirmed both versions of antitoxin group 52P, associated with toxins from markedly different groups (31H (HicA) and 54H (BroN)), were able to counter toxin group 31H (Figures 3.10, 3.9; Table 3.2). Although the antitoxin group was linked to two different toxins and the two versions of the antitoxin shared only 83.2% amino acid identity, both versions inhibited the activity of this toxin. We were unable to confirm the toxicity of toxin group 54H (BroN) (Figure 3.9, Table 3.1), hence we could not confirm inhibition of this toxin group by these antitoxins. Finally, two variants of the toxin group 51H were tested (HigB); a shorter protein (53 aa) which was observed with antitoxin group 39P and a longer protein (103aa) observed with antitoxin group 147P. The C-terminus of the longer toxins was 83% identical to the shorter protein. The two antitoxins shared 71% amino-acid identity. We were only able to confirm the toxicity of the shorter 51H toxin. Nonetheless, we tested both antitoxins 39P and 147P with the shorter 51H toxin, and found that both antitoxins were functional and able to inhibit the toxin (Figure 3.10).

3.4.6 Orphan antitoxins are abundant in the dataset

We sought to determine whether the antitoxins of the TA pairs were also present on the *K. pneumoniae* species complex genomes as orphan genes uncoupled to a candidate toxin gene. The predicted antitoxin sequences were aligned against all the genomes and a total of 2,253 occurrences of orphan antitoxins belonging to 105 of the 233 antitoxin groups defined in this study were identified in the genomes (96 type II and 9 type IV) (Figure 3.11A, Appendix D). Of these, 25% were known antitoxins found in TADB or Interpro (26/105). For 80% (77/96) of type II and 89% (8/9) of type IV antitoxin groups, fewer than 26 orphan copies were identified in the entire genome collection, i.e. occurrences of unpaired antitoxins were rare and were found in fewer than 10% of genomes (Figure 3.11A). Conversely two antitoxin groups, containing the type II Fic and HipA toxin domains, were observed as unpaired in more than 80% of the genomes (>207 copies) across the species complex. In 35 of the 105 orphan antitoxin groups, orphans were detected in a species that was different to that of the original valid TA pair (Appendix D). For instance, antitoxin group 89P of the HipA toxin was originally identified in *K. quasipneumoniae*. However, orphan antitoxins were observed only in *K. variicola* (Figure 3.11A). Similarly, antitoxin group 115P belonging to a PemK-containing toxin was originally identified in *K. variicola*, but orphan antitoxins were observed in *K. quasipneumoniae* as well. Altogether there were no significant differences in the number of

orphan antitoxins per strain between the three species, with a median of nine orphans per strain across the three species (Figure 3.11B) (pairwise Wilcoxon rank sum test, FDR corrected).

Figure 3.11: Orphan antitoxins in K. pneumoniae genomes. *A Number of orphan antitoxins identified from each antitoxin group, coloured by K. pneumoniae complex species. The toxin Pfam profile of the toxin of the valid TA pair is in brackets. Antitoxin of type IV toxins are highlighted.* ***B*** *Orphan antitoxins per strain stratified by K. pneumoniae complex species.* ***C*** *Illustration of context analysis applied to each orphan antitoxin. The flanking sequences around each orphan antitoxin were compared to the flanking sequences of the valid TA pair. Each flank was classified according to whether or not it matched the sequence of the original TA pair.* ***D*** *Number of occurrences of orphan antitoxins classified by the similarity of their contexts to the valid TA pairs' .* ***E*** *Presence of a CDS in the orphan antitoxin's toxin-context.*

To assess the origin of orphan antitoxins, the upstream and downstream sequence surrounding the antitoxin were aligned with those found in valid TA pairs (Figure 3.11C) (See Section 3.3.6). 39% of the orphan antitoxins (879/2,253) shared the same toxin-context as the valid TA pair. Of these, 92% also shared the same non-toxin-context, indicating that they are in the same genetic context as the valid TA pairs from the same group (Figure 3.11D). 65% of orphans which did not share the toxin-context of the original TA pair (893/1374) did share the non-toxin context. In 20% of cases (470/2,253) neither the toxin-context or the non-toxin-context matched the valid TA pair, i.e. the orphan antitoxins were surrounded up- and downstream by unrelated sequences to any of the detected TA pairs.

To confirm whether these were truly orphan antitoxins, a CDS within the toxin-context was searched for that could function as the toxin. In 49% of orphans (1,107/2,253) a CDS within the context region was identified that does not contain a known toxin domain and could be a candidate for a novel toxin (Figure 3.11E). In 43% of cases (947/2,253) a toxin containing the original Pfam profile used in the search was found but the CDS was discarded due to the conservative structural requirements applied for a TA system (Figure 3.11E). These may be false negatives in the original analysis, or otherwise TAs which have diverged from the expected structure for a functional TA pair. In 8% of cases (171/2,253) the predicted antitoxin was truly orphan as a CDS longer than 50 aa could not be identified in the context region that may function as a toxin. In 1% of cases (28/2,253), the orphan antitoxin was close to the contig edge or proximate to a region with more than eight unknown nucleotides (N/X) and therefore the presence or absence of a toxin in its proximity could not be confirmed.

3.4.7 The association between toxins and antimicrobial resistance genes, virulence genes or plasmid replicons

Several of the sporadically distributed toxin groups were associated with clinically relevant AMR or virulence genes as well as plasmid replicons linked to the spread of AMR in *K. pneumoniae* and *E. coli* (Figures 3.12A,B, Fisher's exact test $p<0.01$, FDR corrected). These included 24H (Gp49) and 72H (HD) toxin groups which were significantly associated with multiple AMR genes, including those conferring resistance to aminoglycoside, amphenicol, sulfonamide, tetracycline and beta-lactams, with 13-29% of toxin genes found on the same contig as the respective AMR genes (Figure 3.12C). 100% and 30% of toxin CDSs' of toxin groups 24H and 72H respectively were on the same contig with an IncA/C plasmid replicon (Figure 3.12D). These contigs shared 99% (24H) and 97% (72H) sequence identity with the *K. pneumoniae* IncA/C-LS6 plasmid (JX442976), originally isolated from carbapenem-resistant *K. pneumoniae* [373], as well as AMR plasmids pNDM-KN (24H), pRMH760, pIMP-

PH114 and pR55 (72H) (Appendix D) [374–377]. Two toxin groups with a RES domain, 3H and 12H, were associated with multiple virulence genes (Figure 3.12B, Fisher's exact test p<0.01, FDR corrected) and one of these groups (3H) with the presence of an IncHI1B plasmid replicon. Contigs containing these two toxins showed over 99% sequence identity to *K. pneumoniae* virulence plasmids pK2044 and pLVPK (Appendix D) [378,379]. Five other toxin groups which were associated with AMR or virulence genes were also associated with the presence of plasmid replicons (Fisher's exact test p<0.01, FDR corrected) (see Figures 3.12A-C).

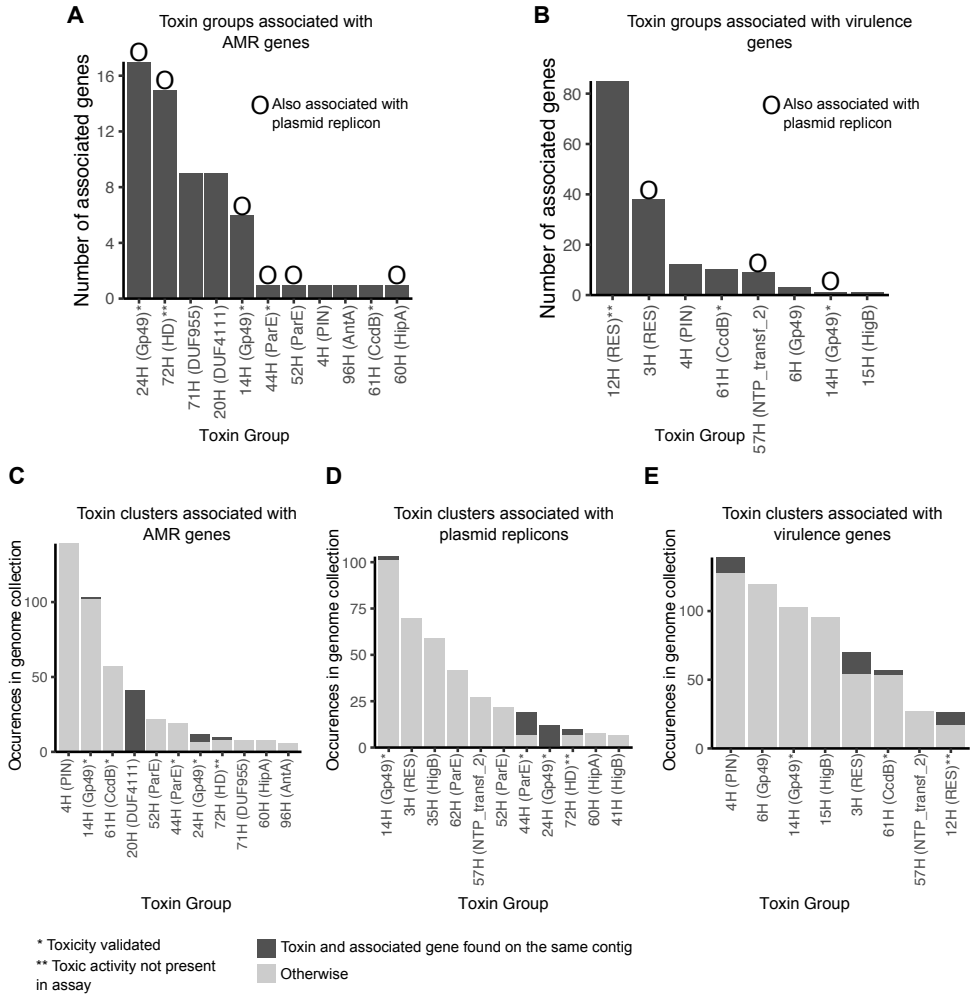

Figure 3.12: Toxin groups associated with AMR genes, virulence genes and plasmid replicons. *Number of unique AMR **(A)** and virulence **(B)** genes associated with each of the toxin groups. Circles above bars indicate the toxin group was also associated with the presence of a plasmid replicon. **C-E** Number of occurrences of toxins in the genome collection, for the toxin groups associated with AMR genes (**C**), plasmid replicons (**D**) and virulence genes (**E**). An occurrence of a toxin is coloured in dark if it was observed on the same contig with one or more of the associated genes, light otherwise.*

3.5 Discussion

In this chapter was presented a systematic in-depth analysis of the diversity and evolution of TA systems in a large collection of a clinically important member of the *Enterobacteriaceae,* the *K. pneumoniae* species complex. TA systems are highly prevalent in the species complex*,* however, the underlying processes of the evolution of TA systems are likely to be context-dependent. The toxins of these TA systems could be classified based on their distribution patterns as ubiquitous, species associated, sporadically distributed (often with associations to clinically important genes) or rare. The evolution of ubiquitous toxins is likely vertically inherited, as higher nucleotide identity was observed between toxins of the same species than between species. The same effect was not observed for the sporadic toxins, suggesting that some TA systems are more mobile than others. Importantly, the classification presented in this study was based on the dataset used, which was aimed to capture the diversity of the *K. pneumoniae* species complex*.* It is possible that further sampling of under-represented lineages would increase power and refine the classification.

The pairing of antitoxin to toxin is not fixed; for each toxin a range of candidate antitoxins were identified and found in different arrangements, putatively able to inhibit the same toxin. Sampling of more genomes would lead to a large diversity in antitoxins relative to toxins, suggesting the potential number of interactions between toxins and antitoxins is large. Notably, some toxins were more stably coupled to a single antitoxin and observed in a single arrangement, while other toxins were observed with a wide range of antitoxins and operon arrangements. This highlights that the co-evolution between toxin and antitoxin is dependent on the system and context. This has functional implications as the antitoxin and its interaction with the toxin can affect the functioning of the TA system ([380]. Some antitoxins play a role in the regulation of the TA module as the toxin-antitoxin pair regulate the expression of the TA operon [346,371]. Furthermore, the interaction of the toxin with the antitoxin will determine the specificity of the inhibition and therefore would affect the dynamics of both activation and deactivation of the TA operon. Finally, antitoxin instability is often the result of degradation by proteases [346], therefore the inhibition of an antitoxin in response to stress can depend on the antitoxin sequence as it would determine the specificity of interaction with proteins that lead to its degradation [381].

Even more, a number of toxin or antitoxin groups were observed as specific to a species, i.e. a toxin-antitoxin pairing was observed only in one particular genetic background. This suggests it may be beneficial to possess a specific toxin-antitoxin pair under one genetic background compared to another.

Altogether 76% of the identified candidate antitoxins were novel and not identified in the existing toxin-antitoxin database TADB or Interpro [318,319,382]. Furthermore, there was additional sequence diversity within each antitoxin group that we found. These results emphasise the potential large diversity of antitoxins that could inhibit these toxins and our lack of knowledge of the complete range and diversity of these systems.

Using an *E. coli* model system, the toxicity of 10 of 17 tested toxins was confirmed (~59%) and the inhibitory activity of 10 of 14 tested antitoxins (~71%). Nine of the tested antitoxins are novel and we were able to confirm the inhibition of five of them. We also found candidate antitoxins downstream of the toxin, and confirmed the inhibitory activity of three of them, highlighting exceptions to the common setup in which the antitoxin is encoded upstream of the toxin. These results could form the basis of future studies investigating how different autoregulatory principles enabled by upstream or downstream antitoxins might affect the biology of a TA system. While some of these candidate antitoxins could be false predictions, the observation of known or confirmed antitoxins both upstream and downstream to toxins suggests we cannot rule out any antitoxin candidate. Importantly, a negative result in our assays does not rule out toxic or inhibitory activity of these proteins, but rather could be the result of confounding effects in our assays for example biological differences between *E. coli* K-12 and *K. pneumoniae*, lack of protein expression or incorrect folding in the heterologous host. Furthermore, our assays do not indicate whether these systems are expressed in the host bacterium or whether they have a physiological role in the host cell.

There is an abundance of orphan antitoxins present in the population which are unpaired to a functional toxin. These include a number of the antitoxins we expressed and were able to confirm their inhibitory activity (92 orphan copies of 39P, 17 orphan copies of 24P and 45 orphan copies of 45P, Figure 3.11). Sources of orphan antitoxins may be degrading TA pairs that are in different genetic locations, older degraded TA systems or otherwise, these could be candidates for new toxins which share the same antitoxin as we have identified. Alternatively, some orphan antitoxins may be paired to a known toxin but were discarded in our analysis due to the conservative structural criteria we defined for a TA system, suggesting that the prevalence of TA system in the *K. pneumoniae* species complex presented here may be under-estimated.

These orphan antitoxins may be serving a new purpose. For example, they may serve as anti-addiction modules, preventing the fixation of plasmids or other MGEs [383]. They may be interacting with the toxins of active TA pairs and affecting their function. Alternatively, they

could also be conserved as remnants of a degraded TA locus that have acquired functions in transcriptional regulation of other genes in the genome [384].

The importance of this type of analysis is not limited to TA systems, and presents general trends to distinguish between groups of genes of other gene systems. Pan-genome analysis of bacterial datasets is often focused on the description of core compared to accessory genes without focusing on the precise details within these two categories. Here we showed a more refined description of genes based on their distribution across the *K. pneumoniae* population and in the context of linkage to other genes. This finer grained analysis can be applied in other settings and lead to novel, highly relevant insights on evolutionary dynamics of poorly understood genetic elements.

4 Building a collection of 10,000 *E. coli* isolates and defining the gene content in the collection

4.1 Introduction

As of today, there are more than 130,000 *E. coli* and *Shigella* genomes available on public databases. Indeed, recent studies have utilised the availability of these genomes to better understand the population structure and the pan-genome of the species [92,93]. The analysis presented in the Chapter 3 revealed interesting patterns regarding the distribution of a single genetic system in a collection of 259 *K. pneumoniae* genomes. The next two chapters will expand on the analysis presented on TA systems in *K. pneumoniae,* to investigate the distribution of all genes in a collection of 10,000 *E. coli* isolates taken from public databases.

While genomic data is widely available online, the process of building a comprehensive and high-quality collection of genomes is not trivial. The genomic data is stored across different databases which are associated with specific data types. The Sequence Read Archive (SRA) is the main repository which contains all the sequence read data worldwide, and is a collaboration between three read archives worldwide (European Nucleotide Archive; ENA, National Center for Biotechnology Information; NCBI and the DNA Data Bank of Japan; DDBJ) [385]. In some cases, the raw read data is not submitted but only an assembled genome. In these cases, the data will be found elsewhere, for instance, in the NCBI Assembly database. Even more, specific databases have been set up for particular purposes [93,386]. Enterobase, mentioned in Section 1.4, is a database which integrates, assembles and analyses the genomic data of specific enteric pathogens from the SRA, while providing researchers with relevant metadata and software to make these data more accessible [93]. Importantly, when collating the data from these multiple sources, genomes are often duplicated or there are database specific identifiers which need to be matched. Finally, the metadata associated with each genome is often restricted to the publication and is not directly linked to the database from which the genome was downloaded. All of these make the primary process of collating the data challenging.

Following data collation, multiple steps need to be applied to obtain a high-quality collection of genomes and their genes. This includes applying quality control (QC) measures on the

downloaded reads to ensure they are of good quality and that there was no contamination. Enterobase, for example, applies its own QC pipeline before importing data from the SRA and after assembly [93]. The reads need to be assembled and annotated for their gene content. Finally, a pan-genome analysis is applied to obtain the gene content across multiple genomes (detailed in Section 1.4.1.4). The most widely used tools for genome assembly, annotation and pan-genome analysis were published anywhere from five to twelve years ago [292,305,356,387]. As the number of genomes has grown exponentially (Figure 1.7), the most commonly used tools can become obsolete as they do not scale well for a very large number of genomes. For instance, a pan-genome genome analysis requires an all-against-all comparison of the CDSs across all isolates being compared. In an analysis of 10,000 isolates, each with 5,000 genes, this would require 1.25 quadrillion pairwise comparisons. For this reason, some (but not all) existing pan-genome analysis tools use an initial step to remove redundant sequences [305,388]. Even so, with a very large dataset of a diverse organism like *E. coli*, the number of unique sequences is large enough that removing redundant sequences does not solve the complexity issues. Therefore, existing studies using very large datasets have compromised on the level of resolution of the analysis applied and were generally limited to high-level descriptive studies with few downstream analyses [92,93].

4.2 Aims

The aim of this Chapter was to build a comprehensive and high-quality collection of *E. coli* and *Shigella* isolates taken from public databases. The work in the Chapter is divided into the following steps which were required to obtain a complete collection of 10,000 *E. coli* isolates and their gene content:

- The data collection process
- The characteristics of the dataset including associated metadata, population structure and AMR and virulence profiles.
- Definition of the gene content across this collection

4.3 Methods

4.3.1 Data collection

The data collection process for this project is summarised in Figure 4.1 and is detailed in the Results section, including specific modifications to the tools used and all the QC measures applied. All scripts for downloading and processing the genomes are available at

https://github.com/ghoresh11/ecoli_genome_collection. The final collection of genomes consisted of 10,159 presumptive *E. coli* and *Shigella* genomes.

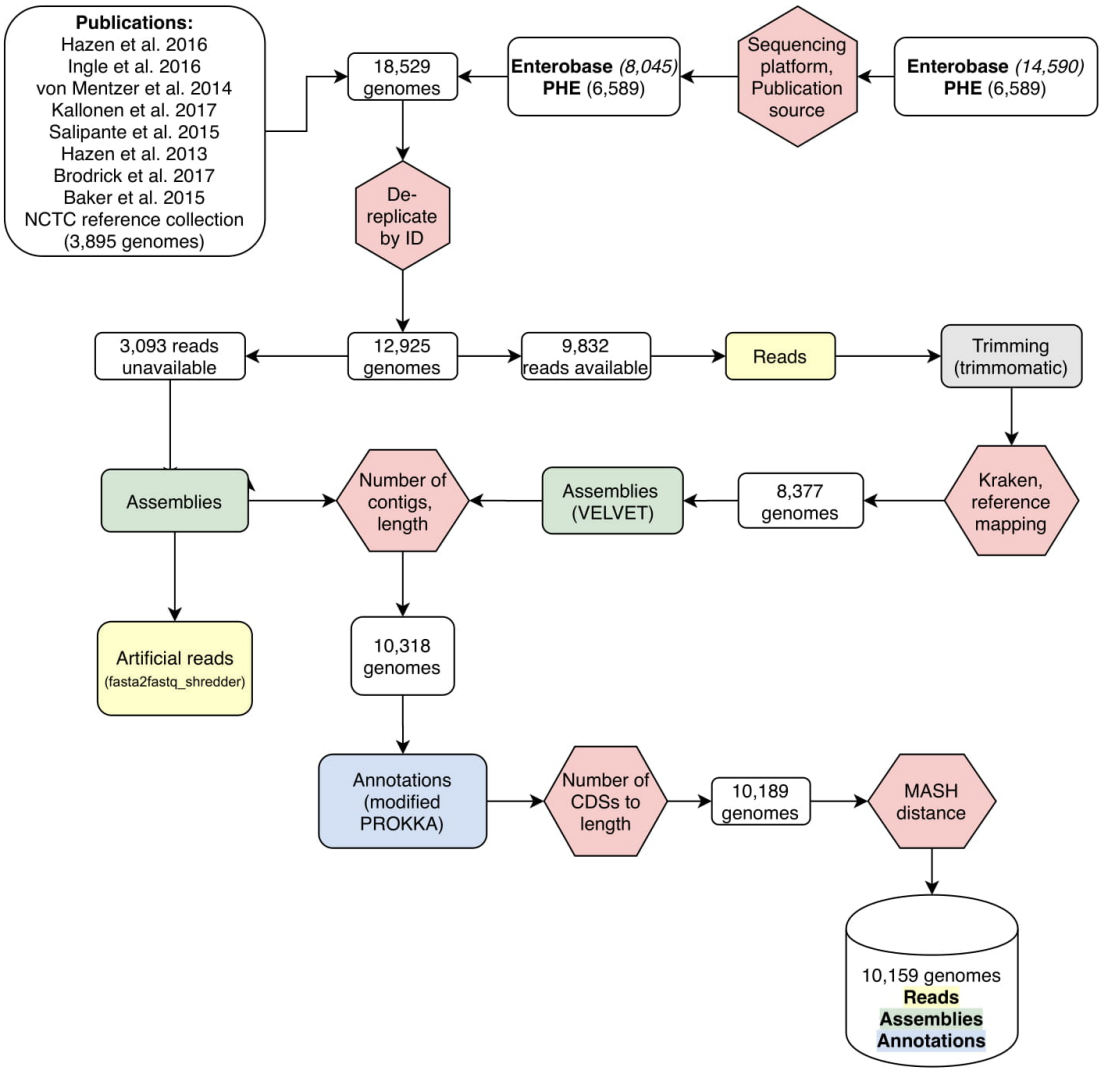

***Figure 4.1: Workflow for collating the* E. coli *genome collection**. Steps taken to obtain a final curated, comprehensive and high-quality collection of genomes which include, for all genomes, reads, assemblies and annotation files. QC steps are in red hexagons.*

4.3.1.1 Reads

Reads were downloaded from the SRA using fastq-dump (v2.9.2). Reads which had been Illumina sequenced were trimmed using trimmomatic (v0.33) [389] with the *TruSeq3-PE-2* adaptors, a minimum length of 36 bp, and parameters LEADING=10, TRAILING=10, SLIDING WINDOW=4:15 and quality encoding Phred33. When reads were unavailable, assemblies were shredded into artificial reads (fasta2fastq_shredder.py) with 100bp paired reads from a 350bp insert every 3 bases along a linear genome.

4.3.1.2 Assemblies

Reads were assembled by VELVET (v1.2.09) [356] using the prokaryotic assembly pipeline (v2.0.1) with default setting [357].

4.3.1.3 Gene calling

Predicted CDSs, referred to as "genes" were called using a modified version of Prokka (v1.5). Prodigal (v2.6) was trained using a random selected set of 100 genomes from the entire dataset using the "prodigal.py" script available in Panaroo [292,306]. The training file was then used as the input for Prokka for the predicted genes in the entire dataset. This was compared against running Prokka without using a training file for all genomes. Panaroo was used to compare the gene content of two annotation files by building a synteny graph of the genes [306].

4.3.2 MLST

The ST of all genomes was determined by running "mlst_check" (https://github.com/sanger-pathogens/mlst_check) according to the Achtman MLST scheme downloaded from PubMLST on Jan 22nd, 2019 [390].

4.3.3 Genome Clustering using PopPUNK

Population Partitioning Using Nucleotide K-mers (PopPUNK) (v. 1.1.3) was used to group the assemblies into PopPUNK Clusters [277]. PopPUNK uses Mash to calculate the pairwise distance between every two assemblies. Mash estimates the Jaccard distance between two sequences using a reduced set of k-mers of a defined size *k* [279]. PopPUNK applies Mash with increasing values of *k*. The "core" (π) and "accessory" (*a*) distances between two assemblies are estimated in PopPUNK by fitting a function which measures the probability of any two sequences matching between the two assemblies across the increasing values of *k* used for Mash (the function: $p_{match} = (1-a)(1-\pi)^k$). The "core" and "accessory" distances were inferred in this analysis using the *k* values 18, 21, 24, 27 and 31 as these values generated a good fit. Following the distance calculation, the pairwise "core" and "accessory" distances were fitted into clusters using two-dimensional Gaussian mixture models to split the points into K two-dimensional Gaussian distributions and to identify the "core" and "accessory" distance values which represent isolates belonging to the same "strain" or "lineage". The model fitting was applied using six different values of K (5, 8, 11, 14, 17 and 20). The scores generated by PopPUNK for all values of K were compared and these are summarised in Table 4.1. The value of K=11 was chosen for the clustering as it had the overall lowest entropy and comparably high overall score. A network between all assemblies is constructed where each

node is an assembly and an edge is drawn between two assemblies only if their “core” and “accessory” distance is within the “within strain” cluster in the result of the two-dimensional Gaussian mixture models. Each connected component in this network is defined as a “PopPUNK Cluster”.

Table 4.1: PopPUNK Clustering statistics. Statistics retrieved from clustering genomes using different values of K when running PopPUNK. Green: The chosen value of K with the lowest entropy.

K	Components	Density	Transitivity	Score	Entropy
5	920	0.1444	0.9929	0.8496	0.0082
8	1120	0.1405	0.9852	0.8467	0.009
11	1185	0.139	0.982	0.8455	0.0042
14	1918	0.1	0.8973	0.8075	0.0055
17	1856	0.1048	0.9093	0.814	0.0053
20	3361	0.0208	0.6273	0.6143	0.0138

4.3.4 Phylogenetic analysis

The core gene phylogeny was inferred from the core gene alignment generated using Roary for each PopPUNK Cluster [305], and a tree from the SNPs, extracted using SNP-sites [332] (v2.3.2), was constructed using FastTree [391]. Treemer (v0.3) [392] was used to select ten genomes from each PopPUNK cluster as representatives of that cluster and representative of the diversity within that cluster. Treemer greedily prunes leaves off the phylogeny by choosing a random lead from the closest pair of leaves in every iteration, until the number of selected leaves in the tree is reached. Similarly, only a single representative sequence was chosen using Treemer from each of the 50 PopPUNK clusters to generate a minimal tree containing only 50 sequences. In both cases, the core gene phylogeny was inferred from a core gene alignment generated using Roary on the 500 representative genomes [305]. A maximum likelihood tree from the informative SNPs, chosen using SNP-sites (v2.3.2) [332], was constructed using RAxML (v8.2.8) [282] with 100 bootstrap replicates.

4.3.5 Phylogroup assignment

EzClermont (v0.4.5) was used to assign the phylogroup of the 500 representative genomes selected in the previous section [393]. EzClermont applies *in-silico* PCR of marker genes to assign phylogroup according to the phylotyping scheme presented in [271]. PopPUNK clusters were assigned a phylogroup according to the most common phylogroup assignment of the ten representative strains. Phylogroup assignments were corrected based on the phylogeny.

4.3.6 Identification of AMR and virulence genes

A collection AMR genes were obtained from the modified version of ARG-ANNOT available on the SRST2 website (https://github.com/katholt/srst2/tree/master/data, downloaded on 08.03.18) [288,290]. Virulence factors were downloaded from the Virulence Finder Database (https://bitbucket.org/genomicepidemiology/virulencefinder_db/src, downloaded 24/08/18). Read files of genomes (real and artificial) were searched for the presence or absence of genes against the downloaded databases using ARIBA (v2.14) with default settings [283]. A gene was marked as present only if 80% of the database entry was covered, otherwise it was marked as absent.

4.3.7 Pathotype assignments

Each isolate was assigned a pathotype according to the presence and absence of specific virulence genes, as well as the source of isolation (Figure 1.4). If the source of isolation was either “blood” or “urine” it was assigned to “ExPEC”. If any variant of shiga-toxin was present it was assigned to “STEC”. If *eae* was present it was assigned to aEPEC/EPEC. If both shiga-toxin and *eae* were present it was assigned to “EHEC”. If either *aatA*, *aggR* or *aaiC* were present it was assigned EAEC. If *est* or *elt* were present it was assigned to ETEC. If *ipaH9.8* or *ipaD*, characteristic of the invasive virulence plasmid pINV, were present it was assigned to EIEC. A pathotype was assigned to a PopPUNK Cluster if at least half of the isolates of the cluster were assigned to the same pathotype.

4.3.8 Pan-genome analysis

4.3.8.1 Pan-genome analysis on each PopPUNK cluster

A pan-genome analysis using Roary [305] was applied on each PopPUNK Cluster separately using the default identity cut-off of 0.95 with paralog splitting disabled [305]. The gene accumulation curves were generated using the *specaccum* function in the vegan (v2.5.6) library with 100 random permutations [359].

4.3.8.2 Combining the pan-genomes of all PopPUNK Clusters

The outputs of the pan-genome analysis of each PopPUNK Cluster were combined to generate a final collection of gene clusters of the entire dataset according in the following steps:

1. Gene cluster definitions, from the Roary analysis within each PopPUNK cluster, were assumed to be the best approximation of the representation of the genes that are well-defined within a closely related group of genomes. Note that each gene cluster has multiple members, i.e. sequences (Figure 4.2, Step 1). A representative sequence was chosen for each gene cluster as the sequence that had the most common length within that gene cluster (the modal length). If there was no mode, a sequence with the median length was chosen.
2. A pan-genome analysis using Roary was applied on all PopPUNK Clusters in a pairwise manner using an identity threshold of 0.95 and with paralog splitting disabled. Namely, a pan-genome analysis was conducted including all genomes of PopPUNK Clusters 1 and 2, 1 and 3, 1 and 4 etc, leading to a total of 1,081 Roary analyses (47 choose 2). This generated gene clusterings for all pairs of PopPUNK Clusters. Note that each gene cluster in the combined Roary analysis had multiple sequences from both PopPUNK Clusters (Figure 4.2, Step 2).
3. A graph was constructed such that each node was one gene cluster from the original Roary outputs from Step 1, named the “combined Roary graph” (Figure 4.2, Step 3).
4. An edge was drawn between a gene cluster of PopPUNK Cluster “A” to a gene cluster of PopPUNK Cluster “B” if there was a gene clustering in the combined Roary analysis such that 80% of the sequences of the gene cluster of “A” were in the new combined clustering and 80% of the members of the gene cluster of “B” were also in the combined clustering (Figure 4.2, Step 4).
5. The following corrections were applied to remove likely incorrect connections between gene clusters in the combined Roary graph (Figure 4.2, Step 5):
    1. Density based clustering was applied on each connected component of the combined Roary graph using the Jaccard similarity between every two nodes with the `dbscan` method of the python package sci-kit learn[394] with parameters epsilon=0.5 and min_samples=6. Edges between a gene cluster of PopPUNK Cluster A and a gene cluster of PopPUNK Cluster B that do not belong to the same dbscan cluster were removed.
    2. A nucleotide MSA using mafft (v7.310)[364] with default settings was applied to all representative sequences of each gene cluster in a connected component of the combined Roary graph. If the alignment of two representative sequences

had more than 20% mismatches along the length of the longer sequence, the edge between them in the combined Roary graph was removed.

6. To correct for over splitting, the representative sequences of all the gene clusters of the original Roary outputs were aligned to each other using blastp (version 2.9). Representative sequences which were more than 95% identical, over 80% of their length, were merged.
7. Following corrections, the connected components of the combined Roary graph were recalculated and these were the final set of gene clusters in the entire dataset (Figure 4.2, Step 6).

4.3.9 Statistical analysis

Statistical analyses were performed in R (v3.3+). Ape (v5.3) [395] and ggtree (v1.16.6) [396] were used for phylogenetic analysis and visualisation. The ggplot2 (v3.2.1) package was used for plotting [360].

4.4 Results

4.4.1 Constructing a collection of 10,000 *E. coli* isolates

A collection 18,156 *E. coli* genomes, isolated from human hosts, were downloaded and curated to create a final collection of 10,159 genomes as summarised in (Figure 4.1).

4.4.1.1 Initial collection of 18,156 genomes

For an initial collection of human *E. coli* genomes for which complete metadata is available, whole genome sequences were downloaded and the metadata combined from recent publications describing specific *E. coli* pathotypes. These included 70 EPEC isolates from [115], 398 EPEC isolates from [119], 373 ETEC isolates from [117], 1,509 ExPEC isolates from [397], 302 ExPEC isolates from [121], 113 EHEC and EPEC from [116], 538 ExPEC isolates from [174] and 25 ExPECs from [398]. Additionally, 140 isolates were taken from the Murray collection [399], which includes isolates collected from the pre-antibiotic era. Furthermore, 313 genomes were available from the NCTC reference collection which have been long read sequenced (https://www.sanger.ac.uk/resources/downloads/bacteria/nctc/).

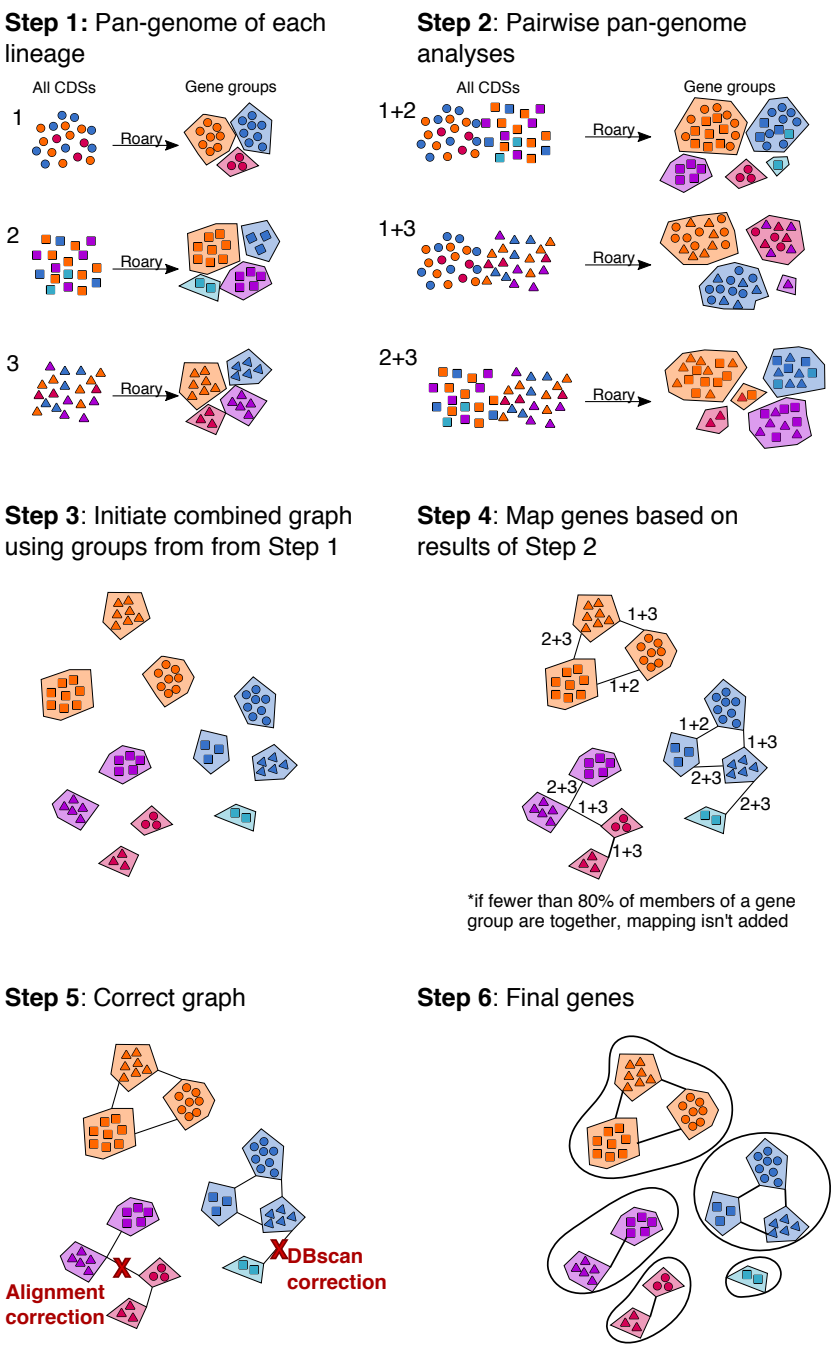

Figure 4.2: Method for combining the pan-genome analysis of all PopPUNK Clusters. *Step 1: a pan-genome analysis is applied on each PoPPUNK Cluster separately, generating gene clusters from all the CDSs of all genomes in that cluster. Step 2: A pan-genome analysis using Roary was applied on all PopPUNK Clusters reciprocally, generating new gene clusters. Step 3: A graph is constructed where the original gene clusters are the nodes. Step 4: An edge between two gene clusters was added if the members of both gene clusters were grouped together in the pairwise pan-genome analysis. Step 5: Edges were removed from the graph using density-based clustering and sequence alignments. Step 6: Connected components were extracted as the final gene cluster definitions.*

These genomes were supplemented to include other genomes available from public databases for which there was only partial associated metadata available. 14,590 genomes (isolated from human hosts) were downloaded from EnteroBase [400] on August 1st, 2018. EnteroBase searches the NCBI short read archive every day to download (and assemble) newly submitted Illumina reads or complete genomes (See Section 1.4). These genomes were filtered to include only genomes which were sequenced with Illumina, Pacbio or Minion platforms and were open for use, leading to a total of 8,045 genomes. Enterobase's data usage policy states metadata, assemblies and genotyping can only be used for academic purposes following their release. Therefore, the remaining genomes in the dataset were mostly from either publications or otherwise from public surveillance institutions from which we were able to obtain approval to use. These include Public Health England (PHE), the Food and Drug Administration (FDA) and the CDC. An additional 6,589 raw read sequences from Public Health England Routine surveillance bioproject (PRJNA315192) were downloaded on September 17th, 2018.

All downloaded reads were assembled (See Section 4.3.1.2). Artificial reads were generated for assemblies for which reads were unavailable (See Section 4.3.1.1). Annotation files were generated using a modified version of PROKKA, detailed below [293]. By the end of the data collection process, reads, assemblies and annotations were available for all genomes.

4.4.1.2 Modifying the annotation tool PROKKA to remove errors in gene calling between genomes

Prokka combines the use of five other tools to identify features in the assemblies. Importantly, Prokkka uses Prodigal to predict CDSs, or "genes" as they will be referred to in this thesis for simplicity [292,293], By default, Prokka will use the input genome to define properties for gene calling such as the start codon usage, ribosomal binding site motif usage etc. [292]. In this thesis, a collection of 100 randomly sampled genomes from the complete collection of genomes were used to train Prodigal to define these properties (See Section 4.3.1.3). All the genomes were then annotated using the same training properties. This ensured the gene calling was done in a consistent manner for all genomes.

In most cases, the gene content between the modified and default versions Prokka varied by less than 4%, with 96.5% of genes being called the same using both versions (Figure 4.3A). However, there were a number of outlier genomes for which the difference in gene content was much higher. The difference in these cases was mostly driven by genes within each

genome which were no longer called when using the same training file across all genomes (Figure 4.3B). In general, the genes which were differentially called were shorter, had a more varied GC content, were often present on shorter contigs and closer the contig edge, and more often began with an alternative start codon (Figure 4.3C-G).

4.4.1.3 Filtering to a high-quality collection of 10,159 genomes

Genomes were removed from the collection in multiple steps along the collection process when they did not pass the QC measures (Figure 4.1).

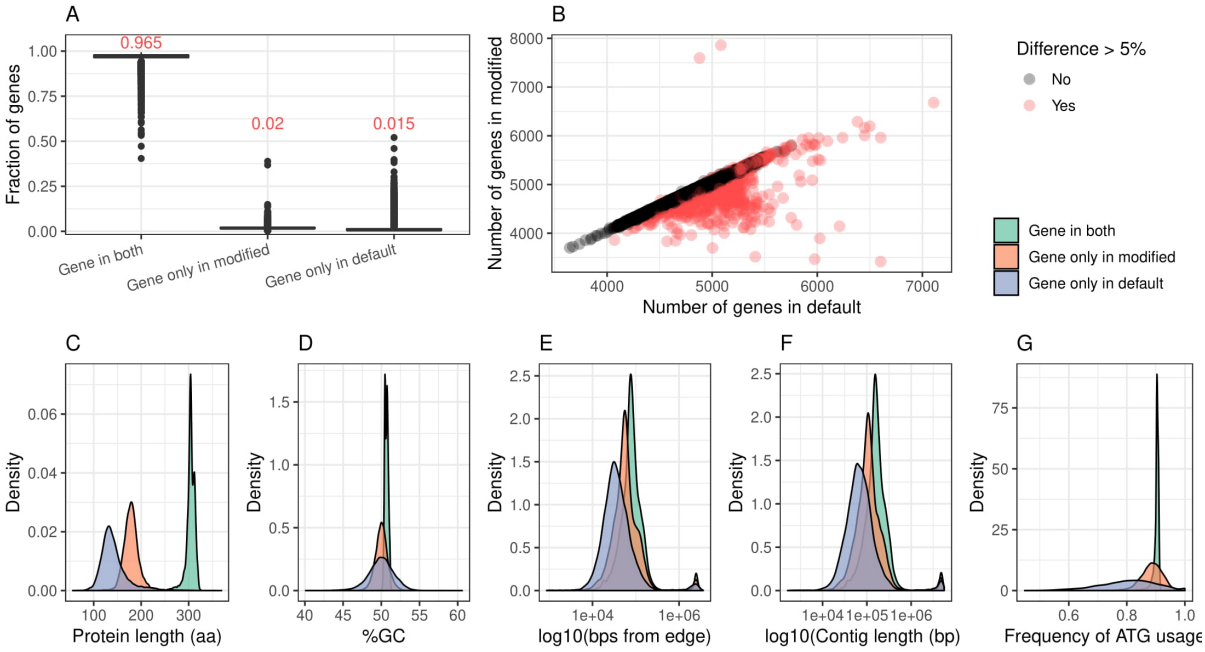

Figure 4.3: Effect of modifying Prokka on the CDS prediction. *The default version generated CDS properties for each genome individually, the modified used the same properties for all genomes.* ***A*** *Fraction of genes in each genome which was found in both runs, only in the modified run and only in the default. Red text: the average fraction of genes in each group across the 10,000 genomes.* ***B*** *Relationship between the number of genes in the default run compared to the modified run for each genome. Red: outliers from A for which there is more than 5% difference in gene content between both runs.* ***C-G*** *Protein length (**C**), GC content (**D**), distance from contig end (**E**), contig length (**F**) and frequency of ATG usage (**G**) of genes that were called in both, modified and default Prokka runs.*

Read filtering: Kraken was used on the reads to determine what organism had been sequenced [401]. Kraken uses a k-mer based search of the reads on a taxonomy tree of RefSeq genomes to find the most likely taxon for each read. If fewer than 30% of reads were assigned to *E. coli* or *Shigella* spp., the genome was removed (Figure 4.1). Following that, reads were mapped to an *E. coli* reference strain cq9 (GCF_003402955.1) and QC stats were

calculated. Samples were removed based on the according to the distributions of QC values across all reads (Percentage of reads mapped to the reference >60%, the mean insert size <80bp, percentage of bases mapped that were mismatches was >0.03, percentage of heterozygous SNPs<3%).

Assembly filtering: Assembled genomes were filtered to remove those with more than 600 contigs or those that had a total combined contig length of less than 4 Mbps or larger than 6 Mbps (Figure 4.4A,B, 4.1).

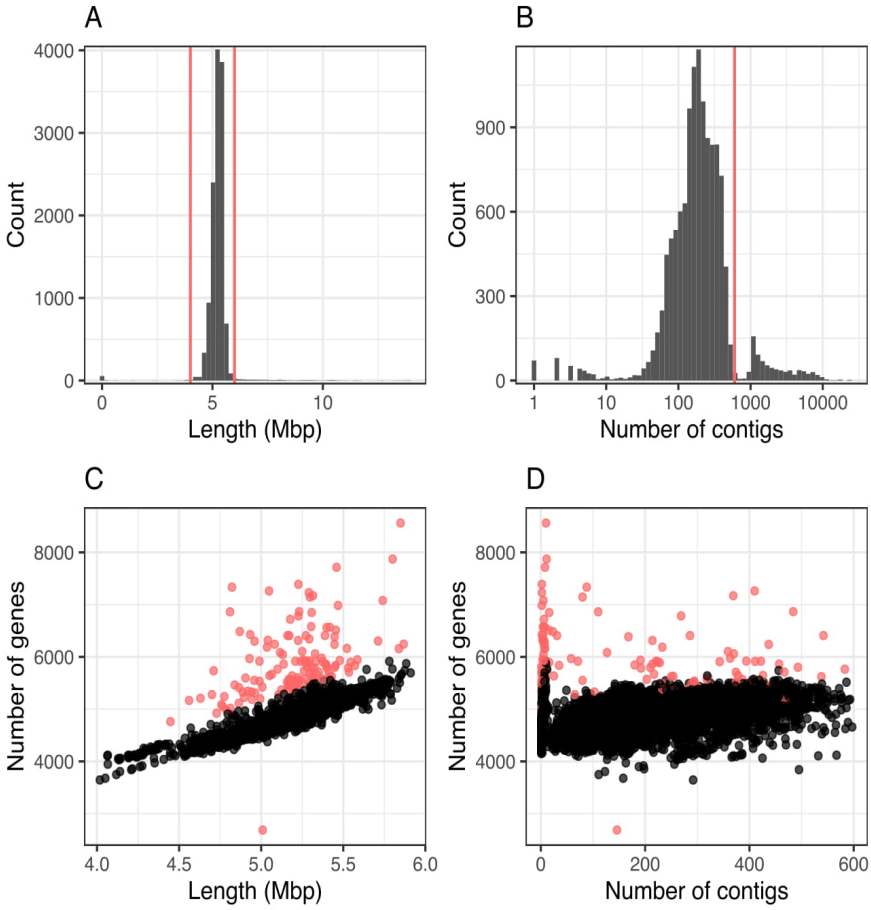

Figure 4.4: Quality control measures used to filter* E. coli *genomes. ***A*** *Distribution of genome lengths in the collection. Red lines: genomes shorter than 4 Mbps or longer than 6 Mbps were removed.* ***B*** *Distribution of number of contigs per genome in the collection. Red line: genomes with more than 600 contigs were removed.* ***C*** *Correlation between genome length and number of predicted CDSs using Prokka. Red: Genomes which deviate from the expected number of genes were removed.* ***D*** *Relationship between the number of contigs and number of predicted genes. Red: Genomes which deviate from the expected number of genes presented in C.*

Annotation filtering: The number of genes from each genome was retrieved from the annotations. There was a linear correlation between the size of the genome and the number of genes called (Figure 4.4C). Genomes which deviated from linear correlation by 500 genes were removed (Figure 4.1). These genomes tended to have fewer contigs, i.e. they were long-read sequenced (Figure 4.4D).

Average Nucleotide Identity based filtering: Mash distances were calculated between all the assemblies [279]. Mash uses a minimised database of k-mers to represent each genome (based on the Minhash sketch), and returns the Jaccard distance between the k-mers of every two genomes. A network was constructed so that there was an edge between every two genomes only if their Mash distance was smaller than 0.04 (equivalent to 96% Average ANI) [279]. Isolates from the same species should have an ANI of approximately 95-96%, i.e. Mash distance smaller than 0.04 [402]. Therefore, genomes were removed if they were disconnected from the largest connected component which should represent the *E. coli* species (Figure 4.1).

4.4.2 Characteristics of the filtered dataset

4.4.2.1 Most of the genomes are from developed countries, collected in surveillance in clinical settings

The vast majority of genomes were available from public resources which conduct regular surveillance of *E. coli* in clinical settings. These PHE (5,207 genomes), FDA (883 genomes) and the CDC (561 genomes) (Figure 4.5A). The availability of surveillance data from the United Kingdom and the United States lead to a biased collection from these countries which represented 70% (7,085/10,158) and 15% (1,548/10,158) of the dataset respectively. The rest of the genomes originated mostly from other countries in Europe, with only a small fraction of genomes available from Asia, Africa and Oceania (Figure 4.5A). The continent and country of 336 genomes was unknown.

The source of isolation for 38% of the samples considered here were taken from faeces, blood and urine (Figure 4.5B). However, the remaining samples were simply recorded as having been isolated from unknown “human sources”. Isolates from Africa and Asia include only those collected from faecal samples, whereas isolates from Europe and North America include those causing both intestinal and extra intestinal disease (Figure 4.5B).

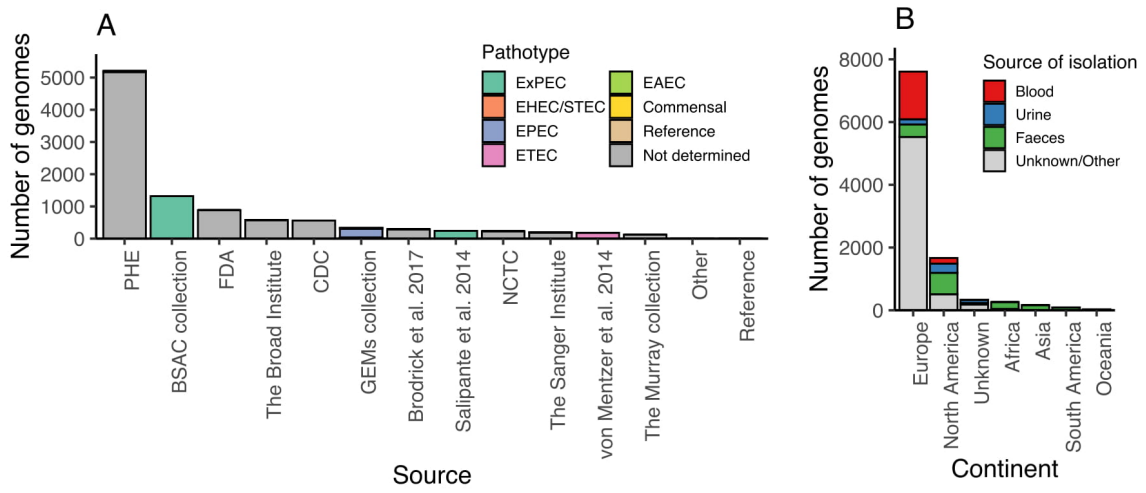

***Figure 4.5: Source of* E. coli *genomes. A** Source of the* E. coli *genomes in the collection, coloured by the pathotype associated with the specific studies. **B** Continents from which the* E. coli *genomes were collected, coloured by source of isolation.*

4.4.2.2 Only 5% of all genomes are the cause of diarrheal disease in developing countries

The pathotype for isolates taken from urine and blood samples was assigned as ExPEC (2,299 genomes, 15%). The metadata of 522 (5%) isolates was available and thus the pathotype was known, based on the publication (Figure 4.5A). Within these isolates, the representation of diarrheal disease causing *E. coli* pathotypes, EPECs and ETECs, was very low with only 3% and 2% of the genomes belonging to these pathotypes, taken from the The Global Enteric Multicenter Study (GEMS collection) and from [117] (Figure 4.5A). For the remainder of the genomes, the pathotype could not deterministically be assigned (7,335 genomes). This is due to pathotypes not being defined by clear one to one relationship of presence or absence of specific virulence genes, but by clinical manifestation or phenotype. In Section 4.4.4.7 of this thesis, the virulence profiles of genomes are described as predictive of their pathotype (See Section 1.1.2.3, and Figure 1.4).

4.4.2.3 Six STs represent more than 50% of the genomes in the collection

993 different STs were identified in the collection. 87 STs (9%) alone account for 80% of the isolates. Six STs, 11, 131, 73, 10, 95 and 21, account for 50% of the isolates (Figure 4.6A,B). Many of the latter represent important STs linked to human health. For instance, ST11 (30% of all genomes) is associated with EHEC serotype O157:H7, a major foodborne pathogen that can be contracted by eating contaminated foods, specifically beef products, as since it lives in

the guts of cattle and is the cause of HUS (See Section 1.1.2.2). The collection also includes STs of non-O157 EHECs, including STs 17 (2%) and 21 (2%). STs 131 (8%), 73 (4%), and 95 (3%) are all STs known to be associated with extra-intestinal disease[174,397,403]. ST10 (3%) is a broad host range ST which has been observed in all *E. coli* pathotypes and across hosts [404].

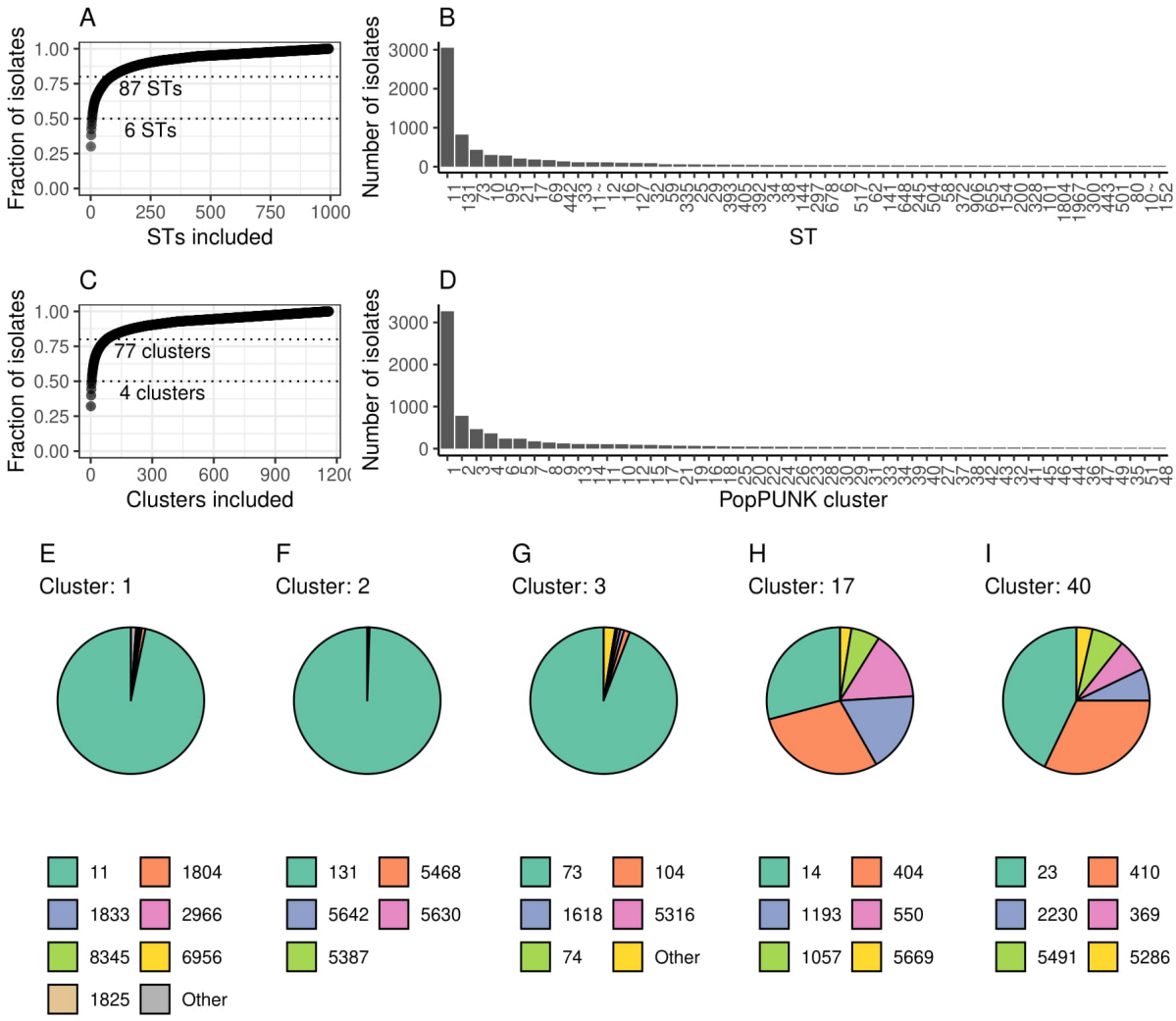

***Figure 4.6: Distribution of STs and PopPUNK Clusters in the collection**. **A,C** Coverage of genome collection by increasing the number of STs (A) or PopPUNK clusters (C) included in the study. Dotted lines: Number of STs (**A**) or PopPUNK clusters (**C**) which account for 0.5 and 0.8 of all isolated in the genome collection. **B,D** Number of genomes in the fifty largest STs (**B**) and PopPUNK clusters (D). **E-I** Examples of ST distributions in five of the PopPUNK Clusters - Cluster 1 (**E**), 2 (**F**), 3 (**G**), 17 (**H**) and 40 (**I**).*

The bias in the collection towards STs which are known to cause severe disease such as HUS or invasive infections emphasises the sampling bias; 80% of isolates originate from developed

countries where diarrheal disease caused by EPEC and ETEC is less common. 790 STs (~80% of the STs) are represented by five isolates or fewer and are rarely observed. Thus, this collection is inherently biased towards clinical isolates which are under surveillance in the UK and US, and does not represent the human *E. coli* population.

4.4.3 PopPUNK can be used to group the collection into isolates belonging to the same lineage

In order to examine the gene pool of the *E. coli* genomes considered here, the genomes were grouped into clusters of closely related isolates using PopPUNK [277]). PopPUNK uses a k-mer based comparison of genomes to measure the deviation in gene sequence termed the "core distance", and the deviation in gene content, termed the "accessory distance" between two genomes (See Section 4.3.3). In *E. coli*, it was shown that the "core distance" estimated by PopPUNK correlates with the pairwise SNP distance between the two genomes being compared, and the "accessory distance" correlates with the Jaccard distance based on the presence and absence of CDSs extracted from a pan-genome analysis [277]. Genomes which were sufficiently similar in both their "core distance" and their "accessory distance" were included in the same PopPUNK Cluster (See Section 4.3.3).

This approach was taken in order to handle the biased sampling of the genomes. For instance, the dataset is over-represented with ST11; had all isolates been treated with the same weight in the analysis, the results would be biased to ST11. By examining the gene content within each subpopulation individually and then merging these results while adding weights for the sampling bias, conclusions can be drawn.

The grouping produced 1,185 PopPUNK Clusters. The partition of the genomes using PopPUNK mostly agreed with partitioning the genomes by ST (rand index of 0.923). Therefore, the distribution of PopPUNK cluster sizes was similar to that of the STs with a few large clusters representing most of the population (Figure 4.6A,B). A single cluster, PopPUNK Cluster 1, contained 34% of all genomes (3,326/10,158) (Appendix E). This cluster was mostly comprised of ST11 (Figure 4.6E), i.e. O157:H7 EHEC. Similarly, PopPUNK Cluster 2 contained 8% of all genomes (781/10,158) consisted mostly of ST131 (Figure 4.6F). The third largest cluster, PopPUNK Cluster 3, contained 5% of all genomes (463/10,158) and was mostly composed of ST73 (Figure 4.6G). See Appendix E for a summary of all other PopPUNK Clusters. There were exceptions for which a higher diversity of STs within a PopPUNK Cluster was observed. For instance, PopPUNK Cluster 17 which had 79 isolates, consisting of four almost equally distributed STs (14, 404, 1193 and 550) (Figure 4.6H). PopPUNK Cluster 40,

which had 28 isolates, was composed of two equally common STs (410 and 23) along with another four which were less common (Figure 4.6I).

For this analysis, PopPUNK Clusters of fewer than twenty isolates were removed. There were 50 PopPUNK Clusters in total which met this requirement and together they contained 7,693 genomes (76% of the collection) and 271 different STs (27% of collection) (Appendix E). Whilst the effect of this is a further reduction in the diversity of the dataset, it is not possible to characterise the gene pool of groups for which there were too few representatives. Additionally, this approach would further filter out contaminants and isolates which may not be *E. coli*.

4.4.4 Characteristics of the selected 50 largest PopPUNK Clusters

4.4.4.1 Genetic diversity

The median "core distance" and median "accessory distance" estimated within each of the remaining PopPUNK Clusters were correlated, with higher deviations in the core indicating higher deviations in gene content, i.e. in the accessory genome (linear regression, p=1.342e-11, R^2=0.61) (Figure 4.7). However, differences between the PopPUNK Clusters were evident, with some PopPUNK Clusters presenting higher diversity in their accessory genome relative to their core genome, and vice versa. For instance, PopPUNK Cluster 40, which contains isolates of STs 410 and 23, had high diversity in its accessory genome relative to the core genome. There was no connection between the size of the PopPUNK Cluster and the median "core" or "accessory" distances (not shown).

4.4.4.2 Population structure

The phylogeny of the 50 selected PopPUNK Clusters was examined by selecting ten genomes from each PopPUNK cluster that captured most of the diversity of that cluster (See 4.3.4), leading to a total of 500 genomes representing the complete dataset. The core genome of these 500 genomes was extracted and the phylogenetic tree of the core gene alignment was built (Figure 4.8). PopPUNK separated the genomes into clearly distinct lineages based on their core genome. The effect of the "accessory distance" between every two isolates was minimal as there was a correlation between "core" and "accessory" distance across the isolates (Figure 4.7). The exception to this was PopPUNK Cluster 12 which was split into two closely related clades. One clade was more closely related to PopPUNK Cluster 28 whereas the other to PopPUNK Cluster 35. The "core" and "accessory" distances estimated by PopPUNK showed that indeed the "core" distance between PopPUNK Clusters 12, 28 and 35

were low and these could be viewed as a single clade according to their core distances. However, PopPUNK Clusters 12, 28 and 35 deviate in their accessory gene content from PopPUNK Cluster 12 whereas the two clades of PopPUNK Cluster 12 are sufficiently low in their accessory distance. That said, PopPUNK Cluster 12 presented the highest median "core distance" and median "accessory distance" between every two isolates (Figure 4.7).

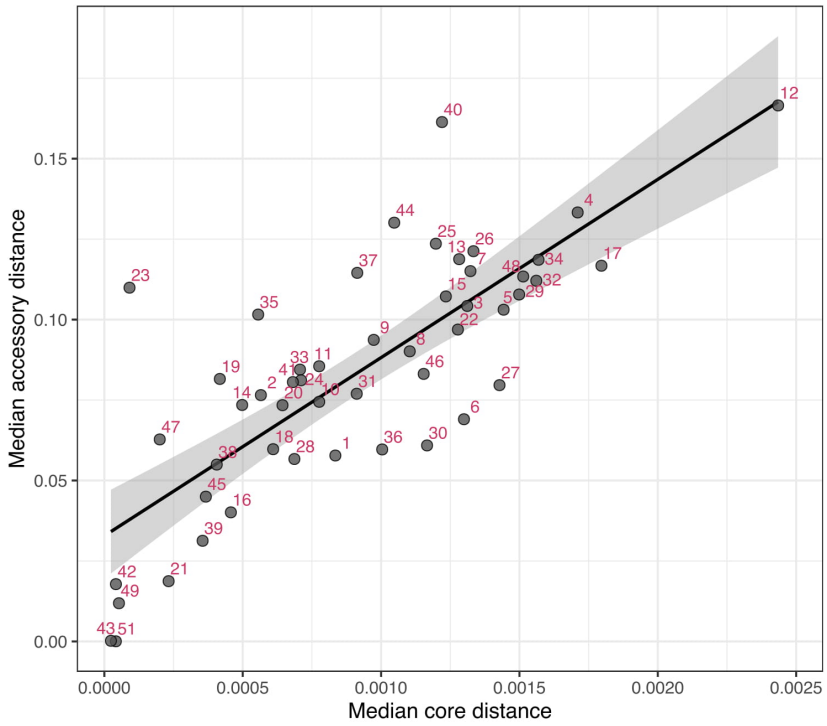

Figure 4.7: PopPUNK Clusters' genetic diversity. *Median "core distance" and "accessory distance" between all isolates of the same PopPUNK Cluster. Line fitted using linear regression, showing 0.95 confidence interval.*

Although the dataset was substantially reduced to include only PopPUNK Clusters with 20 genomes or more, the remaining genomes spanned the complete *E. coli* population, defined by having PopPUNK Clusters representing the well described *E. coli* phylogroups (18 from B1, 12 from B2, 4 from A, 5 from D, 4 from F, 3 from E, 1 from C, 2 of *Shigella* representing *S. sonnei* (45) and *S. flexneri* (30) and one phylogroup which was undefined according to the Clermont 2013 phylotyping scheme (18) [271,393]).

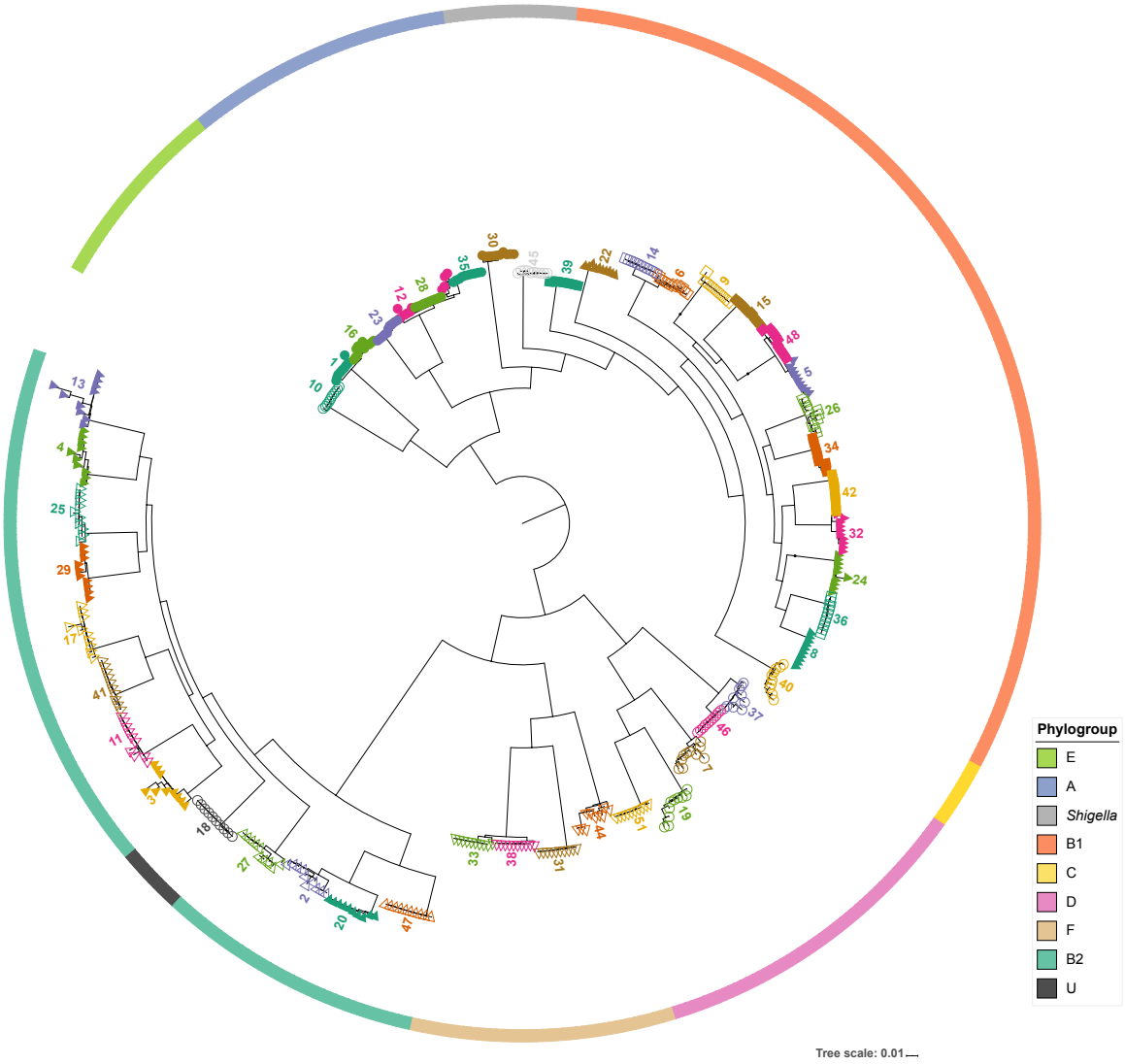

Figure 4.8: Population structure of the PopPUNK Clusters. *Core gene phylogeny of 10 representatives from each of the 50 PopPUNK clusters chosen using Treemer [392]. Coloured bar indicates the phylogroup assignment of the representatives of that PopPUNK Cluster.*

4.4.4.3 Pathogenic and geographic association

The PopPUNK Clusters broadly divided into those enriched for isolates collected from faecal samples (2, 5, 6, 14, 21, 26, 34, 42, 43, 48, 49 and 51) and those collected from blood and urine samples (2, 3, 4, 7, 11, 13, 17, 19, 20, 25, 29, 31, 33, 37, 40, 41, 46, and 47), i.e. those causing intestinal or extra-intestinal disease (Figure 4.9A). Only PopPUNK Clusters 26, 34 and 48 of the intestinal causing disease clusters were enriched for samples collected from Africa and Asia (Figure 4.9B). These clusters mostly represented EPEC and ETEC isolates which had been collected from faecal samples in developing countries as part of the GEMS collection, in contrast to the other PopPUNK Clusters containing faecal samples which include

STECs or EHECs collected in the developed world. PopPUNK Cluster 12, which consisted of 78% isolates from ST10, was the only PopPUNK Cluster that spanned all continents and consisted of all types of isolation source samples (faecal, blood, urine or unknown).

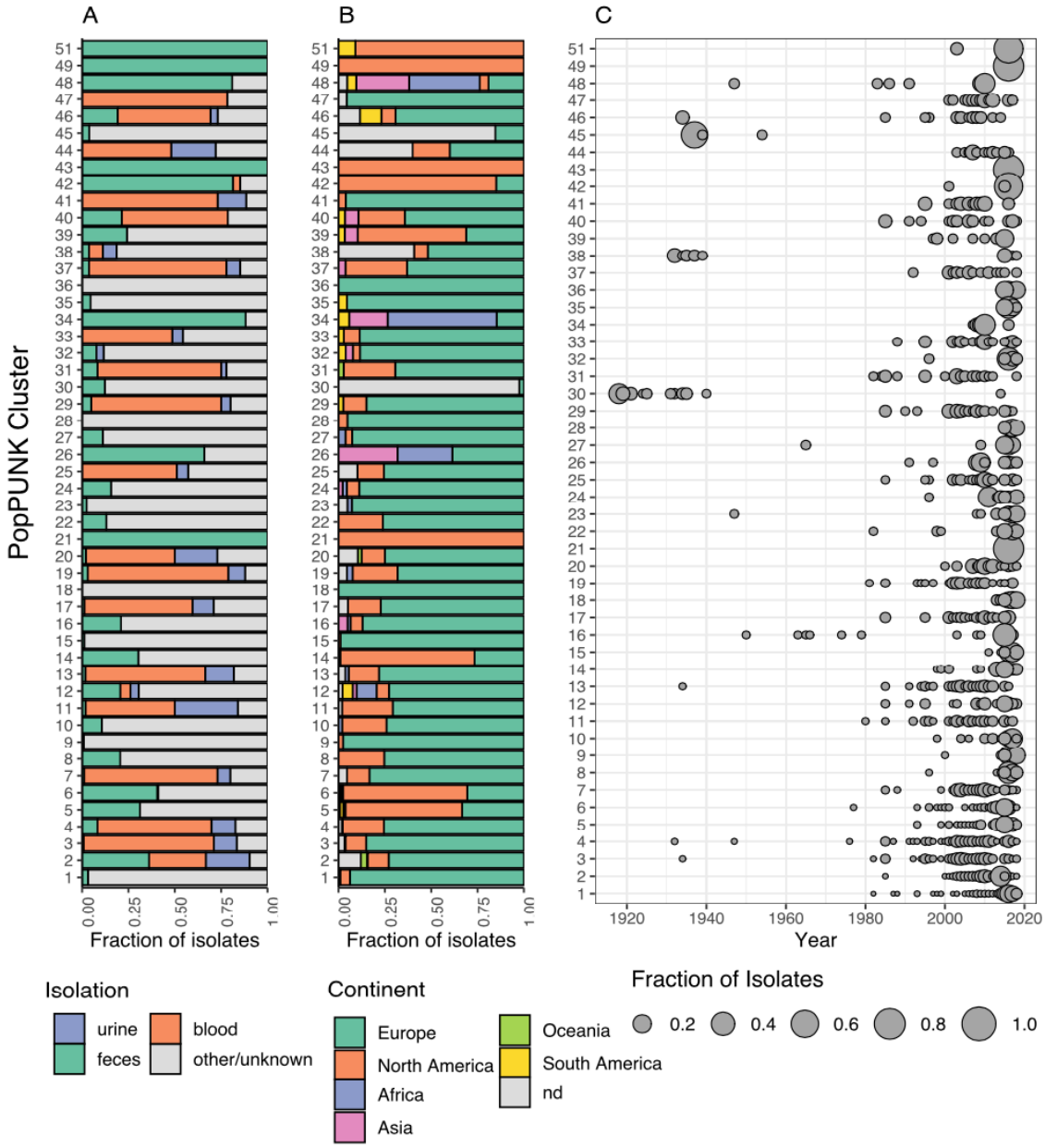

Figure 4.9: Metadata associated with the PopPUNK Clusters. A,B *Source of isolation (**A**) and continent (**B**) of isolates from the fifty PopPUNK Clusters.* ***C*** *Fraction of genomes from each of the PopPUNK Clusters collected from each year (where metadata was available).*

4.4.4.4 Sampling time

A number of PopPUNK Clusters consisted of older isolates taken from the Murray collection. Notably, PopPUNK Cluster 30, with contains *S. flexneri* isolates, had a higher proportion of

isolates sampled before 1980 relative to the rest of the collection (Wilcox summed rank test, p<0.05, Bonferroni corrected, Figure 4.9C). 39% of the rest of the genomes for which sampling date was available, were collected in the last 10 years.

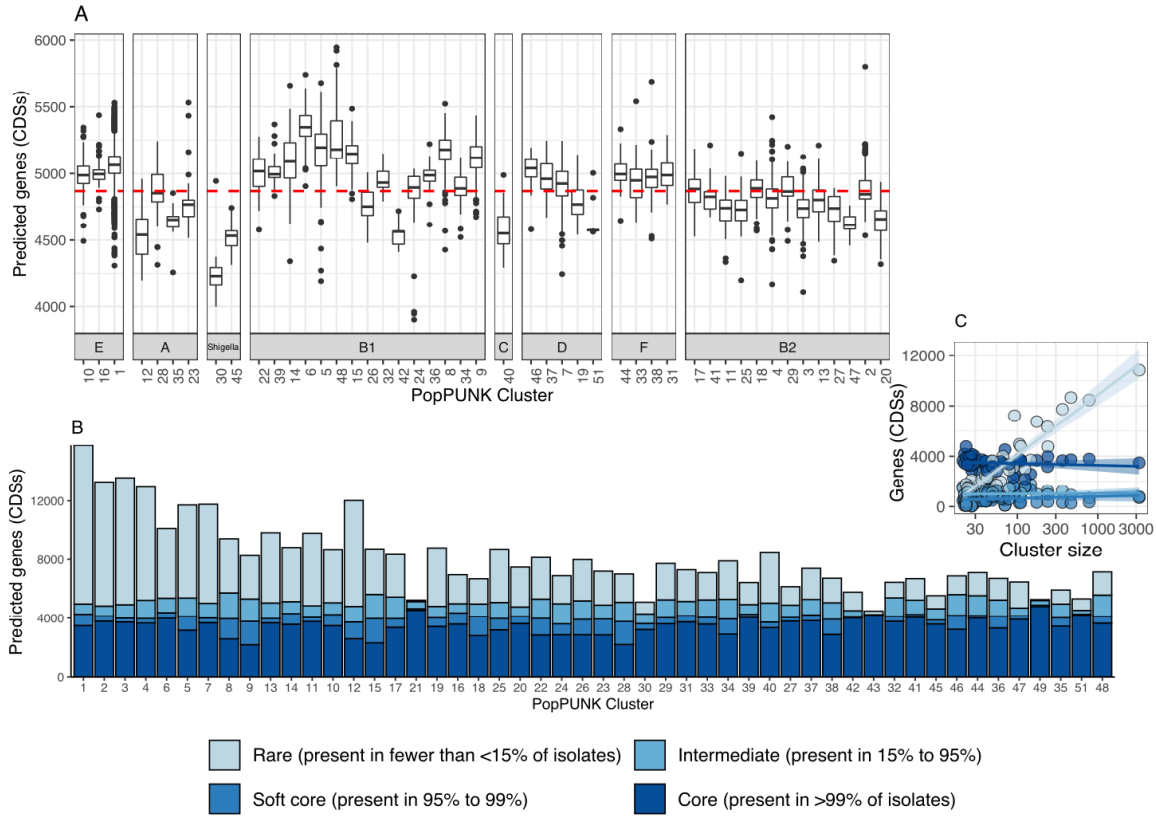

Figure 4.10: Gene content in the 50 PopPUNK clusters. A *Number of genes (predicted CDSs) per isolate across the PopPUNK clusters, divided by their phylogroup. Dashed line: mean number of predicted genes across the entire population.* ***B*** *Number of core (>99% of isolates), soft-core (95%-99% of isolates), intermediate (15%-95% of isolates) and rare genes (<15% of isolates) in each PopPUNK Cluster. Clusters on the x-axis are ordered by their size.* ***C*** *Size of PopPUNK Cluster against number of core, soft-core, intermediate and rare genes. Line is fitted using a generalised log-linear model with 0.95 confidence interval.*

4.4.4.5 Genome size and number of predicted genes

The number of genes in a single isolate and the size of the genome varied significantly between the PopPUNK clusters (Figure 4.10A). The mean number of genes corrected across all PopPUNK Clusters was 4,869 genes and a genome length of 5.2 Mbp. Smaller genomes had fewer genes as we used the correlation between genome length and the number of genes as a measure of QC, thus these measures are interchangeable (See Section 4.4.1.3). Isolates from the *Shigella* PopPUNK Clusters 30 and 45 had the smallest genomes with a median of

4,231 genes per isolate and a genome size of only 4.3 and 4.7 Mbp. PopPUNK Clusters 12, 40 and 48, had the second smallest genome lengths with a mean of ~4,500 genes and genome length of ~4.85Mbp. On the other hand, PopPUNK clusters 5, 6, 8, 15, and 48, all from phylogroup B1, had a mean of over 5,100 genes per isolate (200 genes more than the population mean). The number of predicted genes/length of the genome was affected by the phylogroup (Figure 4.10A). Isolates from phylogroups E, F and B1 tended to have larger genomes with a few exceptions. Isolates from phylogroup C, B2 and A tended to have smaller genomes, whereas within phylogroup D a wider range of genome sizes was observed.

4.4.4.6 Antimicrobial resistance profiles

A total of 153 known resistance genes were identified in the collection (See Section 4.3.6), conferring resistance to beta-lactamases, aminoglycosides, macrolides, sulfonamides, fluoroquinolones and other antimicrobial classes (Appendix E) [286]. The number of known resistance genes found within each isolate ranged from no resistance genes detected to a maximum of 18 resistance genes present in a single isolate, conferring resistance to up to nine different antimicrobial classes in a single isolate (Figure 4.11A). The median number of resistance genes per isolate in the complete dataset was one gene. This was because 99% of isolates possess the multidrug resistance efflux pump gene *mdfA[405]* (Figure 4.11B).

Multidrug resistance in an isolate has been defined as resistance to three classes of antibiotics or more [406]. All but six PopPUNK Clusters (21, 28, 36, 43, 47 and 49) had at least one isolate which was MDR. An MDR PopPUNK Cluster was defined as one where half of the isolates or more were MDR, i.e. resistant to three classes of antibiotics or more (Figure 4.11A, Appendix E). 16 of the 50 PopPUNK Clusters investigated in this thesis were MDR. Half of these were PopPUNK Clusters which were isolated predominantly from blood and urine sample, i.e. ExPECs (Clusters 2, 20, 44, 40, 17, 7, 37 and 9). These include PopPUNK Clusters 2 and 20 which both contain isolates of the global ExPEC lineage ST131. Three of the ExPEC MDR PopPUNK Clusters belong to phylogroup D (Clusters 19, 7 and 37). These three PopPUNK Clusters possessed the same set of genes which confer resistance to ESBLs, sulfonamides, tetracycline and aminoglycosides (Figure 4.11B). Three other MDR PopPUNK Clusters predominantly contained EPEC isolates from the GEMS collection (Clusters 26, 34 and 48). The remaining five PopPUNK Clusters (Clusters 32, 35, 18, 16 and 24) were isolated from unknown sources. Resistance to carbapenems was most common within PopPUNK Cluster 44 of phylogroup F with 44% of the isolates of this Cluster possessing the carbapanemase *bla*KPC-2. Resistance in PopPUNK Clusters 44 as well as PopPUNK Cluster 37 were generally high, with most of the isolates in these PopPUNK Clusters resistant to seven classes of antibiotics or more, comparable and even higher to the resistance observed for

ST131 in PopPUNK Clusters 2 and 20. Resistance to colistin was not observed within any of the isolates in this dataset.

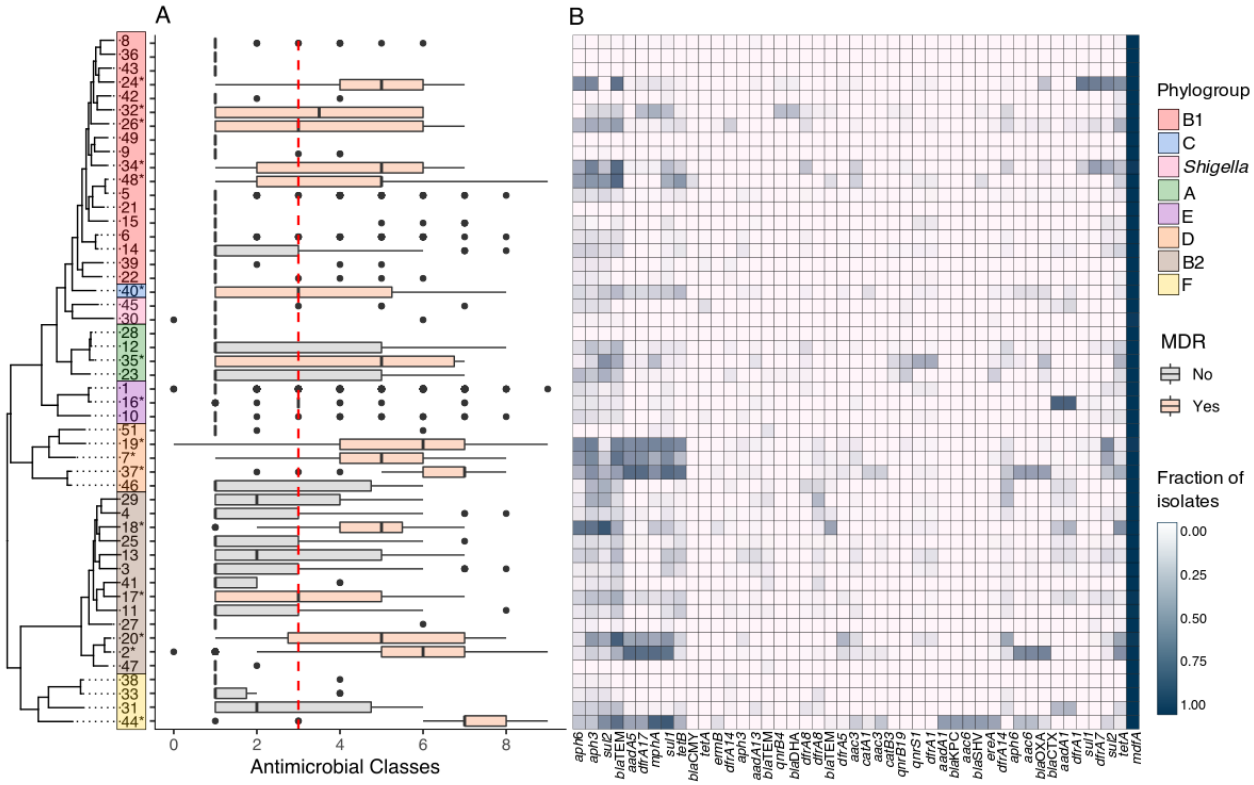

Figure 4.11: Antimicrobial resistance profiles of the PopPUNK Clusters. A *Number of antimicrobial classes each isolate is resistant to, stratified by PopPUNK Cluster. Dashed red line indicates threshold for multidrug resistance.* ***B*** *Heatmap presenting the frequency of each resistance gene within each of the 50 PopPUNK Clusters. (Presenting only genes which were found in at least 10% of isolates of one PopPUNK Cluster.) Darker squares indicate higher prevalence of a gene in the PopPUNK Cluster. Phylogenetic tree constructed by selecting one isolate from each PopPUNK Cluster using Treemmer* ***[392]*** *(See Methods 4.3.4). Asterisk by PopPUNK Cluster name indicates MDR cluster.*

The presence and absence patterns of antibiotic resistance genes are presented in Figure 4.11B. Particular resistance genes are widespread in the dataset, these include *sul2* and *blaTEM*. Certain resistance gene combinations tended to co-occur multiple times in distantly related PopPUNK Clusters. For instance, resistance genes *aac6*, *blaOXA* and *blaCTX* co-occur in the MDR PopPUNK Clusters 20, 37 and 44. The genes *aadA1* and *dfrA1* are present together in PopPUNK Clusters 31, 17, 18 and 16. Finally, most of the resistance genes observed were in fact observed rarely and only present in very low frequencies in this dataset.

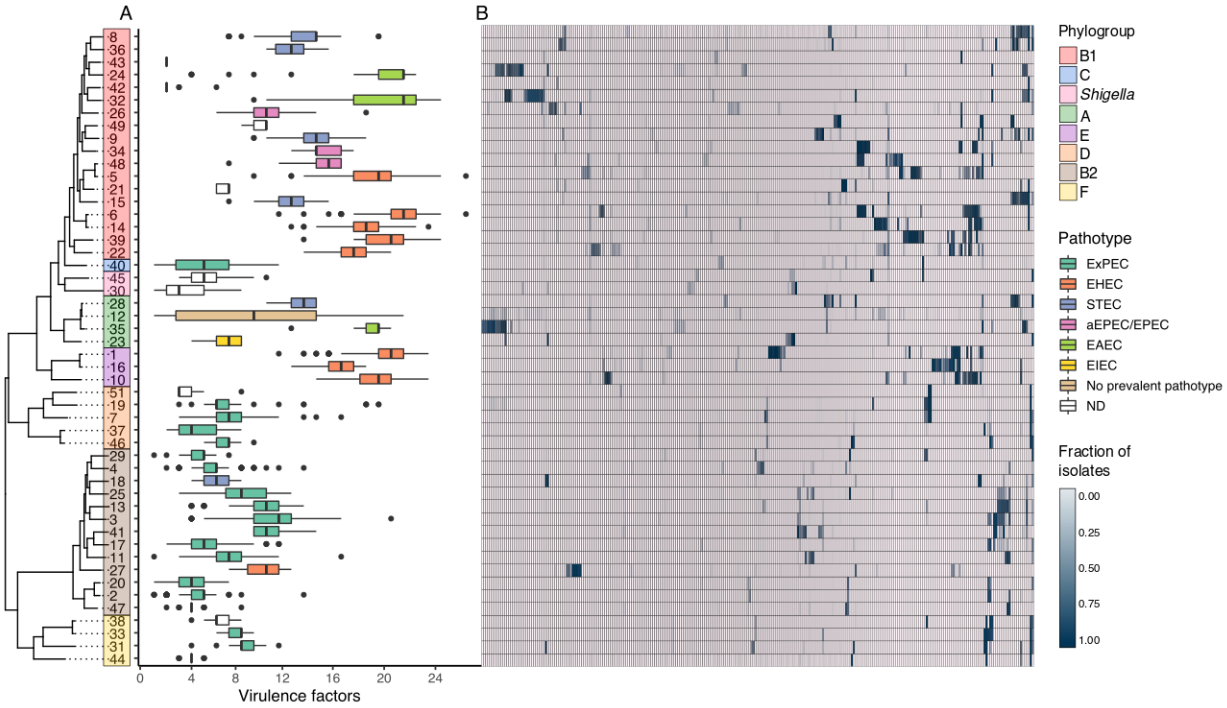

Figure 4.12: Markers of virulence in the PopPUNK Clusters. A *Number of virulence genes per isolate, stratified by PopPUNK cluster and coloured according to the most prevalent predicted pathotype in the cluster. ND = "Not Determined"* ***B*** *Heatmap presenting distribution of the virulence genes across the 50 PopPUNK clusters. Darker squares indicate higher prevalence of a gene in a lineage. (Presenting only genes which were found in at least 10% of isolates of one PopPUNK Cluster.) Phylogenetic tree constructed by selecting one isolate from each PopPUNK cluster using Treemmer [392] (See Methods XX).*

4.4.4.7 Markers of virulence

Consistent with the collection of *E. coli* isolates being from human hosts and mostly from clinical samples, 439 known virulence factors were observed in our dataset. The isolates had a median of nine known virulence factors in a single genome, with a maximum value of 26 virulence factors present in a single isolate (Figure 4.12A).

A combination of the source of isolation as well as the presence of key virulence factors were used to find the most prevalent predicted pathotype of each PopPUNK Cluster (See Section 4.3.7). 41 of 50 PopPUNK Clusters were identified as predominantly containing one of the defined *E. coli* pathotypes (See Section 1.1.2.2) (Figure 4.12A). Two of the PopPUNK Clusters without a prevalent pathotype were PopPUNK Clusters 30 and 45 which represent the *Shigella* species. PopPUNK Cluster 12, which mostly consists of *E. coli* isolates typing as ST10, was the only PopPUNK Cluster which contained isolates assigned to different pathotypes with no single pathotype dominating (11% ExPEC, 29% EAEC, 24% EPEC, 9% STEC, 2% EHEC,

1% ETEC, and 24% Not Determined (ND)). Indeed, PopPUNK Cluster 12 had the highest variability in number of virulence genes per isolate, relative to the rest of the clusters (Figure 4.12A). The remaining six PopPUNK Clusters which were not assigned a pathotype (21, 38, 42, 43, 45, 49 and 51) had relatively few virulence factors per isolate as well as low levels of predicted resistance, perhaps representing non-virulent lineages (Figure 4.12A).

Phylogroups B2, F, and D predominantly contained ExPEC isolates. PopPUNK Cluster 27 was the only cluster in phylogroup B2 which contained 67% EHEC isolates and 33% aEPEC/EPECs. PopPUNK Cluster 18, also nested within phylogroup B2 but not assigned a phylogroup according to the Clermont typing scheme, contained 100% STEC isolates. All PopPUNK clusters of phylogroup E contained predominantly EHEC isolates (Figure 4.12A, Appendix E). Phylogroups A and B1 had more diversity of pathotypes, containing PopPUNK Clusters which were assigned to the range of diarrheagenic pathotypes (EPEC, EHEC, EAEC and EIEC). PopPUNK Cluster 24 of phylogroup B1 also contained 38% isolates which were *stx* and *eae* positive. These are isolates of *E. coli* serotype O104:H4 taken from the 2011 German outbreak, which were classified as both shiga-toxin producing EAEC [407] (See Section 1.1.2.2). PopPUNK Cluster 40, the only cluster assigned to phylogroup C, was the only ExPEC cluster within the B1-C-A clade (Figure 4.12A).

The number of virulence factors per isolate differed between the phylogroups depending on their predominant pathotype (Figure 4.12A). Phylogroups containing ExPEC isolates (B2, D, F and C) had fewer virulence factors per isolate, relative to phylogroups containing the PopPUNK Clusters of the diarrheagenic *E. coli* (E, B1 and A). This could be a result of biases in the virulence factor database and our lack of complete understanding of ExPEC virulence factors.

The virulence factors identified in this dataset were more commonly specific to a PopPUNK Cluster and were generally not widespread across the whole dataset. PopPUNK Clusters which had a large number of virulence genes per isolate tended to possess a set of virulence factors which were otherwise not shared with other PopPUNK Clusters. This is exemplified in Figure 4.12B for PopPUNK Cluster 27, 10, 35 and more. Exceptions to this exist for virulence factors which were shared across PopPUNK Clusters which were assigned to the same pathotype, such as the ExPEC PopPUNK Clusters in Phylogroup B2 or the EHEC PopPUNK Clusters in phylogroups E and B1.

4.4.4.8 Relationship between resistance and virulence

The PopPUNK Clusters divided into clear groups based on their pathotype when comparing the median number of antimicrobial classes each isolate was resistant to against the median number of virulence factors identified per isolate for each PopPUNK Cluster (Figure 4.13). PopPUNK Clusters which were not assigned to a pathotype were resistant only to a single class of antimicrobials, i.e. these were predicted to be non-virulent and non-resistant. PopPUNK Clusters containing mostly ExPEC isolates ranged in the number of antimicrobial classes they were resistant to, with the most resistant PopPUNK Clusters, 2, 44 and 37, containing predominantly ExPEC isolates. However, more than half of the ExPEC PopPUNK Clusters (11/19) showed only low levels of resistance. Shiga-toxin producing isolates, EHECs and STECs, showed low levels of resistance relative to a high load of virulence factors. Exceptions to this were PopPUNK Clusters 16 and 18 which were the only MDR STEC and EHEC Clusters. PopPUNK Cluster 18 was particularly peculiar for an STEC as it is nested within phylogroup B2 and had low number of virulence factors per isolate relative to other STECs. PopPUNK Cluster which contained predominanly EAEC and EPEC isolates were all MDR and highly virulent.

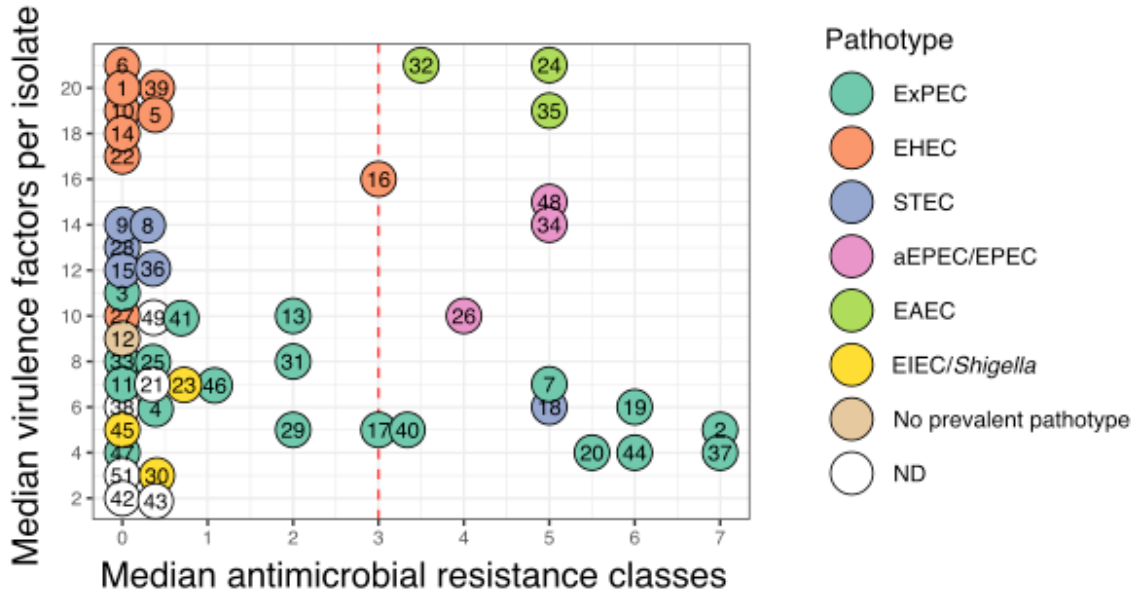

Figure 4.13: Relationship between resistance and virulence. *Each numbered dot represents a PopPUNK Cluster. Clusters are coloured by the most prevalent predicted pathotype in the cluster.*

4.4.4.9 Pan-genomes

A pan-genome analysis was applied on the isolates of each of the PopPUNK Clusters separately (See Section 4.3.8.1). Genes found within each PopPUNK cluster were divided into 4 categories based on their frequency within the cluster: genes present in more than 99% of isolates of a PopPUNK Cluster were labelled “core”, 95% to 99% of isolates were labelled “soft-core”, 15% to 95% of isolates labelled “intermediate” and “rare” were those present in fewer than 15% of isolates of a PopPUNK Cluster. The number of “core”, “soft-core” and “intermediate” genes in each PopPUNK cluster was stable across the clusters, regardless of the number of genomes in the cluster (Figure 4.10B,C). The number of “rare” genes per PopPUNK Cluster varied and was dependent on the cluster size, with larger PopPUNK clusters possessing more “rare” genes in their pan-genomes than smaller clusters (Figure 4.10C).

The pan-genome analysis on the PopPUNK clusters showed that there was low genetic diversity within PopPUNK clusters 21, 43 and 49. Therefore, these clusters were removed from the analysis, as they contain multiple isolates which were all collected at the same time and were all collected by the FDA (possibly representing an outbreak investigation).

4.4.5 Combining pan-genomes of the PopPUNK Clusters

Following the analysis of the pan-genome of each PopPUNK cluster individually, the outputs of all the analyses were combined in order to provide a description of the gene pool in the entire *E. coli* dataset analysed in this thesis. The precise steps taken are detailed in Section 4.3.7.2. Briefly, a reciprocal pairwise pan-genome analysis was run on every two PopPUNK clusters (Figure 4.2). The grouping of genes in every pairwise pan-genome analysis was examined to determine whether two genes from two separate PopPUNK clusters should be labelled as the same gene in the complete dataset. Since every pairwise comparison between genes was applied, it was possible to identify spurious matches between genes that were identified in single pan-genome analysis but were not supported across other pairwise gene comparisons. In addition, all representative sequences of a gene group were aligned and incorrect gene-groupings removed based on the SNP distances between the members.

4.4.6 Final collection of 55,039 genes

There were 55,039 genes (predicted CDSs) in the dataset after combining the genes of the pan-genomes of the 47 PopPUNK Clusters. As there were 47 PopPUNK clusters, and a varying number of isolates per cluster, each gene had its own frequency within each of the 47 PopPUNK Clusters. For instance, the *intA* gene, encoding a prophage integrase, was

observed in 20 of the PopPUNK Clusters. In two clusters (6 and 9) it was present in over 95% of isolates, in another 8 clusters it was present in intermediate frequency (between 15% and 95%) and in the final 10 clusters it was present in fewer than 15% of isolates (A). In contrast, the gene *wzyE*, a gene involved in antigen biosynthesis, is a core gene which was observed across all PopPUNK Clusters in a frequency of over 95% (Figure 4.14B). Principal component analysis was applied to all gene frequencies across the PopPUNK Clusters (Figure 4.14C). The first and second principal components explained 17.93% and 7.49% of the variance and separated the PopPUNK clusters by the phylogeny.

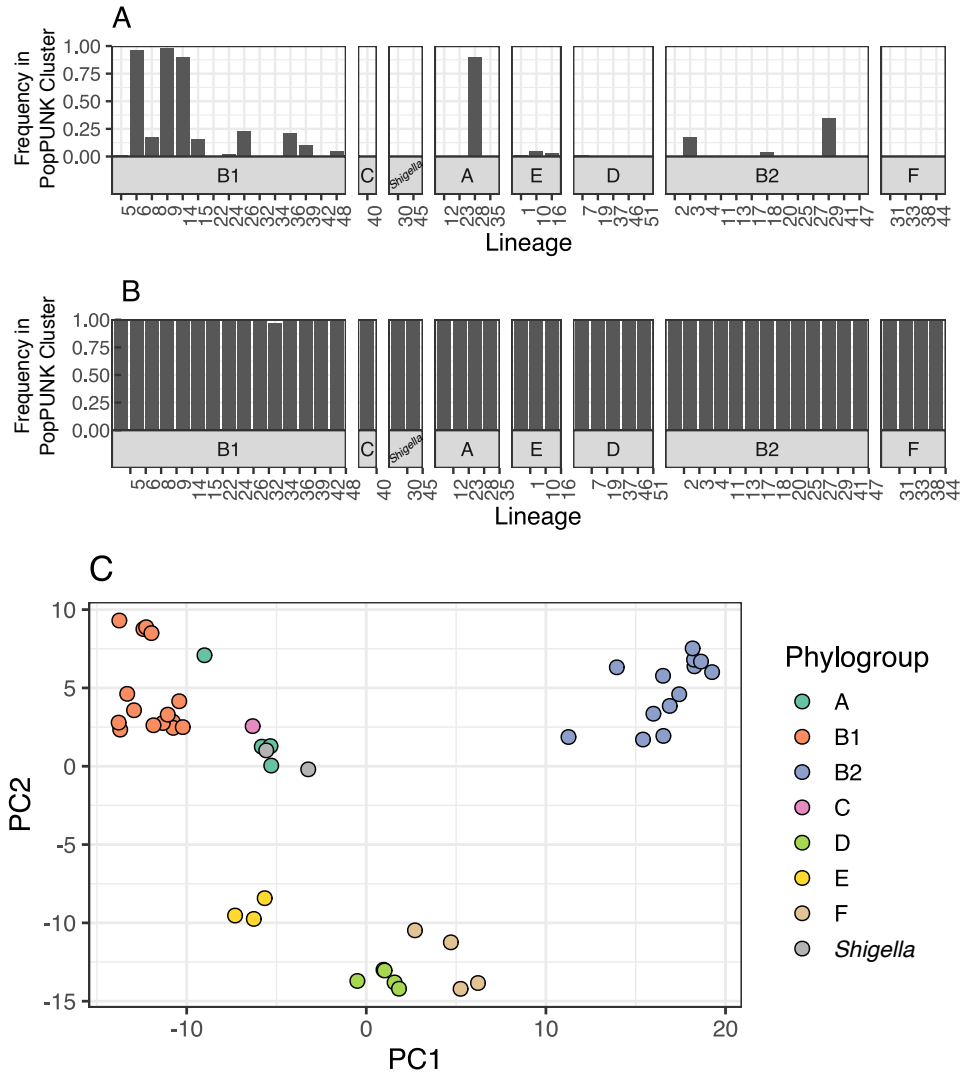

Figure 4.14: Gene frequencies across the PopPUNK Clusters. A,B *Examples of the frequencies of two genes across the 47 PopPUNK Cluster, stratified by phylogroup.* intA *(**A**) is present in some PopPUNK Clusters and is found in different frequencies within them.* wzyE *(**B**) is core across all clusters. **C** PCA plot of the gene frequencies across all clusters.*

4.5 Discussion

The process of building and processing a high-quality dataset of thousands of *E. coli* genomes was described, along with the properties of the lineages that are present within the dataset and their gene (predicted CDS) content. The construction of this collection presented challenges in data accessibility, the scalability of existing tools and general biases in available sequencing data.

Aggregating data from diverse sources along with their associated metadata was a time-consuming effort. Genome identifiers and data formats across publications and databases do not always match leading to many conversions which are error prone and require knowledge of programming. In addition, computational resources are required in order to apply thousands of assembly and annotation calculations. These are all limiting factors to research. This emphasises the need to build new resources which maintain high quality genome collections where users would more easily be able to both retrieve and apply analysis on large collections. Without such resources, we have a mountain of information that is on the one hand available, but on the other hand practically unusable. Enterobase has proved to be one of these valuable resources, collating genomic data, providing assemblies and complete metadata tables for all genomes [93,400]. However, Enterobase currently only includes seven species.

The collection we obtained is biased towards *E. coli* lineages which have clinical significance. Not only that, the vast majority of genomes were available from Europe and North America, such that the pathotypes comprising the dataset are those which predominantly affect these areas. This was exacerbated by the fact that Enterobase's policy on data usage was ambiguous regarding the correct use of genomes which had been uploaded to public databases and have not yet published (or it is hard to confirm if they had been published). In the analysis presented here all genomes which were not taken directly from publications or from institutions from which approval was acquired were removed. This led to the removal of thousands of genomes. Finally, in the final collection, lineages or PopPUNK Clusters which had fewer than 20 isolates were also removed. Of the 1,185 PopPUNK Clusters, only 50 remained. This emphasises our lack of understanding of the true diversity of *E. coli* as a species. Hence, sampling should be increased in under-represented areas in the world as well as sampling of non-clinical isolates.

Existing tools were often designed to handle smaller collections or were not suitable for the analysis of a biased and diverse collection. Division of the dataset into groups of closely related isolates had been applied before when analysing diverse collections [408]. Indeed, Roary was

designed to define the pan-genome of groups of closely related isolates, and thus was suitable when investigating the pan-genome of each PopPUNK Cluster [305]. However, an option to merge results of multiple pan-genome analyses had not been implemented and hence was built in this thesis. Additionally, Prokka, a commonly annotation tool, was not originally designed for genome comparison but rather for the annotation of a single genome [293]. A modified version of Prokka needed to be designed in order to remove artefacts when comparing multiple genomes. With more genomes, new methods need to be designed that are scalable when analysing diverse and large datasets.

Biological differences between the PopPUNK Clusters (lineages) were revealed from the initial investigation presented in this chapter. There were clear differences in the genome size between different phylogroups and PopPUNK Clusters. Higher variability in genome size with a phylogroup or PopPUNK Cluster could be an indication of higher rates of gene gain and loss within that cluster, as observed in phylogroup D. A larger genome size may also help to equip a lineage to survive in a range of niches as observed for PopPUNK Clusters of phylogroups E, F and B1 [4] (Figure 4.10A). Considering the major discrepancies in genome size between PopPUNK Clusters, it is interesting that the size of the core-genome across all the clusters is stable. This suggests that within a closely related group of genomes there is a specific number of genes, approximately 4,000 genes, that are required to define a lineage of closely related isolates (Figure 4.10B). The number of rare genes in a pan-genome was dependent on the cluster size, suggesting that the pan-genome of all lineages is open and is driven by continuous discovery of rare variants. A PCA plot of the gene frequencies as extracted from the complete dataset suggests that the phylogeny is driving the differences in gene content between the PopPUNK clusters. Questions regarding the distributions of different genes and the levels of gene sharing between the PopPUNK Clusters are further examined in Chapter 5 of this thesis.

5 Redefining the *E. coli* pan-genome reveals new patterns of gene gain/loss and gene sharing between lineages

5.1 Introduction

HGT is common in *E. coli* and is a major contributor to resistance and pathogenicity. *E. coli* has a high plasmid load, with many resistance genes present on these plasmids [4]. The virulence genes which are used as markers to identify the different *E. coli* pathotypes are also horizontally transmitted, either by plasmids, phage or other MGEs [102,221–223]. Additionally, recombination rates have been estimated to be high in *E. coli* [11,77,212,213]. All of the above emphasise the importance of HGT to the lifestyle and pathogenicity of *E. coli* (See Section 1.2.4 of Introduction for more details).

Genome size, plasmid load and recombination rates, along with rates of gene gain and loss, have all been shown to differ across *E. coli* lineages and phylogroups [4,77,92,212]. Indeed, there are differences in the distribution of the pathotypes across the phylogroups and it has been shown that particular genetic backgrounds are required for the acquisition of specific virulence factors [409,410]. Phylogroup F, B2 and D predominantly contain ExPEC isolates whereas phylogroups B1 and E predominantly contain diarrheal *E. coli* pathotypes (See Section 4.4.47, Figure 4.12). Phylogroup A contains isolates from the different *E. coli* pathotypes and has been termed a “generalist” phylogroup [411]. Concordantly, phylogroup A, as well as C, have been estimated to have high rates of HGT with high rates of gene gain/loss and high recombination rates [77,92,212,213]. Conversely, reduced recombination rates were estimated within the global MDR ExPEC lineage, ST131 of phylogroup B2 and the common EHEC lineage ST11/O157:H7 of phylogroup E, suggesting a clonal expansion of these lineages due to their clinical success [213].

These existing studies examining HGT in *E. coli* were mostly focused on high-level descriptions of the relationships between the phylogroups and have not looked at the resolution of specific *E. coli* lineages [92,212]. Even more, these studies have mostly considered only the core genome when estimating recombination rates [77,213], or otherwise, when measuring dynamics of gene gain/loss dynamics or gene sharing, have treated all genes of the gene pool equally [11,92,212]. These approaches are likely to mask particular signals

in the data. When considering only the core genome, the added information of the accessory genome, which represents the main fraction of the gene pool which undergoes HGT, is entirely missing in the analysis. When treating all genes equally in gene gain/loss or gene sharing calculations, rare events would be lost in the background. For instance, if 90% of genes are shared according to phylogenetic relatedness whereas only 10% are not, the signal for the unique 10% would not be observed when events are summed across the entire gene pool. Therefore, a higher resolution approach needs to be applied to understand the dynamics of different genes and how these dynamics differ across lineages.

In the previous chapter, a high-quality collection of *E. coli* genomes was built and the lineages of the collection, termed PopPUNK Clusters, were defined and characterised for their resistance and pathogenic profiles. The described dataset is novel in its resolution as it includes 47 well-characterised lineages (PopPUNK Clusters) with multiple representatives, and the frequency of each gene of the gene pool within each PopPUNK Cluster is known. This dataset provides the ability to identify different types of genes in the *E. coli* gene pool based on their distribution across the 47 lineages, and to unravel the differences between these lineages.

5.2 Aims

The work presented in this Chapter is a novel approach to classifying and analysing the patterns of gene sharing and gene gain and loss in the collection of 7,500 *E. coli* isolates presented in Chapter 4. The specific aims of this chapter were:

- Define a novel approach for describing the *E. coli* pan-genome.
- Unravel the properties of genes from the newly defined gene-classes in terms of their function and dynamics of gain and loss.
- Understand the differences between the PopPUNK Clusters in terms of their gene content and the levels of gene sharing between them.

5.3 Methods

5.3.1 Gene classification into “occurrence classes”

The genes were classified into “occurrence classes” based on their distribution patterns in the dataset. Each gene was assigned to an occurrence class based on its frequency within genomes belonging to the same phylogenetic clusters, termed PopPUNK Clusters. Within each PopPUNK Cluster, a gene was defined as “core” if it was present in more than 95% of

the isolates of that cluster, “intermediate” if present in 15% to 95% of isolates of the cluster, and “rare” if present in up to 15% of the isolates of the cluster. Three main occurrence classes, “Core”, “Intermediate” and “Rare”, contained all the genes that were always observed as being “core”, “intermediate” or “rare” respectively across all PopPUNK Clusters in which they were present. However, within these four occurrence classes, whilst the frequency was maintained within a cluster, genes were seen to be “core”, “intermediate” or “rare” across different numbers of clusters. Hence, to capture the distribution of all genes these occurrence classes were further subdivided into a total of eleven subclasses based on the number of PopPUNK Clusters in which a gene was observed and the frequency of that gene within those clusters (Figure 5.1).

“Dataset core” genes were present and “core” in all PopPUNK Clusters. “Multi-cluster core”, “multi-cluster intermediate” and “multi-cluster rare” genes were present in multiple PopPUNK Clusters in their respective frequencies. “Cluster specific core”, “Cluster specific intermediate” and “Cluster specific rare” genes were present only in one PopPUNK Cluster in their respective frequencies. The final main occurrence class “Varied” included all the genes which were observed as either combination of “core”, “intermediate” or “rare” across multiple PopPUNK Clusters. These combinations were “core, intermediate and rare”, “core and intermediate”, “core and rare” and “intermediate and rare” (Figure 5.1).

5.3.2 Measuring the genetic composition of each PopPUNK Cluster

The number of genes from each of the eleven occurrence classes was counted in each of the 7,693 *E. coli* genomes remaining in the collection described in Section 4.4.6. The mean number of genes and the standard deviation of the number of genes from each occurrence class was calculated per PopPUNK Cluster using built-in R functions. To measure the genetic composition of a typical *E. coli* genome within our dataset, the mean and standard deviations were calculated on the mean counts of all the 47 PopPUNK Clusters.

5.3.3 Phylogenetic analysis

5.3.3.1 Phylogenetic tree construction

A representative sequence from all 47 PopPUNK Clusters was chosen using Treemmer [392]. Treemer greedily prunes leaves off the phylogeny by choosing a random leaf from the two most closely related pairs of leaves in every iteration, until the desired number of leaves in the tree is reached. The core gene alignment of the 47 selected isolates was generated using

Roary [305], and a tree from the SNPs, taken using SNP-sites [332] (v2.3.2), was constructed using RaXML (v8.2.8) using a GTR+gamma model with 100 bootstrap replicates [282].

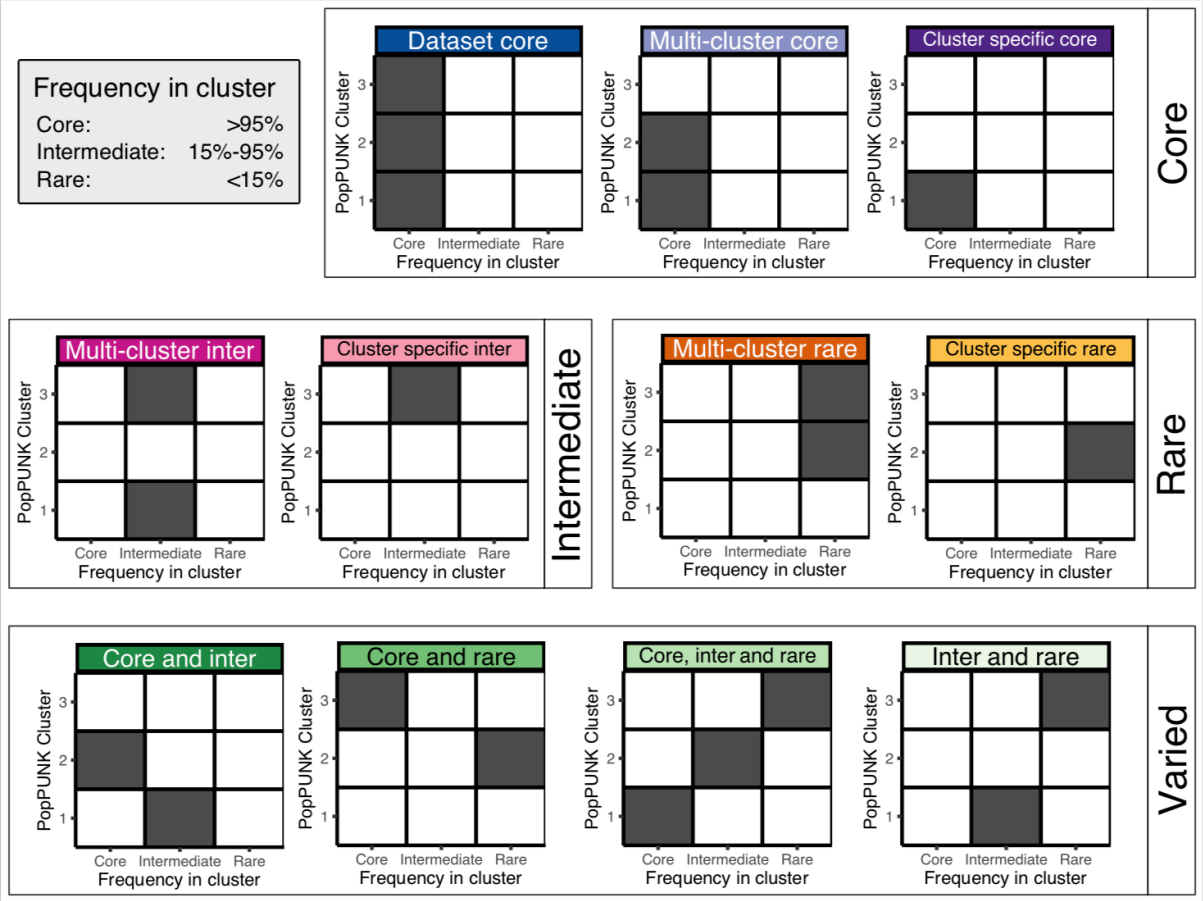

Figure 5.1: Gene classification into occurrence classes*. The figure presents a hypothetical example of comparing a total of 3 PopPUNK Clusters written on the y-axis. The x-axis represents the frequency of the gene in each of the three clusters being compared. A gene is considered "core" in a cluster if it was present in >95% of isolates of the cluster, "intermediate" if it was present in 15%-95% of the the isolates of the cluster, and "rare" if present in <15% of isolates of the cluster. Each panel is an example of a gene from the given occurrence class. A dark square indicates the gene is present in the cluster and the frequency of that gene in the cluster. As there are three clusters, each gene can be observed in any combination of frequencies across the three clusters. "Core" genes were observed in core frequencies in all (dataset core), some (multi-cluster core) or one (cluster specific core) cluster. "Intermediate" genes were observed in intermediate frequencies in some (multi-cluster intermediate) or one (cluster specific intermediate) cluster. "Rare" genes were observed in rare frequencies in some (multi-cluster rare) or one (cluster specific rare) cluster. "Varied" genes were observed in different frequencies across multiple clusters. For instance, the "Core and intermediate" gene*

presented is core in cluster 2 and rare in cluster 1 (and absent in 3). The "Core and rare" gene is core in cluster 3 and rare in cluster 2 (and absent in 1) etc.

5.3.3.2 Phylogenetic distance calculations

The phylogenetic distance between every two PopPUNK Clusters was measured as the patristic distance using the function 'cophenetic' from the R package APE (v5.3) [395]. The patristic distance is the sum of the total distance between two leaves of the tree, which represent the PopPUNK Clusters in this thesis, and hence summarises the total genetic change in the core gene alignment represented in the tree.

5.3.3.3 Ancestral state reconstruction

The leaves or tips of the phylogenetic tree constructed in Section 5.3.3.1 represent the 47 PopPUNK Clusters. Presence of a gene in a PopPUNK Cluster (tree leaf) was defined as the gene being observed at least once in at least one isolate of the PopPUNK Cluster. The presence or absence of a gene in an ancestral node, i.e. an internal node, was determined using accelerated transformation (ACCTRAN) reconstruction implemented in R [412]. ACCTRAN is a maximum parsimony-based approach which minimises the number of transition events on the tree (from absence to presence and vice versa) while preferring changes along tree branches closer to the root of the tree.

5.3.3.4 Counting gain and loss events

Gain and loss events were counted based on the results of the ancestral state reconstruction. If there was a change from absence to presence from an ancestor to a child along a branch in the phylogeny, a gain event was counted. If there was a change from presence to absence a loss event was counted. The total number of gain and loss events was counted for each gene as well as on each branch for all occurrence classes.

5.3.4 Functional assignment of COG categories

The predicted function and COG category of each gene cluster were assigned using eggNOG-mapper (1.0.3) on the representative sequence of each of the gene clusters [413]. Diamond was used for a fast-local protein alignment of the representative sequences against the eggNOG protein database (implemented within eggNOG-mapper). The COG (Clusters of Orthologous Groups) classification scheme comprises 22 COG categories which are broadly divided into functions relating to cellular processes and signaling, information storage and processing, metabolism and genes which are poorly categorised [414]. When no match was found in the eggNOG database, the genes were marked as "?" in their COG category.

Sub-sentences of all lengths were extracted from each of the functional predictions for each gene cluster using the function “combinations” from the python package “itertools”, while ignoring common words. For instance, for the functional prediction “atp-binding component of a transport system”, the words “of”, “a” and “system” were ignored, and the extracted sub-sentences were “atp-binding component”, “atp-binding component transport” and “component transport”. The number of times each sub-sentence appeared in each occurrence class was counted. Overlapping sub-sentences which only had a difference of 3 or smaller in their total counts per occurrence class were merged in the final count to include only the longer sub-sentence. For instance, if “atp-binding component transport” was counted 100 times and “atp-binding component” was counted 103 times, the final count would only include the longer sub-sentence “atp-binding component transport” with a count of 100.

5.3.5 Identifying gene variants

The function makeblastdb from the Blast+ package (v2.9) was used to construct a database from the 50,039 genes of the *E. coli* pan-genome taken from Chapter 4 of this thesis [285,321]. Blastp was used to apply a pairwise all-against-all comparison of all the protein sequences. If two proteins shared more than 95% sequence identity over 95% of the total length of the shorter sequence, they were considered “partner genes”, with one being the “shorter variant” and the other the “longer variant”.

5.3.6 Gene property calculations

The length of each gene cluster was calculated as the mean length of all the members of that gene cluster. The GC content was calculated using Biopython (v1.72) on all the members of a gene cluster and the mean was taken as the value for that gene cluster. The fraction of members in a gene cluster that had ATG as their start codon was measured as the “ATG fraction”. If an alternative start codon was used in more than 50% of the members of a gene cluster then that cluster was considered as starting with an alternative start codon. The contig length was calculated for all the members of a gene cluster and the mean was calculated across all members.

5.3.7 Statistical analysis

Statistical analyses were performed in R (v3.3+). Ape (v5.3) [395] and ggtree (v1.16.6) [396] were used for phylogenetic analysis and visualisation. ggplot2 was used for all plotting [360].

5.4 Results

5.4.1 A novel approach for examining the *E. coli* pan-genome

In a standard pan-genome analysis, genes are classified into four categories: core, soft-core, intermediate and rare. These definitions are based on the frequency of the genes in the dataset. For instance, the default settings in Roary are that genes found in over 99% of the genomes are “core”, between 95% and 99% “soft-core”, between 15% and 95% “intermediate” and fewer than 15% “rare” [305]. In Section 4.4.4.9 of this thesis, these definitions were used to describe the pan-genomes of each of the 47 PopPUNK Clusters individually. Roary was originally designed for a pan-genome analysis of a single *Salmonella enterica* serovar (Typhi), thus the default approach used in Section 4.4.4.9 was valid for a pan-genome analysis on each PopPUNK Cluster which represents a group of closely related isolates. When expanding the pan-genome analysis to examine the pan-genome of an entire species, which in this case includes 47 different PopPUNK Clusters, new definitions needed to be established. Hence, a new set of rules was defined to classify the genes into four broad “occurrence classes”: “core”, “intermediate”, “rare” and “varied” genes. These four occurrence classes could be further subdivided into eleven sub-classes as detailed below. These definitions were based on the number of PopPUNK Clusters in which a gene was present (1 to 47), and the frequency of the gene in the clusters in which it was present (Figure 5.1).

Core genes were always observed in high frequencies (>95%) in one or multiple PopPUNK Clusters (Figure 5.1). These genes represented 9% (4,998/50,039) of the *E. coli* pan-genome (Figure 5.2A). Core genes included 1,426 genes (3% of all genes) which are the “dataset core” genes as they were core in all 47 of 47 PopPUNK Clusters (Figure 5.2B,C, 5.1). On the other side of the spectrum, there were 2,040 genes (4% of all genes) which were “cluster specific core” genes as they were core in a single PopPUNK Cluster. A set of 1,532 genes (3% of all genes) were defined as “multi-cluster core” as they were core to a subset of the PopPUNK Clusters (2-45 PopPUNK Clusters).

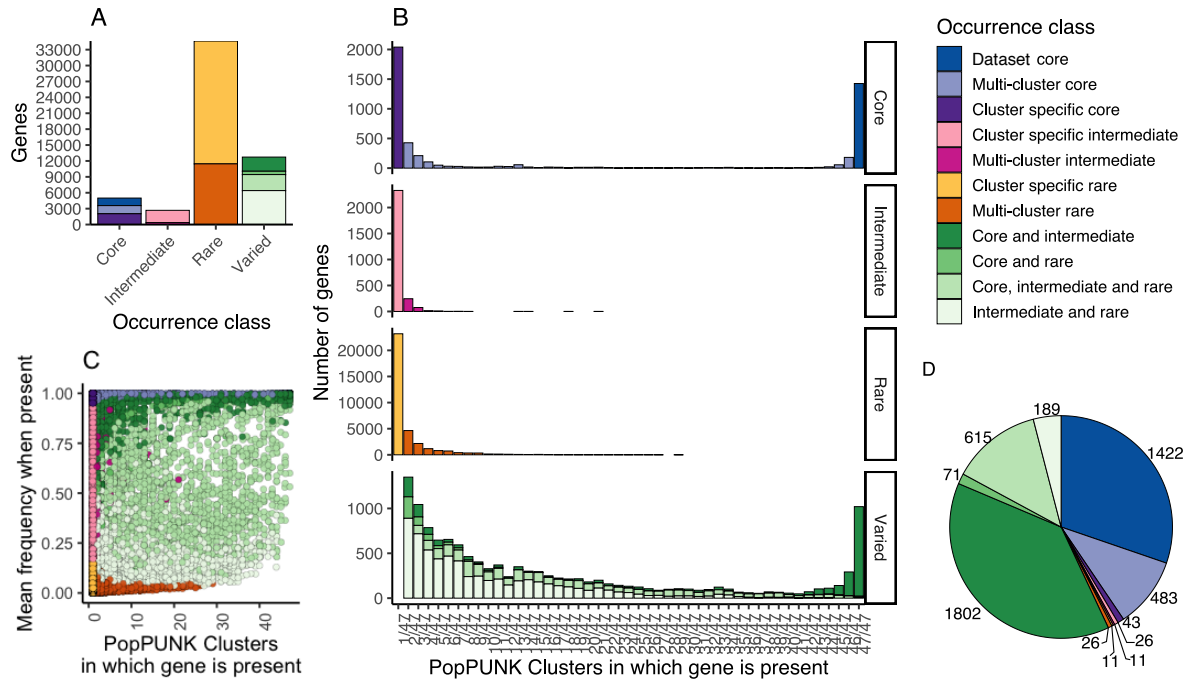

Figure 5.2: Distribution of the* E. coli *gene-pool based on the rules defined. A *Number of genes from each of the occurrence classes.* ***B*** *Distribution of the number of genes in each occurrence class relative to the number of PopPUNK Clusters in which they were found.* ***C*** *Mean frequency of each gene in the PopPUNK Clusters in which it was observed, plotted against the number of PopPUNK Clusters it was observed in, coloured by occurrence class.* ***D*** *The relative abundance and count of genes from each of the occurrence classes in a single representative* E. coli *genome in our dataset.*

Intermediate genes, representing 5% of all genes, were always observed in intermediate frequencies (15%-95%) in one or multiple PopPUNK Clusters (Figure 5.1, 5.2A). 87% of these genes (2,329/2,685) were only observed in a single PopPUNK Cluster and were termed “cluster-specific intermediate” genes (Figure 5.2B,C). The remaining intermediate frequency genes (356) were termed “multi-cluster intermediate”. These were mostly shared between a maximum of five PopPUNK Clusters (97%, 346/356) and their mean frequency within those clusters ranged from 16% to 94% of isolates, representing the full range of possible frequencies for intermediate genes. There were four genes (1%, 4/356) which were observed in intermediate frequencies in more the 10 PopPUNK Clusters. One gene was of particular interest as it was observed in 20 PopPUNK Clusters and its mean frequency across the these clusters was 0.57, appearing be a truly intermediate frequency gene (Figure 5.2C). A closer examination of the precise frequencies in which this gene was observed across the 20 clusters confirmed that it was indeed observed in 30-70% of isolates in all the clusters, with most PopPUNK Clusters having 50-60% of isolates possessing this gene. Further analysis on the

sequence of this gene revealed that this is a a short protein, only 53 aa long, which could not been assigned to any known function using functional annotation tools.

Rares genes were always observed in low frequencies (<15%) in one or multiple PopPUNK Clusters (Figure 5.1). This occurrence class represented the largest fraction of the entire gene pool consisting of a total of 34,624 genes, representing 63% (34,624/55,039) of the entire gene pool (Figure 5.2A). Of these, 67% were “cluster specific rare” genes (23,175/34,624) as they were observed only in a single PopPUNK Cluster (Figure 5.2B,C). The remaining “rare” genes were observed in multiple PopPUNK Clusters, termed “multi-cluster rare”. 76% (8,800/11,449) of these were observed in five PopPUNK Clusters or fewer. There were 651 (5%) genes which were observed in rare frequencies across 10 PopPUNK Clusters or more, i.e. rare genes across multiple PopPUNK clusters were more common than intermediate genes across multiple clusters.

Varied genes were observed in different frequencies across multiple PopPUNK Clusters (Figure 5.1). These genes represented 23% of the gene pool (12,732/55,039) (Figure 5.2A). These were further divided depending on the precise combination of frequencies in which they were found: “Core and intermediate”, “Core, intermediate and rare”, “Core and rare” or “Intermediate and rare” (Figure 5.1). Varied genes which were observed in more PopPUNK Clusters were more commonly observed in higher frequencies within those clusters and thus belonged to the group of “Core and intermediate" genes (Figure 5.2B,C). On the other hand, varied genes which were observed in fewer PopPUNK Clusters were more commonly observed in low frequencies within those clusters and thus belonged to the group of “Intermediate and rare" varied genes (Figure 5.2B,C).

5.4.2 The typical composition of an *E. coli* genome

A typical *E. coli* genome contained 1,422±4 genes (~30%) “dataset core” genes (core across the entire dataset) (Figure 5.2D; see Section 5.3.2). There were 483±66 (~10%) “multi-cluster core” genes which were core to a subset of the population and 43±55 (1-2%) genes which were “cluster specific core” genes, present and core only in a single PopPUNK Cluster (Figure 5.2D). A typical genome also contained 11±7 (~0.3%) “multi-cluster intermediate” and 26±23 (0.5-1%) “cluster specific intermediate” genes (Figure 5.2D). Similarly, there were 26±11 (~0.5%) “multi-cluster rare” genes 5.2D) and 11±9 (~0.3%) “cluster specific rare” genes in each genome (Figure 5.2D). Although the “rare” and “intermediate” genes made up more than 60% of the entire gene pool (34,543/55,039), they each represented fewer than 1% of the genes within a single isolate (Figure 5.2D). The “varied” genes represented approximately

60% of all the genes in a typical *E. coli* genome (Figure 5.2D). Most of these were “core and intermediate” genes (1802±87, ~40%) (Figure 5.2D). Additionally, each genome contained 71±13 (1-2%) “core and rare” genes, 614±116 (10-15%) “core, intermediate and rare” genes, and 189±51 (3-5%) “intermediate and rare” genes.

5.4.3 Rates of gene gain and loss differ across the occurrence classes

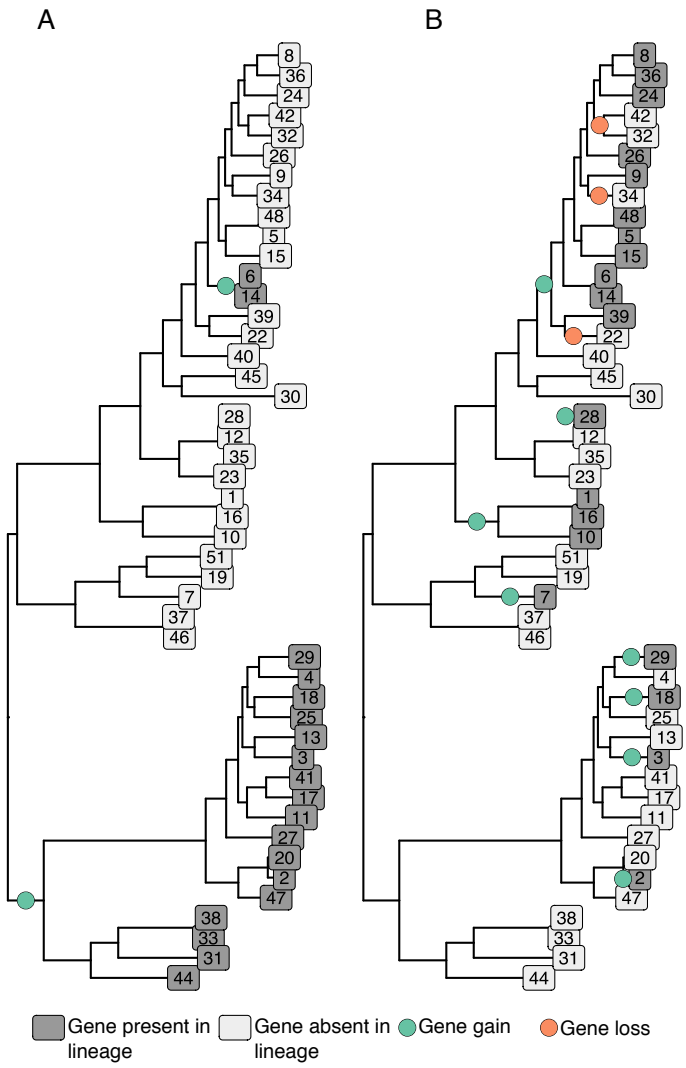

Figure 5.3: Example of the distribution patterns of two genes, along with the number of gain and loss events required to explain their distribution across the tree tips. *A gene is defined as present in a tip (dark grey) if at least one genome of the lineage had the gene. Gain (green circle) and loss (red circle) were estimated using ancestral state reconstruction.* ***A*** *A “multi-cluster core” gene which is associated with two clades and required only 2 gain events to explain its distribution.* ***B*** *An “intermediate and rare” gene which was not clade associated required 8 gain and 3 loss events to explain its distribution along the tree tips.*

The presence and absence patterns of genes which were present across multiple PopPUNK Clusters were used to count the number of gain and loss events estimated to have occurred along the tree branches. This was achieved using a parsimony-based ancestral state reconstruction approach to infer the minimum number of gain and loss events required to explain the distribution of a gene on the tree tips. (See Sections 5.3.3.3-4). For instance, if a gene was present in only two clades (regardless of its frequency when present), its distribution along the tree tips could be explained by two gain events on two branches (Figure 5.3A). If a gene was distributed across

the tree tips with no clear pattern, many more gain or loss events were required to explain its distribution on the tree tips (Figure 5.3B).

The number of gain and loss events which occurred for each gene varied across the occurrence classes (Figure 5.4A). For comparison, the specific combinations of gain and loss events across all genes for each of the occurrence classes are summarised in Figure 5.4B-H. These will be referred to in the following sections. Note that due to the method by which genes from the different PopPUNK Clusters were grouped as described in Section 4.3.8 of this thesis, gene loss could indicate either complete loss, truncation by more than 20% of the gene length or diversification beyond the 95% sequence identity threshold used to group genes together.

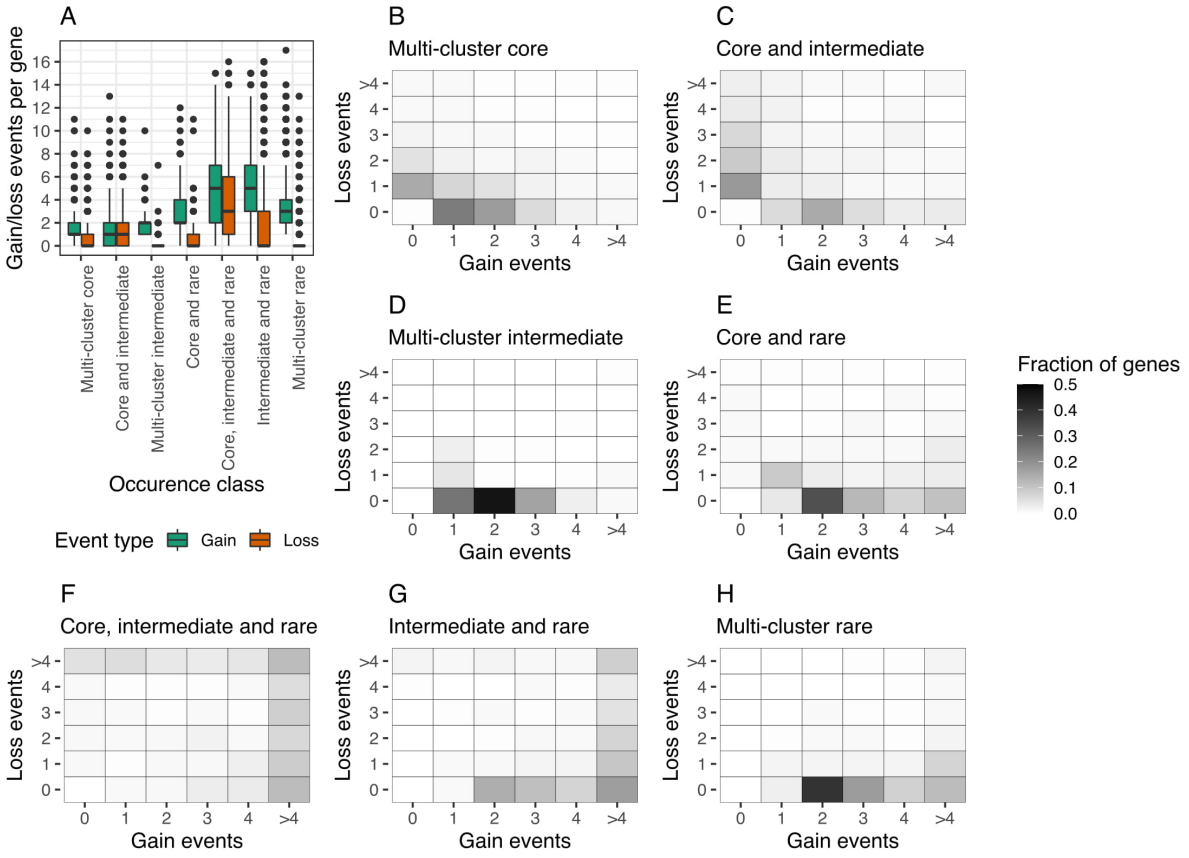

Figure 5.4: Gain and loss events per gene. A *Number of gain and loss events per gene stratified by occurrence class.* ***B-H*** *Fraction of genes which have undergone specific combinations of gain and loss events for each occurrence class. The shade of each square indicates the fraction of genes from the occurrence class which have undergone the specific combination of gain and loss events.*

5.4.4 “Multi-cluster core” genes represent the shifts in core genome of *E. coli* clades

The median number of gain and loss events estimated for “multi-cluster core” genes was a single gain event and a no loss events (Figure 5.4A). The majority (68%) of the presence and absence patterns of these genes could be explained by up to two gain or loss events along the tree branches (Figure 5.4B). Most prominently, a single gain event and no loss events was observed for 24% of these genes, i.e. these genes were gained in a single point in time and were fixed within the lineages downstream from the point of introduction. On the other hand, 15% of these genes were estimated to have been lost in a single event that led to the absence of the gene from a subset of the PopPUNK Clusters. While some genes were estimated to have been gained and lost on more occasions, these were the exception rather than the rule for this occurrence class (Figure 5.4A,B).

The number of gain and loss events predicted to have occurred on each branch were counted (Figure 5.5A,B). Gain and loss events of “multi-cluster core” genes most commonly occurred along the internal branches which define the phylogroups (Figure 5.5A, B, C). A large number of gain events occurred on the branches leading to phylogroups B2 (104 gain events), E (66), F (52) and two clades of phylogroup D (90 and 64) (Figure 5.5A,C). Two PopPUNK Clusters within phylogroup E, Clusters 1 and 16, were also estimated to undergo a large number of gene gain events (97). The branches leading to the clades of phylogroups A, *Shigella*, B1 and C were not estimated to have undergone a large number of gene gain events. Phylogroup B2 was the only phylogroup which had undergone excessive gene loss in addition to gene gain (52 loss events) (Figure 5.5B,C). Otherwise, gene loss occurred most commonly along the tree tips (Figure 5.5C). Most prominently, PopPUNK Clusters 30 and 45 which represent *S. sonnei* and *S. flexneri* respectively, as well as PopPUNK Cluster 18 which has not been assigned to any of the phylogroups, have undergone the largest number of recent loss events (90, 52 and 65) (Figure 5.5B, D).

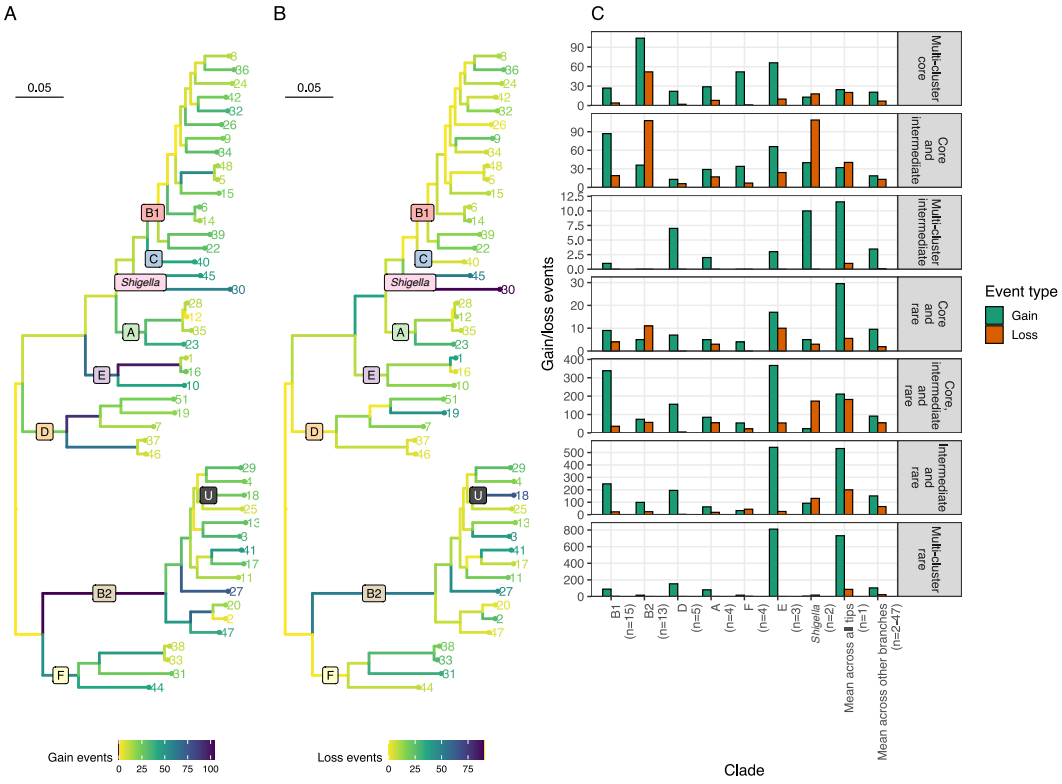

Figure 5.5: Gain and loss events per branch. A,B *Example for "multi-cluster core" genes on the precise counts of "gain" and "loss" events across all genes of this occurrence class predicted to have occurred on each branch. Darker branches indicate a larger number of events occurring on the branch.* ***C*** *Summary of the total number of gain and loss events on key branches for all the occurrence classes. The top panel for the "multi-cluster core" genes summarises panels* ***A*** *and* ***B****.*

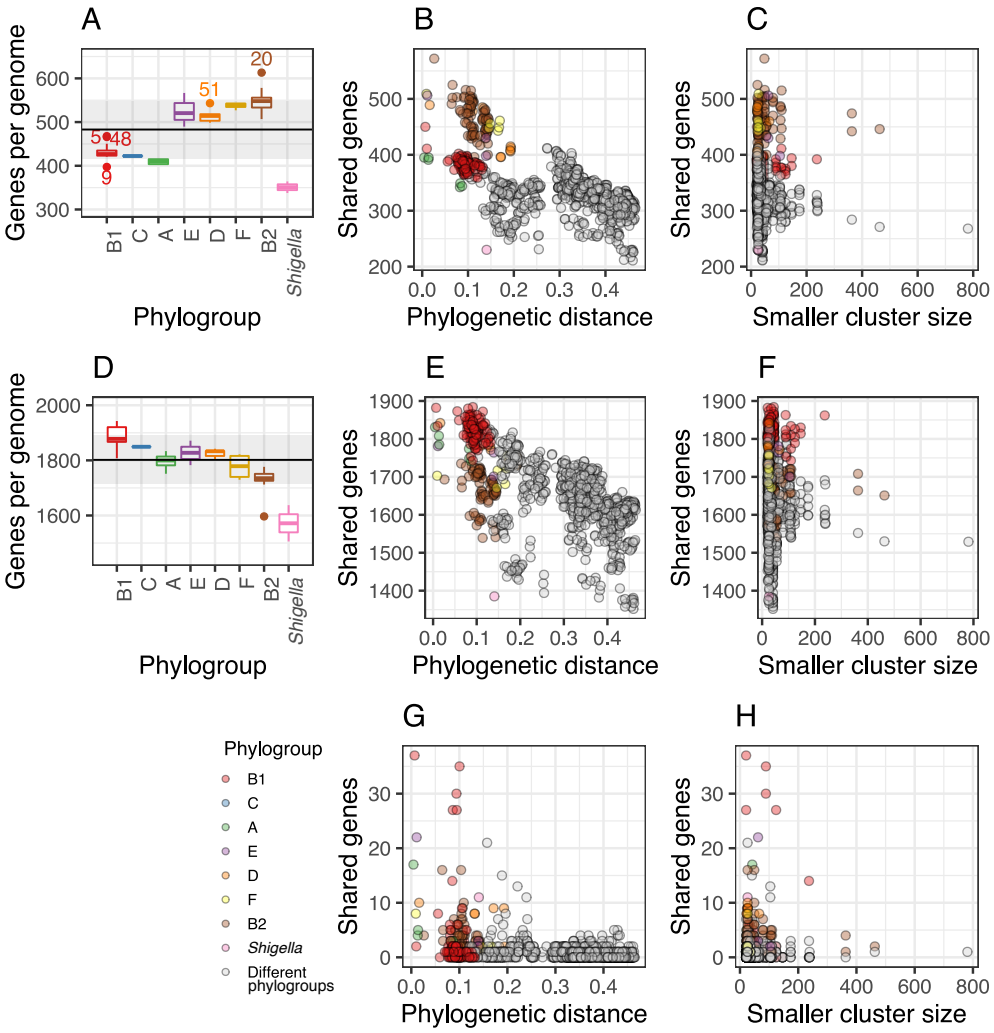

Figure 5.6: Properties of high frequency genes in the* E. coli *dataset. A,D *Number of "multi-cluster core" genes (**A**) and "core and intermediate" genes (**D**) per genome in each of the 47 PopPUNK Clusters, grouped by phylogroup.* ***B, E, G*** *Relationship between the number of genes shared between every two PopPUNK Clusters and phylogenetic distance between them for "multi-cluster core" genes (**B**), "core and intermediate" genes (**E**) and "multi-cluster intermediate" genes (**G**). Coloured dots indicate that the two PopPUNK Clusters being compared are from the same phylogroup, whereas gray dots indicate that the two clusters being compared are from different phylogroups.* ***C, F, H*** *Relationship between the number of genes shared between every two PopPUNK Clusters and the size of the smaller PopPUNK Cluster of the two being compared for "multi-cluster core" genes (**C**), "core and intermediate" genes (**F**) and "multi-cluster intermediate" genes (**H**).*

In agreement with the above, while the mean number of "multi-cluster core" genes was 483 per genome, isolates belonging to phylogroups B1, C, and A tended to have ~400 multi-cluster core genes per genome compared with ~500 for those belonging to phylogroups E, D, F and

B2 (Figure 5.6A). PopPUNK Clusters of *Shigella* spp. had the fewest number of "multi-cluster core" genes per genome with ~350 multi-cluster core genes per genome (Figure 5.6A).

The above analysis suggests that "multi-cluster core" genes represent the changes in the core genome between the clades. Accordingly, the number of "multi-cluster core" genes shared between every two PopPUNK Clusters was correlated negatively with the phylogenetic distance between them (linear regression, R^2=0.42, p<2e-16), i.e. two PopPUNK Clusters which were close phylogenetically shared more "multi-cluster core" genes (Figure 5.6B). There was no connection between the size of the two PopPUNK Clusters being compared and the number of "multi-cluster core" genes they shared (linear regression, R^2=0, p=0.51) (Figure 5.6C).

5.4.5 "Core and intermediate" represent the "soft-core" genome

The properties of the "core and intermediate" genes, which represented 40% of the genes in a single *E. coli* genome and 5% of the entire gene pool (Figure 5.2A,D), prove that these genes present similar distribution patterns, patterns of gain and loss and predicted functions to the defined "multi-cluster core" and "dataset core" genes.

59% of these genes (1,566/2,674) were observed in 40 PopPUNK Clusters or more, and in high frequencies within those clusters (Figure 5.2B,C). In fact, 37% of the "core and intermediate" genes were ubiquitous, i.e. they were present in 47 of 47 PopPUNK Clusters (Figure 5.2B). The median number of gain and loss events occurring per gene for "Core and intermediate" genes was a single gain event and a single loss event (Figure 5.4A). Similar to the "multi-cluster core" genes, 51% of the presence and absence patterns of these genes could be explained by up to two gain and loss events (Figure 5.4C). Gain events of "core and intermediate" genes were largest for the branches leading to phylogroups B1 and E (87 and 66) (Figure 5.5C). Rates of gene loss were generally higher in this occurrence class compared with the "multi-cluster core genes". Similarly, loss events predominantly occurred within *Shigella* and phylogroup B2 (Figure 5.5C). Indeed, these phylogenetic clusters had the lowest number of "core and intermediate" genes per genome relative to the other phylogroups whereas PopPUNK Clusters of Phylogroup B1 had the highest number of these genes per genome (Figure 5.6D).

"Core and intermediate" genes were also more commonly shared between closely related isolates (linear regression, R^2=0.39, p<2e-16) (Figure 5.6E), and there was no connection between the size of the two PopPUNK Clusters being compared and the number of core and

intermediate genes shared between them (linear regression, R^2=0.0007, p=0.18) (Figure 5.6F).

The distribution of predicted functions of this set of genes was similar to the predicted functions of the “dataset core” genes (Figure 5.7). COG categories were assigned to all the genes with eggNOG-mapper on the representative protein sequence of each gene cluster [413,414] (See Section 5.3.6). 34% of the “core and intermediate” genes were assigned to be involved in metabolism, similar to 40% of the “dataset core” genes. 14% and 13% were predicted to be involved in “information storage and processing” and “cellular processes and signalling” relative to 19% and 20% of the “dataset core” genes. Even more, the relative abundance of the specific COG categories was similar between the “dataset core” and the “core and intermediate” genes (Figure 5.7).

5.4.6 “Multi-cluster intermediate” genes are shared between closely related PopPUNK Clusters, but have different functional profiles to the “core” genes

In 89% of cases, “multi-cluster intermediate” genes were gained in 1-3 events and not lost (Figure 5.4D, 5.4C). Additionally, above a certain phylogenetic distance, the number of “multi-cluster intermediate” genes shared between every two PopPUNK Clusters drops to zero, meaning that these genes were only shared between closely related isolates (Figure 5.6G). Shared “multi-cluster intermediate” genes were only observed within PopPUNK Clusters which had fewer than 200 isolates (Figure 5.6H). These findings together suggest that these genes are confined to a phylogenetically close subset of the population, yet were gained multiple times within this subset. Unlike the “core and intermediate” genes, 77% “multi-cluster intermediate” genes were assigned a category of “poorly characterised” in their function prediction, and fewer than 1% were predicted to have a function related to cell metabolism (Figure 5.7). While these genes are shared between closely related PopPUNK Clusters as was observed for the “multi-cluster core” and the “core and intermediate” genes, they evidently differ in their functional profiles.

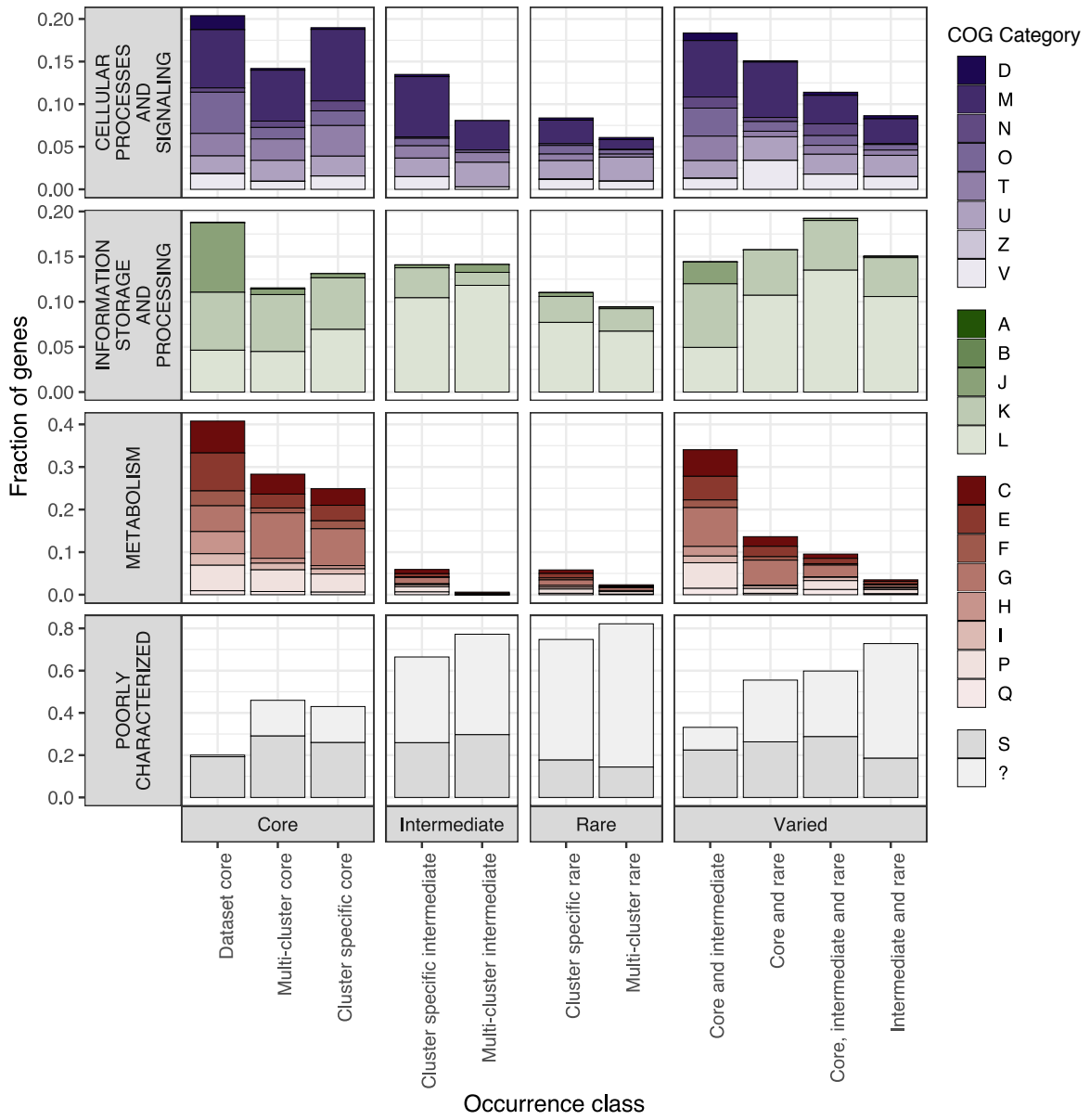

Figure 5.7: Fraction of genes from each occurrence class which were assigned each of the COG categories. *D (Cell cycle control, cell division, chromosome partitioning), M (Cell wall/membrane/envelope biogenesis), N (Cell motility), O (Post-translational modification, protein turnover, and chaperones), T (Signal transduction mechanisms), U (Intracellular trafficking, secretion, and vesicular transport), Z (Cytoskeleton), V (Defense mechanisms), A (RNA processing and modification), B (Chromatin structure and dynamics), J (Translation, ribosomal structure and biogenesis), K (Transcription), L (Replication, recombination and repair), C (Energy production and conversion), E (Amino acid transport and metabolism), F (Nucleotide transport and metabolism), G (Carbohydrate transport and metabolism), H (Coenzyme transport and metabolism), I (Lipid transport and metabolism), P (Inorganic ion transport and metabolism), Q (Secondary metabolites biosynthesis, transport, and catabolism), S (Function unknown) and "?" (unassigned).*

5.4.7 Low frequency genes are gained and lost at high rates, and their sharing is independent of the phylogeny

Shared low frequency genes include “multi-cluster rare”, “intermediate and rare” and “core, intermediate and rare” and “core and rare” genes as these were most commonly found in a small number of PopPUNK Clusters and in a low frequency within those clusters (Figure 5.2B,C). Unlike their high frequency counter-parts (“multi-cluster core”, “multi-cluster intermediate” and “core and intermediate” genes), the estimated number of gain and loss events predicted to have occurred for these occurrence classes was often estimated to be as high as four events and more (Figure 5.4A,E-H). “Multi-cluster rare” genes were not commonly lost as they were generally observed across a smaller number of PopPUNK Clusters and hence were mostly commonly gained 2-3 times along the tree tips (Figure 5.2B, 5.3H, 5.4C). Gain events of low frequency genes mostly occurred recently along the tree tips (Figure 5.5C). Phylogroup E was an exception which presented a large number of acquisition events of low frequency genes. A large number of gain events of “Core, intermediate and rare” genes were predicted to have occurred on the branch leading to Phylogroup B1.

The number of genes shared between every two PopPUNK Clusters for the “multi-cluster rare”, “intermediate and rare” and “core, intermediate and rare” genes did not correlate with the phylogenetic distance between the clusters (linear regression, R^2<0.03, Figure 5.8A-C). On the other hand, the number of shared genes was positively correlated with the size of the two PopPUNK Clusters being compared, with larger clusters sharing more genes (linear regression, R^2=[0.566,0.349,0.22], p<2.2e-16) (Figure 5.8D-F). This is because more genomes need to be sampled in order for the same low frequency gene to be observed in two PopPUNK Clusters. However, the number of genes shared plateaued after a particular PopPUNK Cluster size (Figure 5.8D-F). This number was smaller when the PopPUNK Clusters were from two different phylogroups, compared to when they were from the same phylogroup (Figure 5.8D-F).

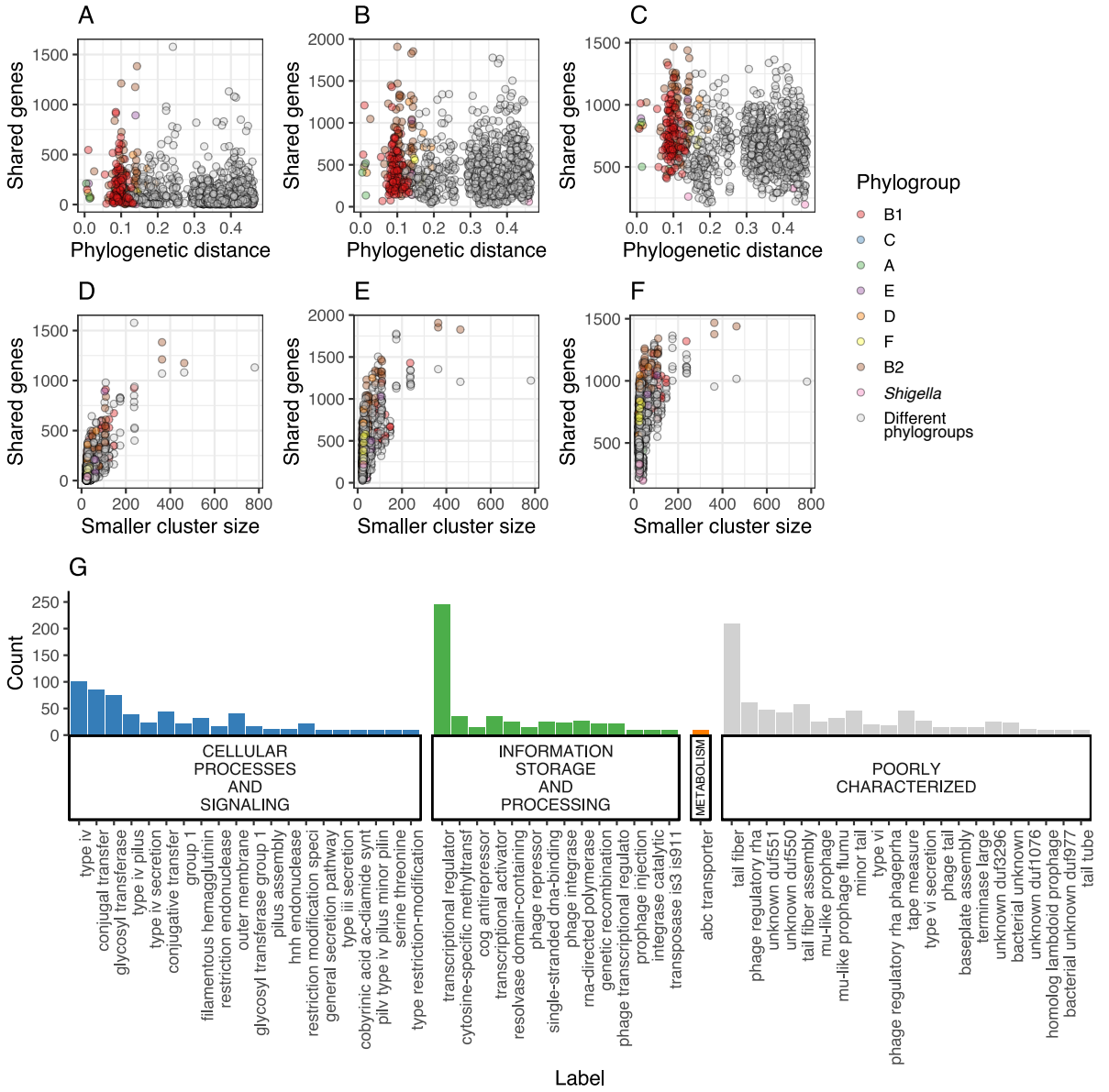

Figure 5.8: Properties of low frequency genes in the* E. coli *dataset*.** ***A-C *Relationship between the number of genes shared between every two PopPUNK Clusters and phylogenetic distance between for "multi-cluster rare" genes (**A**), "intermediate and rare" genes (**B**) and "core, intermediate and rare" genes (**C**). Coloured dots indicate that the two PopPUNK Clusters being compared are from the same phylogroup whereas gray dots indicate the two clusters being compared are from different phylogroups.* ***D-F*** *Relationship between the number of genes shared between every two PopPUNK Clusters and the size of the smaller PopPUNK Cluster of the two being compared between for "multi-cluster rare" genes (**D**), "intermediate and rare" genes (**E**) and "core, intermediate and rare" genes (**F**).* ***G*** *Most common phrases taken from the predicted functional annotations of the "multi-cluster rare", "intermediate and rare" and "core, intermediate and rare", divided into the four main COG categories.*

A large fraction of genes from these three gene categories were assigned a COG category of “Poorly Characterised” (Figure 5.7). The most common predicted terms for these genes were prophage related (Figure 5.8G). These included terms such as “tail fiber”, “baseplate assembly”, “terminase” and “Mu-like prophage”. Other common annotations in the other COG categories included “conjugal transfer”, “type IV pilus”, “restriction endo-nuclease”, “integrase catalytic”and “transposase”.

5.4.8 PopPUNK Clusters of broad host range lineage ST10 and MDR lineage ST410 share more low frequency genes with distantly related PopPUNK Clusters than expected

To explore the distribution of low frequency genes further I identified PopPUNK Clusters which share a large number of low frequency genes with other PopPUNK Clusters that are distantly related to them. The median number of low frequency genes each cluster shares with all other clusters that are distant from it (patristic distance higher than 0.4) was compared against the size of the cluster (Figure 5.9A). As expected, there was a linear relationship between the size of the PopPUNK Cluster and the median number of low frequency genes that a cluster shared with distantly related PopPUNK Clusters (log linear regression, R^2=0.547, p=2.965e-08). However, there were also a number of PopPUNK Clusters that shared more low frequency genes with distant PopPUNK Clusters than expected for their size. These include PopPUNK Clusters 12 and 40 (Figure 5.9A). Accordingly, the branches leading to these PopPUNK Clusters had been predicted to have undergone a large number of gain-events of “core, intermediate and rare” and “intermediate and rare” genes relative to the rest of the tips (Cluster 12: 682 and 1574 events, Cluster 40: 333 and 836 events, Tip-mean: 182 and 530 events, not shown). 78% of the isolates from PopPUNK Cluster 12 are of ST10, members of which are known to have a broad host-range. 30% of the isolates in PopPUNK Cluster 40 are from ST410 known as an MDR lineage and another 43% are from ST23. Multidrug resistance was common amongst the other PopPUNK Clusters which deviated from the expected number of shared genes, including PopPUNK Clusters 19, 26 and 34 with resistance observed to aminoglycoside, sulfonamides, beta-lactams and more (See Appendix E). Clusters 26 and 34 predominantly contained EPEC isolates from the GEMs collection (see Section 4.4.4.7).

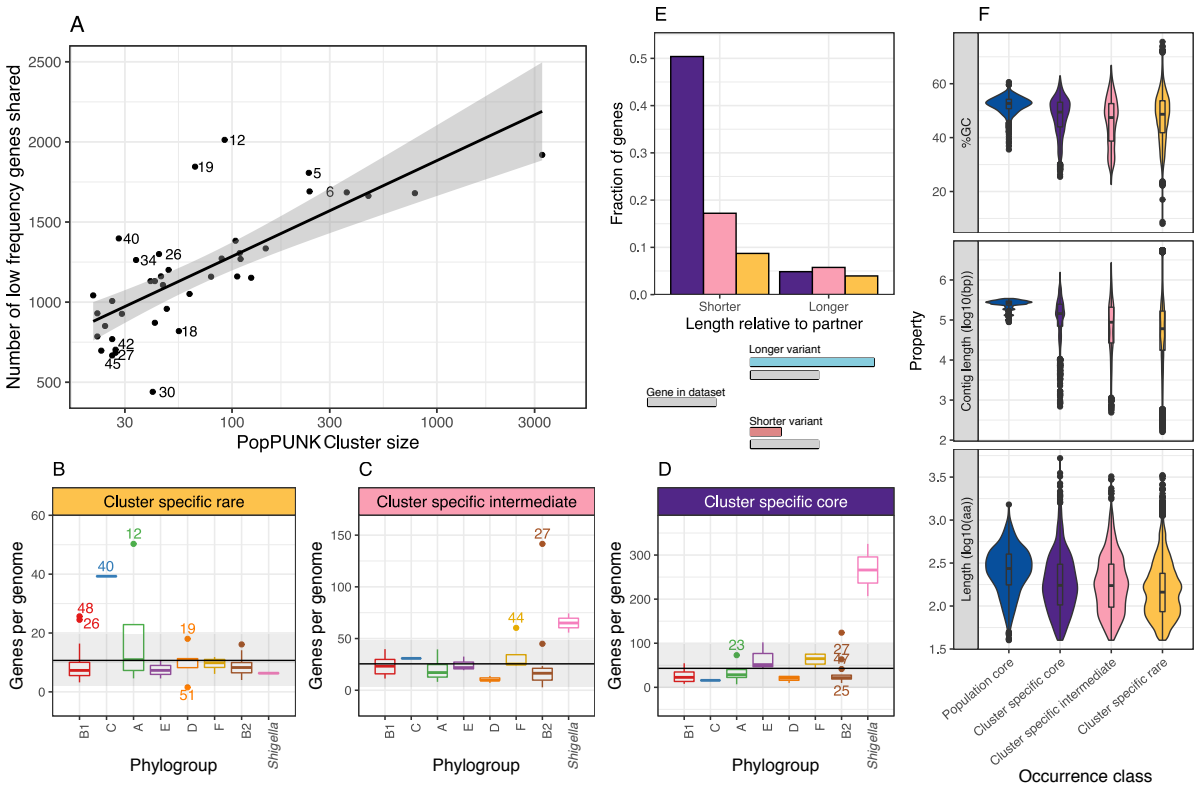

Figure 5.9: Cluster specific genes in the* E. coli *dataset. A *Median number of low frequency genes shared by each PopPUNK Cluster, with other clusters which are phylogenetically distant from it, relative to the size of the cluster. Line fitted using linear regression, shaded area is the 95% confidence interval.* ***B-D*** *Number of "cluster specific rare" genes (**B**), "cluster specific intermediate" genes (**C**), and "cluster specific core" genes (**D**) per genome in each of the 47 PopPUNK Clusters, grouped by phylogroup.* ***E*** *Fraction of cluster specific genes that were found to either be a short variant or a long variant of another gene in the dataset.* ***F*** *Distribution of GC content, contig length and protein length of the genes of cluster specific occurrence classes, compared to the "dataset core" genes.*

5.4.9 Hyper-sharing PopPUNK Clusters possess more "cluster specific rare" genes in a single genome relative to the rest of the clusters

PopPUNK Clusters 48, 26, 40, 12 and 19 had more "cluster specific rare" genes per genome relative to the rest of the PopPUNK Clusters (Figure 5.9B). There was an overlap between clusters which had a high number of "cluster specific rare" genes in each genome and the clusters which shared more low frequency genes with distant PopPUNK Clusters in the dataset. Similar to the "multi-cluster rare" genes, the "cluster specific rare" genes were most commonly predicted to be phage derived or otherwise had other annotations related to HGT such as "conjugational transfer", "restriction modification", "resolvase" and more (not shown).

5.4.10 PopPUNK Clusters which shared fewer low frequency genes than expected also had the largest number of “cluster specific core” genes

PopPUNK Clusters which were not assigned a phylogroup based on the Clermont phylotyping scheme (18) and the *Shigella* PopPUNK Clusters (30,40) shared fewer low frequency genes with distantly related PopPUNK Clusters than expected for their size. The branches leading to these PopPUNK Clusters were estimated to have undergone a large number of gene loss events of “multi-cluster core” genes (Figure 5.5B,C). Additionally, these clusters possessed more “cluster specific core” and “cluster specific intermediate” genes relative to the rest of the PopPUNK Clusters (Figure 5.9A,C,D). While this was expected for *Shigella* spp., PopPUNK cluster 18 is nested within phylogroup B2. This cluster had a mean of 123 “cluster-specific core” genes, relative to a mean of 25 cluster-specific core genes in the rest of the clusters in phylogroup B2. 60% of the isolates of this cluster are from ST504 which has been described in the past as atypical STEC as they have been misclassified as *Shigella* spp. due to the biochemical phenotype these present [415]. Indeed, 100% of the isolates in PopPUNK Cluster 18 were positive for the shiga-toxin gene *stx1B*.

5.4.11 Cluster specific core genes are often truncated variants of other genes in the collection

The sequences of the cluster specific genes, including “cluster specific core”, “cluster specific intermediate” and “cluster specific rare” genes, were aligned against all the other genes in the collection (See Section 5.3.5). Strikingly, 50% of the “cluster specific core” genes were identical along their full length to a region of another gene in the collection (Figure 5.9E). 17% of the “cluster specific intermediate” genes were also identified as shorter variants of other genes in the dataset (Figure 5.9E). Shorter variants of other genes more commonly had an alternative start codon relative to other genes (22% of short variants versus 10% of the rest). Even though only a subset of genes from these occurrence classes were identified as shorter variants of other genes, the length of all cluster specific genes was an order of magnitude shorter than the observed lengths of the “dataset core” genes (Figure 5.9F) The “cluster specific core” genes shared a similar predicted functional profile to those given to the “dataset core and the “multi-cluster core” genes, suggesting these are variants of this same subset of genes (Figure 5.7). Conversely, cluster specific genes had more extreme values in their GC content, particularly the “cluster specific rare genes”, and were more commonly found on shorter contigs (Figure 5.9F).

5.4.12 STEC PopPUNK Cluster 27 and ExPEC PopPUNK Cluster 44 possess a large number of “cluster specific intermediate” genes.

PopPUNK Cluster 44 of phylogroup F and PopPUNK Cluster 27 of phylogroup B2 possessed a high number of “cluster specific intermediate” genes relative to the rest of the clusters (Figure 5.9EC). These clusters had a mean of 60 and 142 “cluster specific intermediate” genes per genome, relative to the mean in the dataset of only 10 “cluster specific intermediate” genes per genome. 100% of the isolates of cluster 44 were from ST648 and were mostly (72%) ExPECs collected from either blood or urine samples. These isolates are multi-drug resistant, with observed resistance to fluoroquinolones, macrolides, aminoglycosides and beta-lactams including ESBLs and carbapenems (See Section 4.4.4.6). ST648 has been described as an emerging multi-drug resistant lineage of phylogroup F, present both in humans and animals [416,417]. Cluster 27, on the other hand, contains 88% isolates from ST583 and 66% of the isolates were collected from fecal samples. Additionally, 66% of isolates from PopPUNK Cluster 27 were positive for shiga toxin gene *stx2B* and 100% positive for *eae* (See Sections 1.1.2.2-3 of Introduction for pathotype definitions). No resistance was observed in this cluster. Thus, these two PopPUNK Clusters with high loads of “cluster specific intermediate” frequency genes are different in their pathogenic and resistance profiles. Their shared property is that they are both out-groups of other clades; PopPUNK Cluster 27 is an out-group of a clade in phylogroup B2 and PopPUNK cluster 44 an out-group in phylogroup F (Chapter 4, Figure 4.8). This resembles the phylogenetic locations of the *Shigella* PopPUNK Clusters 30 and 45 relative to phylogroup B1. PopPUNK Cluster 27 is also similar to these clusters as it shares fewer low frequency genes with distantly related PopPUNK Clusters than expected for its size and the branch leading to PopPUNK Cluster 27 has been estimated to undergone a large number of gain and loss events of “multi-cluster core” genes (Figure 5.9A, 5.4A,B).

5.5 Discussion

An accurate description of the pan-genome of thousands of *E. coli* genomes, when considering all the biases in public genome datasets, required redefining the approach used to understand the distribution of the genes in that dataset. The new approach presented is an extension of previous approaches used for the exploration of the pan-genome in a single species or lineage. In addition to classifying the genes based on their frequency in a lineage, the rules extend to examine the number of lineages, or PopPUNK Clusters, each gene was observed in. The classification presented in this thesis is appropriate given the diversity of the dataset used; Roary, for instance, was designed to handle a dataset with low gene content and sequence diversity and thus would not be applicable to this dataset [305]. Additionally,

this approach corrects for the over-representation of particular lineages in the dataset. For instance, genes which were core and specific to a single PopPUNK Cluster that has a low representation in the dataset would have been mistaken for “rare” genes had we treated all gene-counts equally. However, it is still important to note that the analysis presented here is still an approximation to our understanding of the true distribution of genes in the *E. coli* population. The true representation of each lineage in the natural *E. coli* population is unknown because most of the sequenced isolates in this study, and indeed the public databases have clinical relevance and as such were highly biased in their sampling. Notwithstanding this, as this approach uses two metrics, it provides a higher-resolution to classify the genes in the dataset into occurrence classes which were fully characterised in this thesis, revealing their different functions and dynamics of gain and loss.

There were only 1,426 “dataset core” genes which are the set of genes which are present in every single *E. coli* PopPUNK Cluster and in more than 95% of the isolates of that cluster. These only represent ~30% of the genes in a typical *E. coli* genome. However, there were twice as many genes which were observed in both “core and intermediate” frequencies in multiple PopPUNK Clusters (2,674) and these represent ~40% of the genes in a single *E. coli* genome. The number of PopPUNK Clusters in which these genes were most commonly observed, their mean frequency within those clusters, their predicted functions and their level of association with the population structure revealed that these genes resemble the “dataset core” and the “multi-cluster core” genes, more than they do to the other occurrence classes. Thus, the "core and intermediate" genes represent a level of error that is tolerated using our approach, and they likely represent the "soft-core" genome of the dataset. The fact that these genes were at times observed in intermediate frequencies in particular clusters could be the result of mistakes in sequencing, assembly, annotation or pan-genome pipelines. Alternatively, these genes may be in the process of being lost in some clades. We observed the loss of these genes in PopPUNK Clusters which are undergoing gene degradation like the *Shigella* spp. clusters strengthening the hypothesis that they may be undergoing loss (Figures 5.4C). Importantly, setting a single cut-off between “intermediate” and “core” genes across the entire dataset removes the additional level of understanding of the intricate differences between the genes. Including the “core and intermediate” genes which were observed in 40 PopPUNK Clusters or more as part of the core genome would double the size of the *E. coli* core-genome in this analysis and its relative proportion in a single genome.

Genes which were either core and specific multiple PopPUNK Clusters, i.e. “multi-cluster core” genes, were most commonly found to be gained or lost in a single event on an internal branch in the phylogeny (Figure 5.4,5.5). Genes from these occurrence classes should be further

investigated as they represent the changes in gene content between the clades in the *E. coli* dataset, including the differences between the phylogroups. The fact that these genes had mostly undergone a single gain or loss event suggests that independant shifts in the "core" genome of two or more unrelated lineages are less common. Even so, in 32% of cases changes in the core occur on 3 or more events and in 25% of cases "multi-cluster core" genes are shared between distantly related PopPUNK Clusters. It would be interesting to explore these cases as these could shed light on the commonality of distantly related PopPUNK Clusters and whether they are likely to share similar ecological environments or pressures that lead to the selection of the same genes under different genetic backgrounds.

In most cases, gene sharing of low frequency genes was found to be independent of the phylogenetic distance between the two PopPUNK Clusters being compared. This is an indication of a lack of barrier for movement of these genes between distantly related isolates, for instance, compatibility of phage receptors across the species. Additionally, low frequency genes were estimated to have undergone a large number of gain and loss events along the tree branches, mostly commonly on the tree tips. This means that low frequency genes transfer between distantly related isolates and this happens on short evolutionary timescales. The dependency between the size of the two PopPUNK Clusters being compared and the number of low frequency genes shared between them means that we do not observe sharing of genes due to under-representation of particular lineages rather than lack of sharing between them. This is a likely scenario in the case of low frequency genes as more isolates need to be sampled for these genes to be observed. We have not sampled enough from most of the PopPUNK Clusters in this study in order to truly understand the level of gene sharing of low frequency genes between them. For the largest clusters, we observed a plateau in the number of shared low-frequency genes, meaning that from a specific sample size we were able to capture most of the low frequency genes that are shared between these clusters.

Particular PopPUNK Clusters shared more low-frequency genes with distantly related PopPUNK Clusters than expected for their size and appeared to have an increased ability to acquire genes. Most prominently, these include PopPUNK Cluster 12 which contains isolates from ST10 and PopPUNK Cluster 40 which contains isolates from ST23 and ST410, as well as other PopPUNK Clusters which contain MDR isolates. Interestingly, these same PopPUNK Clusters also contained a high number of "cluster specific rare" genes per genome relative to the rest of the dataset. The correlation between the number of rare genes per genome and enhanced sharing of low frequency genes suggests that a high frequency of rare variants in a single genome can be seen as an enhanced ability to contain low frequency genes in the genome and perhaps to donate them. This assumption appears to be particularly relevant as

many of the cluster specific rare genes were predicted to be mobile elements, vectors of HGT and defense mechanisms that may all contribute to the levels of HGT within these clades and with other clades. ST10 and ST23 are known for their ubiquity as they have been described as both commensal and pathogenic, MDR, as well as isolated from human and animal sources [404,418]. These properties have labelled these lineages as potential facilitators of gene movement in the population [419]. The results in this thesis strengthen these hypotheses. Even more, other PopPUNK Clusters which share similar properties to PopPUNK Clusters 12 and 40 can be viewed as having a high potential to either acquire multidrug resistance or to facilitate movement of genes in the population. Interestingly, PopPUNK Clusters 12 and 40 tended to have smaller genomes relative to the rest of the PopPUNK Clusters in the dataset, suggesting a small genome is not necessarily an indication of a small gene pool or lower levels of HGT (See Section 4.4.4.5)

Particular PopPUNK Clusters shared fewer low-frequency genes with distantly related PopPUNK Clusters than expected for their size. This was particularly apparent in PopPUNK Clusters 30, 45 and 18 which either belong to *Shigella* spp. (30, 45) or were not assigned a phylogroup using the Clermont typing scheme (18). These same clusters had a much larger proportion of “cluster specific core” genes in a single genome and had lost a large number of “multi-cluster core genes”. These results indicate that these lineages are evolving in a separate trajectory to the rest of the PopPUNK Clusters, with little gene sharing and large shifts in their core genome that is specific to them. While on the surface the “cluster specific core” genes appear to represent the acquisition of new genetic material, we found that these genes are commonly short variants of other genes in the dataset, share a similar functional profile to the “dataset core” genes and were an order of magnitude shorter than the “dataset core” genes. Hence, these genes likely represent the process of loss of function and gene degradation rather than gain of function in these clusters. Indeed, major gene degradation has been described in *Shigella* spp. and thus this is an expected result for PopPUNK Clusters 30 and 45 [420]. PopPUNK Cluster 18, on the other hand, contains STEC isolates of ST504 which has been mistaken for *Shigella* spp. in phenotypic testing [420]. Additionally, clusters 30 and 45 have smaller genome sizes relative to the rest of the PopPUNK Clusters, fitting with a model of gene-degradation (Chapter 4, Figure 4.10A). PopPUNK Cluster 18, on the other hand, has a similar genome size to the rest of the clusters in the dataset. This suggests it is undergoing an evolutionary process leading to a phenotype that differs from the rest of the dataset and resembles *Shigella* spp. while maintaining production of shiga-toxin and a large genome.

The new approach for investigating the pan-genome in this study is simple and based on the expansion of the existing approaches however, this analysis provides valuable novel insights regarding gene-sharing and evolutionary dynamics of the lineages in this dataset. Future studies for pan-genomes analysis can use the insights from this study to use more relevant properties beyond the frequency, such as gain and loss rates, clade-association and function, to better define the gene-pools in large collections.

6 Conclusions and Future Directions

K. pneumoniae and *E. coli* are clinically important organisms, which have highly diverse populations which include multiple co-circulating lineages across ecological niches and possess very large gene pools [9,115–117,120]. As such, there is ongoing emergence of novel virulent and multi-drug resistant lineages, led predominantly by acquisition of clinically relevant genes via HGT [74,421,422]. In this thesis, the gene pools of these two organisms were examined in order to answer questions regarding their diversity, the distribution of the genes across lineages and the level of gene sharing between lineages. Firstly, these questions were focused on the examination of the distribution of TA systems across a global collection of *K. pneumoniae* isolates. Secondly, the observations from the analysis on TA systems were expanded to look at the entire gene pool of a collection of 10,000 *E. coli* isolates. The availability of a large number of publicly available genomes, generated worldwide and in different settings was utilised to address these questions. The main conclusions and future directions are detailed in a point by point basis below.

6.1 Other use cases of SLING

In Chapter 2, SLING was presented as a tool that can be used to search for genes which are physically linked in large bacterial genomic datasets. Two use cases were presented. The first, to search for TA systems which represent a simple two-component operon, and the second, to search for RND efflux pumps which represent more complex operons where the order of the genes and operon structure vary across isolates. The usefulness of SLING in describing the diversity of these systems was illustrated. An analysis of 90 *E. coli* genomes revealed different distribution patterns of these operons and indicated that some genes undergo more gain and loss than others.

SLING was designed such that the search is not limited to the two use cases presented in Chapter 2. Searches for other important operons can be constructed as detailed on the SLING Wikipedia page (https://github.com/ghoresh11/sling/wiki/create_db). Possible searches could include examining the distributions of restriction-modification systems, secretion systems and CRISPR systems, all important in their contribution to HGT (See Section 1.2.2.1). Importantly, we showed that by applying SLING in Chapter 3 to identify TA systems in *K. pneumoniae*, we discovered novel antitoxins. Hence, applying SLING to search for other gene systems could lead to the discovery of other novel genes. Finally, SLING can further be used as a discovery tool by applying a reverse search (Figure 2.7). For instance, following the discovery of novel

antitoxins, novel antitoxin HMM profiles can be constructed as the query gene in SLING. The genes found in proximity in the new reverse search are candidate novel toxins.

6.2 Further exploration of the biological implications of toxin-antitoxin pairings, the genetic background of the host and their genetic context

Prior to the investigation presented on TA systems in *K. pneumoniae* in Chapter 3, these systems have mostly been studied on small scales in model organisms [253,335,347–350]. The analysis on a global clinical collection showed that while TA systems are all given the same term, the toxins of these systems can be classified based on their distribution patterns in the dataset, as well as based on the level of diversity of their antitoxin repertoire [345]. In addition to this, a high load of orphan antitoxins was observed in the genomes. Finally, some toxins were commonly associated with the presence of plasmid replicons, AMR genes or virulence genes.

The above conclusions require further experiments to explore their implications. These would include testing toxin and antitoxin combinations in order to understand the effect of these combinations on the functionality of the operon. This should be tested across different host genetic backgrounds, as our analysis found that some toxin-antitoxin pairings are specific to a species. The orphan antitoxins should be included in these experiments. These would include RNAseq experiments which would shed light on whether these systems, including the orphan antitoxins, are expressed in their hosts and under which conditions. This will elucidate whether the orphan antitoxins are functional antitoxins which can serve as a protective system against infection with other TA systems or otherwise, if their presence changes the functionality of other TA operons. The coding sequences in proximity to the orphan antitoxins should be further explored as candidates for discovery of novel toxins. Finally, it would be interesting to apply long-read sequencing on selected strains to identify the genetic context of the TA systems. This would shed light on whether ubiquitous or species associated toxins are chromosomally encoded, or whether they are present on plasmids which have persisted across the species, as well as whether toxins which were associated with clinically relevant genes are present on a plasmid with these genes and helping to maintain them.

6.3 Examination of TA systems on even larger scales

The conclusions presented in Chapter 3 regarding the distribution of TA systems were limited to an analysis on the global population of *K. pneumoniae*. The distribution of these systems across other species and genera is still unexplored. Early studies examining the distribution of TA systems across the entire European Nucleotide Archive (ENA) using BItsliced Genomic Signature Index (BIGSI), a tool which enables to query the ENA easily, in combination with SLING [311,423] have been set up. In this study, SLING was applied on over 3,000 genomes, strengthening its usability in searching for TA systems across large genomic datasets. This type of search is agnostic to species or genus boundaries and examines the distribution of these systems across thousands of genomes. This search could be further expanded to search for these systems in metagenomic datasets as well.

6.4 Therapeutic potential of TA systems

The analysis on TA systems presented in Chapter 3 revealed that TA systems are highly abundant in the *K. pneumoniae* species complex. The discovery of the range of TA systems present in clinical isolates can be used as a potential therapeutic against *K. pneumoniae* infections. This avenue of research is particulatly relevant to further explore in *K. pneumoniae* due to the increasing levels of multi-drug resistance and the inability to treat infections. Assuming these systems are expressed in the host, and that the expression of the toxin inhibits growth or leads to cell death in physiological conditions, new drugs, using peptides or small molecules, can be designed to inhibit the interaction between the toxin and the antitoxin [352]. The classification of the TA systems based on their distribution patterns, presented in Chapter 3, would lead to different outcomes based on the TA system being targeted. Targetting of ubiquitous or species associated TA systems would lead to growth arrest or death of all members of the species complex or one of the species. Alternatively, targettting the sporadic TA systems, which were found to be associated with the presence of AMR and virulence genes and plasmids, can be targeted to prevent the maintenance of these genes and thus lead to their loss. These differences between the TA systems highlight the need to better characterise the distribution of these systems across both clinical and non-clinical isolates in order to better understand their therapeutic potential.

6.5 More reliable databases and scalable tools are required

Through the research presented in this thesis, issues were raised regarding the accessibility of available genomic data and metadata, as well as the scalability of existing tools. While tools

have been developed to address the research questions presented in this thesis, they were either not applicable or not scalable to the size of the datasets used. Existing tools to search for TA systems could only be applied on a small number of genomes, and hence SLING was developed (Chapter 2) [311]. Prokka, the genome annotation tool, was not originally designed for comparative genomics, but rather for the annotation of a single genome (Chapter 4) [293]. The pan-genome analysis tool Roary, was designed to be applied on relatively clonal populations and had to be modified for the purpose of this thesis (Chapter 4) [305]. Furthermore, the data collection process was not straightforward and required programming skills and computational resources (Chapter 4).

The dataset of *E. coli* genomes presented in Chapters 4 and 5 should be made available for others to easily access without barriers of computational ability or resources. Ideally, this would be an online resource which enables users to query an *E. coli* genome in order to investigate its context in the *E. coli* pan-genome, for instance, by providing an R Shiny app [424]. Otherwise, a gene could be used as a query to investigate its distribution across the collection. This can provide the context required when working on a single gene system relative to *E. coli* clinical isolates.

6.6 More systematic sampling of under-represented *E. coli* lineages

The *E. coli* genome collection presented was limited to publicly available data, and hence was heavily biased towards clinical isolates causing disease in the developed world. The vast majority of isolates were collected from Europe and North America and include almost exclusively EHECs and ExPECs. This dataset does not represent natural *E. coli* populations nor does it represent the global clinical burden of *E. coli*, but rather it represents isolates that have been heavily sequenced due to their clinical significance where sequencing is available.

Systematic sampling of *E. coli* isolates is required to cover the full breadth of the *E. coli* diversity. This includes more sampling from under-represented areas of the world, as well as increased sampling of non-pathogenic isolates to better understand commensal *E. coli* populations. One possibility is to expand the research presented to include metagenomic assembled *E. coli* genomes, which could represent commensal populations [425]. Additionally, isolates from other hosts and environments beyond human isolates should be included. The analysis of these genomes can be compared to the dataset presented in this thesis to test whether the same gene distribution or lineages are observed across different

niches. When investigating gene movement, it is essential to include commensal *E. coli* and isolates from different environments as these could be facilitating the movement of genes between lineages and environments.

6.7 Further genomic analysis, as well as functional studies, to understand the differences and commonalities between *E. coli* lineages

In Chapter 5 of this thesis, it was revealed that a large part of the accessory genome is in fact comprised of genes which are core to one lineage or multiple lineages in the dataset. This emphasises the need to expand on traditional pan-genome analyses, as genes which on the surface are part of the accessory genome are core to part of the dataset. These genes are important both to better understand the evolution of the species, and in their potential in diagnostics.

The genes which were identified as core to a subset of the dataset should be further explored both *in-silico* and experimentally for their biological implications. A genomic analysis can be used to investigate genes which differ across the lineages and explore their predicted functions and what implications their presence and absence might have, followed up by functional experiments. The “multi-cluster core” genes should be explored as they represent the shifts in the core genome between *E. coli* lineages. “Multi-cluster core genes” which were acquired independently in multiple evolutionary events are interesting to explore as cases of parallel evolution. The “cluster specific core” should be further explored as they were only observed in a single lineage and were core to that lineage.

For instance, the analysis presented revealed that approximately 100 genes were gained on the branch leading to phylogroup B2, and approximately 70 genes were, on the other hand, lost on that branch. This could be a result of compensatory relationships between these genes, or otherwise, may reveal adaptation of this phylogroup to a particular niche which manifests in major changes in the core genome. Furthermore, the genetic context of these genes can be explored to better understand whether the shifts in the core genome occurred in a single evolutionary event which was beneficial and led to the expansion of this phylogroup, or alternatively, whether this was the the result of the accumulation of many changes, spread across the genome, which occurred over time.

The importance of further investigating these genes is particularly relevant for their potential use in diagnostics and epidemiology. Whole genome sequencing is often not available in clinical laboratories which need to identify the specific causative agent of an infection, nor are they always available during epidemiological investigations. The “cluster specific core” genes and “multi-cluster core” genes were only observed in specific lineages and were core to them, and therefore should be further investigated for their potential as marker genes to identify any of the lineages presented in this thesis using simple assays such as PCR. Importantly, the biased nature of the dataset presented is a major caveat to this potential. The lack of representation of most *E. coli* lineages means that we cannot rule out that the lineage specific genes identified here are not present in any other member of the species. We also cannot rule out that other members of a lineage, not sampled in this study, would possess the lineage specific genomes observed in this collection as sampling needs to be expanded to include other hosts and geographical locations. This emphasises the need to continue to apply similar analyses on much broader collections.

6.8 Examining the routes of movement of the shared low and intermediate frequency genes

Low frequency genes in the *E. coli* dataset were frequently gained and lost and their sharing was independent of the phylogenetic distance between lineages. A number of lineages were identified which shared more of these genes than expected, which were termed “hyper-sharers”.

The routes of movement of these genes should be further investigated. Understanding the precise routes of gene movement in the population would reveal and confirm the hypothesis presented that some lineages are facilitating the movement of genes in the population more than others. By understanding this, we could begin to tackle the problem of the introduction and propagation of novel resistance and virulence genes in the population. Understanding the complete routes by which genes are moving in the population is a harder problem to address, especially given that the dataset is biased and not densely sampled. While the bias in the dataset and under sampling cannot be resolved without more systematic sampling, this question can begin to be tackled in various ways. For instance, comparing gene-trees to species trees for the mobile genes could unravel whether the “hyper-sharers” are the source of many genes, or whether they are a hub and that genes in the dataset are passing through them. Additionally, the genetic context of these genes should be investigated, as a shared context would imply that genes are moving on the same element. Examining the co-

occurrence of these genes across the dataset could shed light on whether some of these genes are moving together as a unit. Furthermore, additional genomic analysis should be applied on the “hyper-sharers” to look at levels of recombination within these isolates, as well as presence of particular MGEs or genes which facilitate HGT that may be contributing to the observed property.

6.9 Further exploration of the rare and intermediate genes

A large proportion of the *E. coli* gene pool is represented by rare genes that were only observed in a single lineage and observed in a low frequency within that lineage. The function of most of these rare genes is unknown. The lineages termed “hyper-sharers” tended to have more of these rare genes within a single isolate genome relative to the other lineages in the dataset. Additionally, we found genes which were found in approximately 50% of isolates, across 50% of the lineages, representing genes which were truly intermediate across the collection. When we examined one such gene, it was a short protein with unknown function.

The origin and function of these rare and intermediate genes should be further explored. BIGSI can be used to search the ENA for these genes to see if they are present in other genera, in addition to searching for them in metagenomic samples. Computational approaches could be used to reveal more of their function, for instance via a “guilt-by-association” approach, followed by functional experiments. Furthering our understanding of these genes is highly relevant. First and foremost, they represent the majority of the *E. coli* gene pool and they have mostly been unexplored. Resistance and pathogenicity, as well as colonisation of different niches, are almost exclusively driven by the accessory genome in *E. coli*, therefore a better characterisation of the genes that make up the majority of the gene pool is highly relevant. As shown in Chapter 4, many AMR and virulence genes were observed in low or intermediate frequencies across the lineages (Figures 4.11, 4.12). This suggests that it is beneficial that only a fraction of the population would possess these genes as in this way the potential for their propagation exists under selective pressure, yet the metabolic burden of possessing them does not inhibit the growth of the whole population. Thus, understanding their function is essential in order to understand the full potential of the gene pool. Secondly, more of these genes were observed in isolates which tended to share more genes with other lineages- it is possible that these genes themselves are contributing to gene movement in the population.

To conclude, this thesis sets the basis for a range of future studies. These include examining TA systems experimentally to investigate the implications of the work presented here, or

otherwise to expand the search even more and examine these systems across more organisms. Within *E. coli*, the high-quality dataset presented sets an opportunity to address more questions regarding the movement of genes between lineages, the differences between the lineages and the function of these genes. This thesis sets a baseline to begin our understanding. With the availability of this resource to the broader scientific community, I hope that the future directions mentioned above and more will be addressed by us and others.

References

1. Jenkins C, Rentenaar RJ, Landraud L, Brisse S. Enterobacteriaceae*. Infectious Diseases*. Elsevier; 2017; 1565–1578.

2. Cohen J, Powderly WG, Opal SM. *Infectious Diseases E-Book*. Elsevier Health Sciences; 2016.

3. McClelland M, Florea L, Sanderson K, Clifton SW, Parkhill J, Churcher C, et al. Comparison of the *Escherichia coli* K-12 genome with sampled genomes of a *Klebsiella pneumoniae* and three *Salmonella enterica* serovars, Typhimurium, Typhi and Paratyphi. *Nucleic Acids Res*. 2000;28: 4974–4986.

4. Wyres KL, Holt KE. *Klebsiella pneumoniae* as a key trafficker of drug resistance genes from environmental to clinically important bacteria. *Curr Opin Microbiol*. 2018;45: 131–139.

5. Martin RM, Bachman MA. Colonization, Infection, and the Accessory Genome of *Klebsiella pneumoniae*. *Front Cell Infect Microbiol*. 2018;8: 4.

6. Kaper JB, Nataro JP, Mobley HL. Pathogenic *Escherichia coli*. *Nat Rev Microbiol.* 2004;2: 123–140.

7. Magill SS, Edwards JR, Bamberg W, Beldavs ZG, Dumyati G, Kainer MA, et al. Multistate point-prevalence survey of health care-associated infections. *N Engl J Med*. 2014;370: 1198–1208.

8. Poolman JT, Wacker M. Extraintestinal Pathogenic *Escherichia coli*, a Common Human Pathogen: Challenges for Vaccine Development and Progress in the Field. *Journal of Infectious Diseases*. 2016: 6–13.

9. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proc Natl Acad Sci U S A*. 2015;112: E3574–E3581.

10. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol*. 2010;8: 207–217.

11. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet*. 2009;5: e1000344.

12. Pendleton JN, Gorman SP, Gilmore BF. Clinical relevance of the ESKAPE pathogens. *Expert Rev Anti Infect Ther*. 2013;11: 297–308.

13. Llaca-Díaz JM, Mendoza-Olazarán S, Camacho-Ortiz A, Flores S, Garza-González E. One-year surveillance of ESKAPE pathogens in an intensive care unit of Monterrey, Mexico. *Chemotherapy.* 2012;58: 475–481.

14. Organization WH, Others. Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. Geneva: World Health Organization. 2017.

15. Friedlaender C. Ueber die Schizomyceten bei der acuten fibrösen Pneumonie. *Archiv für Pathologische Anatomie und Physiologie und für Klinische Medicin.* 1882; 319–324.

16. Brisse S, Verhoef J. Phylogenetic diversity of *Klebsiella pneumoniae* and *Klebsiella oxytoca* clinical isolates revealed by randomly amplified polymorphic DNA, *gyrA* and *parC* genes sequencing and automated ribotyping. *Int J Syst Evol Microbiol*. 2001;51: 915–924.

17. Fevre C, Passet V, Weill F-X, Grimont PAD, Brisse S. Variants of the *Klebsiella pneumoniae* OKP chromosomal beta-lactamase are divided into two main groups, OKP-A and OKP-B. *Antimicrob Agents Chemother*. 2005;49: 5149–5152.

18. Brisse S, Passet V, Grimont PAD. Description of *Klebsiella quasipneumoniae* sp. nov., isolated from human infections, with two subspecies, *Klebsiella quasipneumoniae* subsp. *quasipneumoniae* subsp. nov. and *Klebsiella quasipneumoniae* subsp. *similipneumoniae* subsp. nov., and demonstration that *Klebsiella singaporensis* is a junior heterotypic synonym of *Klebsiella variicola. Int. J. Syst. Evol.* 2014; 64(9): 3146–3152.

19. Blin C, Passet V, Touchon M, Rocha EPC, Brisse S. Metabolic diversity of the emerging pathogenic lineages of *Klebsiella pneumoniae*. *Environ Microbiol*. 2017;19: 1881–1898.

20. Long SW, Wesley Long S, Linson SE, Saavedra MO, Cantu C, Davis JJ, et al. Whole-Genome Sequencing of a Human Clinical Isolate of the Novel Species *Klebsiella quasivariicola* sp. nov. *Genome Announc. 2017;5(42):e01057-17.*

21. Rodrigues C, Passet V, Rakotondrasoa A, Diallo TA, Criscuolo A, Brisse S. Description of *Klebsiella africanensis* sp. nov., *Klebsiella variicola* subsp. *tropicalensis* subsp. nov. and *Klebsiella variicola* subsp. *variicola* subsp. nov. *Res Microbiol*. 2019;170: 165–170.

22. Ørskov I, Ørskov F. 4 Serotyping of *Klebsiella*. *Methods in Microbiology.* Academic Press; 1984; 143–164.

23. Diancourt L, Passet V, Verhoef J, Grimont PAD, Brisse S. Multilocus sequence typing of *Klebsiella pneumoniae* nosocomial isolates. *J Clin Microbiol*. 2005;43: 4178–4182.

24. Podschun R, Ullmann U. *Klebsiella* spp. as nosocomial pathogens: epidemiology, taxonomy, typing methods, and pathogenicity factors. *Clin Microbiol Rev*. 1998;11: 589–603.

25. Wyres KL, Holt KE. *Klebsiella pneumoniae* Population Genomics and Antimicrobial-Resistant Clones. *Trends Microbiol*. 2016;24: 944–956.

26. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A*. 1998;95: 3140–3145.

27. Karkey A, Thanh DP, Boinett CJ, Cain AK, Ellington M, Baker KS, et al. A high-resolution genomic analysis of multidrug-resistant hospital outbreaks of *Klebsiella pneumoniae*. *EMBO Mol Med*. 2015;7: 227–239.

28. Campos AC, Albiero J, Ecker AB, Kuroda CM, Meirelles LEF, Polato A, et al. Outbreak of *Klebsiella pneumoniae* carbapenemase-producing *K. pneumoniae*: A systematic review. *Am J Infect Control*. 2016;44: 1374–1380.

29. Ko W-C, Paterson DL, Sagnimeni AJ, Hansen DS, Von Gottberg A, Mohapatra S, et al. Community-acquired *Klebsiella pneumoniae* bacteremia: global differences in clinical patterns. *Emerg Infect Dis*. 2002;8: 160–166.

30. Shon AS, Bajwa RPS, Russo TA. Hypervirulent (hypermucoviscous) *Klebsiella pneumoniae*: a new and dangerous breed. *Virulence*. 2013;4: 107–118.

31. Struve C, Roe CC, Stegger M, Stahlhut SG, Hansen DS, Engelthaler DM, et al. Mapping the Evolution of Hypervirulent *Klebsiella pneumoniae*. *MBio*. 2015;6: e00630.

32. Pomakova DK, Hsiao C-B, Beanan JM, Olson R, MacDonald U, Keynan Y, et al. Clinical and phenotypic differences between classic and hypervirulent *Klebsiella pneumonia*: an emerging and under-recognized pathogenic variant. *Eur J Clin Microbiol Infect Dis.* 2012;31: 981–989.

33. Long SW, Linson SE, Ojeda Saavedra M, Cantu C, Davis JJ, Brettin T, et al. Whole-Genome Sequencing of Human Clinical *Klebsiella pneumoniae* Isolates Reveals Misidentification and Misunderstandings of *Klebsiella pneumoniae*, *Klebsiella variicola*, and *Klebsiella quasipneumoniae*. *mSphere. 2017;2(4):e00290-17.*

34. Berry GJ, Loeffelholz MJ, Williams-Bouyer N. An Investigation into Laboratory Misidentification of a Bloodstream *Klebsiella variicola* Infection. *J Clin Microbiol.* 2015;53: 2793–2794.

35. Maatallah M, Vading M, Kabir MH, Bakhrouf A, Kalin M, Nauclér P, et al. *Klebsiella variicola* is a frequent cause of bloodstream infection in the Stockholm area, and associated with higher mortality compared to *K. pneumoniae*. *PLoS One*. 2014;9: e113539.

36. David S, Reuter S, Harris SR, Glasner C, Feltwell T, Argimon S, et al. Epidemic of carbapenem-resistant *Klebsiella pneumoniae* in Europe is driven by nosocomial spread. *Nat Microbiol*. 2019;4: 1919–1929.

37. Ellington MJ, Heinz E, Wailan AM, Dorman MJ, de Goffau M, Cain AK, et al. Contrasting patterns of longitudinal population dynamics and antimicrobial resistance mechanisms in two priority bacterial pathogens over 7 years in a single center. *Genome Biol*. 2019;20: 184.

38. Breurec S, Melot B, Hoen B, Passet V, Schepers K, Bastian S, et al. Liver Abscess Caused by Infection with Community-Acquired *Klebsiella quasipneumoniae* subsp. *quasipneumoniae. Emerg Infect Dis*. 2016;22: 529–531.

39. Harada S, Aoki K, Yamamoto S, Ishii Y, Sekiya N, Kurai H, et al. Clinical and Molecular Characteristics of *Klebsiella pneumoniae* Isolates Causing Bloodstream Infections in Japan: Occurrence of Hypervirulent Infections in Health Care. *J Clin Microbiol*. 2019;57.

40. Shankar C, Veeraraghavan B, Nabarro LEB, Ravi R, Ragupathi NKD, Rupali P. Whole genome analysis of hypervirulent *Klebsiella pneumoniae* isolates from community and hospital acquired bloodstream infection. *BMC Microbiol*. 2018;18: 6.

41. O'neill J. Antimicrobial resistance: tackling a crisis for the health and wealth of nations. *Rev Antimicrob Resist.* 2014;20: 1–16.

42. Chen L, Todd R, Kiehlbauch J, Walters M, Kallen A. Notes from the Field: Pan-Resistant New Delhi Metallo-Beta-Lactamase-Producing *Klebsiella pneumoniae* - Washoe County, Nevada, 2016. *MMWR Morb Mortal Wkly Rep.* 2017;66: 33.

43. Bathoorn E, Tsioutis C, da Silva Voorham JM, Scoulica EV, Ioannidou E, Zhou K, et al. Emergence of pan-resistance in KPC-2 carbapenemase-producing *Klebsiella pneumoniae* in Crete, Greece: a close call. *J Antimicrob Chemother*. 2016;71: 1207–

1212.

44. Sonnevend Á, Ghazawi A, Hashmey R. Multihospital occurrence of pan-resistant *Klebsiella pneumoniae* sequence type 147 with an ISE*cp1*-directed *bla*OXA-181 insertion in the *mgrB* gene in the United Arab Emirates. *Antimicrob Agents Chemother*. 2017; 61(7): e00418-17.

45. Lee C-R, Lee JH, Park KS, Jeon JH, Kim YB, Cha C-J, et al. Antimicrobial Resistance of Hypervirulent *Klebsiella pneumoniae*: Epidemiology, Hypervirulence-Associated Determinants, and Resistance Mechanisms. *Front Cell Infect Microbiol*. 2017;7: 483.

46. Wyres KL, Wick RR, Judd LM, Froumine R, Tokolyi A, Gorrie CL, et al. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of *Klebsiella pneumoniae*. *PLoS Genet*. 2019;15: e1008114.

47. Zhang R, Lin D, Chan EW-C, Gu D, Chen G-X, Chen S. Emergence of Carbapenem-Resistant Serotype K1 Hypervirulent *Klebsiella pneumoniae* Strains in China. *Antimicrob Agents Chemother*. 2016;60: 709–711.

48. Gu D, Dong N, Zheng Z, Lin D, Huang M, Wang L, et al. A fatal outbreak of ST11 carbapenem-resistant hypervirulent *Klebsiella pneumoniae* in a Chinese hospital: a molecular epidemiological study. *Lancet Infect Dis.* 2018;18: 37–46.

49. Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, MacLean C. The Ecology and Evolution of Pangenomes. *Curr Biol.* 2019;29: R1094–R1103.

50. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A*. 2009;106: 19126–19131.

51. Shen P, Huang HV. Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. *Genetics*. 1986;112(3):441-457.

52. Paczosa MK, Mecsas J. *Klebsiella pneumoniae*: Going on the Offense with a Strong Defense. *Microbiol Mol Biol Rev.* 2016;80: 629–661.

53. Simoons-Smit AM, Verweij-van Vught AM, MacLaren DM. The role of K antigens as virulence factors in *Klebsiella*. *J Med Microbiol*. 1986;21: 133–137.

54. Follador R, Heinz E, Wyres KL, Ellington MJ, Kowarik M, Holt KE, et al. The diversity of *Klebsiella pneumoniae* surface polysaccharides. *Microb Genom*. 2016;2: e000073.

55. Tomás JM, Benedí VJ, Ciurana B, Jofre J. Role of capsule and O antigen in resistance of *Klebsiella pneumoniae* to serum bactericidal activity. *Infect Immun.* 1986;54: 85–89.

56. Murphy CN, Mortensen MS, Krogfelt KA, Clegg S. Role of *Klebsiella pneumoniae* type 1 and type 3 fimbriae in colonizing silicone tubes implanted into the bladders of mice as a model of catheter-associated urinary tract infections. *Infect Immun.* 2013;81: 3009–3017.

57. Struve C, Bojer M, Krogfelt KA. Characterization of *Klebsiella pneumoniae* type 1 fimbriae by detection of phase variation during colonization and infection and impact on virulence. *Infect Immun.* 2008;76: 4055–4065.

58. Russo TA, Olson R, MacDonald U, Beanan J, Davidson BA. Aerobactin, but not yersiniabactin, salmochelin, or enterobactin, enables the growth/survival of hypervirulent (hypermucoviscous) *Klebsiella pneumoniae* ex vivo and in vivo. *Infect Immun*. 2015;83: 3325–3333.

59. Lam MMC, Wyres KL, Judd LM, Wick RR, Jenney A, Brisse S, et al. Tracking key virulence loci encoding aerobactin and salmochelin siderophore synthesis in *Klebsiella pneumoniae. Genome Med.* 2018;10: 77.

60. Fu Y, Guo L, Xu Y, Zhang W, Gu J, Xu J, et al. Alteration of GyrA amino acid required for ciprofloxacin resistance *in Klebsiella pneumoniae* isolates in China. *Antimicrob Agents Chemother.* 2008;52: 2980–2983.

61. Chen F-J, Lauderdale T-L, Ho M, Lo H-J. The roles of mutations in *gyrA*, *parC*, and *ompK35* in fluoroquinolone resistance in *Klebsiella pneumoniae. Microb Drug Resist.* 2003;9: 265–271.

62. Cannatelli A, D'Andrea MM, Giani T, Di Pilato V, Arena F, Ambretti S, et al. In vivo emergence of colistin resistance in *Klebsiella pneumoniae* producing KPC-type carbapenemases mediated by insertional inactivation of the PhoQ/PhoP *mgrB* regulator. *Antimicrob Agents Chemother*. 2013;57: 5521–5526.

63. Jayol A, Poirel L, Brink A, Villegas M-V, Yilmaz M, Nordmann P. Resistance to colistin associated with a single amino acid change in protein PmrB among *Klebsiella pneumoniae* isolates of worldwide origin. *Antimicrob Agents Chemother*. 2014;58: 4762–4766.

64. Poirel L, Jayol A, Bontron S, Villegas M-V, Ozdamar M, Türkoglu S, et al. The *mgrB* gene as a key target for acquired resistance to colistin in *Klebsiella pneumoniae*. *J Antimicrob Chemother.* 2015;70: 75–80.

65. Groisman EA. The pleiotropic two-component regulatory system PhoP-PhoQ. *J Bacteriol. 2001*;183: 1835–1842.

66. Lee C-R, Lee JH, Park KS, Kim YB, Jeong BC, Lee SH. Global Dissemination of Carbapenemase-Producing *Klebsiella pneumoniae*: Epidemiology, Genetic Context, Treatment Options, and Detection Methods. *Front Microbiol*. 2016;7: 895.

67. Liu Y-Y, Wang Y, Walsh TR, Yi L-X, Zhang R, Spencer J, et al. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study*. Lancet Infect Dis.* 2016;16: 161–168.

68. Gu D-X, Huang Y-L, Ma J-H, Zhou H-W, Fang Y, Cai J-C, et al. Detection of Colistin Resistance Gene *mcr-1* in Hypervirulent *Klebsiella pneumoniae* and *Escherichia coli* Isolates from an Infant with Diarrhea in China*. Antimicrob Agents Chemother*. 2016;60: 5099–5100.

69. Pfeiffer A. Th. Eseherich. die Darmbakterien des Neugeboreiien und Säuglings. *DMW - Deutsche Medizinische Wochenschrift.* 1885: 740–741.

70. Blount ZD. The unexhausted potential of *E. coli. Elife*. 2015;4: e05826.

71. Fratamico PM, DebRoy C, Liu Y, Needleman DS, Baranzoni GM, Feng P. Advances in Molecular Serotyping and Subtyping of *Escherichia coli. Front Microbiol.* 2016;7: 644.

72. Orskov I, Orskov F, Jann B, Jann K. Serology, chemistry, and genetics of O and K antigens of *Escherichia coli*. *Bacteriol Rev.* 1977;41: 667–710.

73. Lim JY, Yoon J, Hovde CJ. A brief overview of *Escherichia coli* O157:H7 and its plasmid O157. *J Microbiol Biotechnol*. 2010;20: 5–14.

74. Muniesa M, Hammerl JA, Hertwig S, Appel B, Brüssow H. Shiga toxin-producing

Escherichia coli O104:H4: a new challenge for microbiology*. Appl Environ Microbiol.* 2012;78: 4065–4073.

75. Chaudhuri RR, Henderson IR. The evolution of the *Escherichia coli* phylogeny. *Infect Genet Evol*. 2012;12: 214–226.

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

77. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, et al. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol*. 2006;60: 1136–1151.

78. Jaureguy F, Landraud L, Passet V, Diancourt L, Frapy E, Guigon G, et al. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. *BMC Genomics*. 2008;9: 560.

79. Clermont O, Gordon D, Denamur E. Guide to the various phylogenetic classification schemes for *Escherichia coli* and the correspondence among schemes. *Microbiology*. 2015;161: 980–988.

80. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*. 2010;11: 595.

81. Selander RK, Caugant DA, Ochman H, Musser JM, Gilmour MN, Whittam TS. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Appl Environ Microbiol*. 1986;51: 873–884.

82. Milkman R. Electrophoretic variation in *Escherichia coli* from natural sources. *Science*. 1973;182: 1024–1026.

83. Selander RK, Levin BR. Genetic diversity and structure in *Escherichia coli* populations. *Science*. 1980;210: 545–547.

84. Whittam TS, Ochman H, Selander RK. Multilocus genetic structure in natural populations of *Escherichia coli. Proc Natl Acad Sci U S A*. 1983;80: 1751–1755.

85. Ochman H, Selander RK. Standard reference strains of *Escherichia coli* from natural populations*. J Bacteriol.* 1984;157: 690–693.

86. Selander RK, Caugant DA, Whittam TS. Genetic structure and variation in natural populations of *Escherichia coli*. Escherichia coli *and* Salmonella typhimurium*: Cellular and Molecular Biology, vol 2*. Washington, D.C.: American Society for Microbiology. 1987: 1625–1648.

87. Herzer PJ, Inouye S, Inouye M, Whittam TS. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli. J Bacteriol.* 1990;172: 6175–6181.

88. Clermont O, Olier M, Hoede C, Diancourt L, Brisse S, Keroudean M, et al. Animal and human pathogenic *Escherichia coli* strains share common genetic backgrounds. *Infect Genet Evol*. 2011;11: 654–662.

89. Clermont O, Dixit OVA, Vangchhia B, Condamine B, Dion S, Bridier-Nahmias A, et al. Characterization and rapid identification of phylogroup G in *Escherichia coli*, a lineage with high virulence and antibiotic resistance potential. *Environ Microbiol*. 2019;21: 3107–3117.

90. Walk ST. The “Cryptic” *Escherichia. EcoSal Plus.* 2015;6

91. Clermont O, Gordon DM, Brisse S, Walk ST, Denamur E. Characterization of the cryptic *Escherichia* lineages: rapid identification and prevalence. *Environ Microbiol*. 2011;13: 2468–2477.

92. Abram KZ, Udaondo Z, Bleker C, Wanchai V. What can we learn from over 100,000 *Escherichia coli* genomes? *bioRxiv*. 2020; 708131; doi: https://doi.org/10.1101/708131

93. Zhou Z, Alikhan N-F, Mohamed K, Fan Y, Agama Study Group, Achtman M. The EnteroBase user’s guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res*. 2020;30: 138–152.

94. Sabarly V, Bouvet O, Glodt J, Clermont O, Skurnik D, Diancourt L, et al. The decoupling between genetic structure and metabolic phenotypes in *Escherichia coli* leads to continuous phenotypic diversity. *J Evol Biol.* 2011;24: 1559–1571.

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

96. Shiga K. Sekiri byogen kenkyu hokoku dai-ichi (first report on etiologic research on dysentery). *Saikingaku Zasshi*. 1897;25: 790.

97. Ewing WH, Starr MP. Edwards and Ewing’s Identification of Enterobacteriaceae: Fourth Edition. *International Journal of Systematic and Evolutionary Microbiology*. 1986;36(4): 581-582

98. Beld MJC, Reubsaet FAG. Differentiation between *Shigella*, enteroinvasive *Escherichia coli* (EIEC) and noninvasive *Escherichia coli. EJCMID.* 2012. pp. 899–904.

99. Pettengill EA, Pettengill JB, Binet R. Phylogenetic Analyses of *Shigella* and Enteroinvasive *Escherichia coli* for the Identification of Molecular Epidemiological Markers: Whole-Genome Comparative Analysis Does Not Support Distinct Genera Designation. *Front Microbiol*. 2015;6: 1573.

100. Chattaway MA, Schaefer U, Tewolde R, Dallman TJ, Jenkins C. Identification of *Escherichia coli* and *Shigella* Species from Whole-Genome Sequences. *J Clin Microbiol*. 2017;55: 616–623.

101. Croxen MA, Finlay BB. Molecular mechanisms of *Escherichia coli* pathogenicity. *Nat Rev Microbiol.* 2010;8: 26–38.

102. Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB. Recent advances in understanding enteric pathogenic *Escherichia coli. Clin Microbiol Rev.* 2013;26: 822–880.

103. Ochoa TJ, Contreras CA. Enteropathogenic *Escherichia coli* infection in children. *Curr Opin Infect Dis.* 2011. pp. 478–483.

104. Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam S, et al. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *Lancet*. 2013;382: 209–222.

105. de la Cabada Bauche J, Dupont HL. New Developments in Traveler’s Diarrhea.

Gastroenterol Hepatol. 2011;7: 88–95.

106. Nguyen Y, Sperandio V. Enterohemorrhagic *E. coli* (EHEC) pathogenesis. *Front Cell Infect Microbiol.* 2012;2: 90.

107. Dean-Nystrom EA, Bosworth BT, Moon HW. Pathogenesis of *Escherichia coli* O157:H7 in weaned calves. *Adv Exp Med Biol*. 1999;473: 173-177.

108. Burger R. EHEC O104:H4 IN GERMANY 2011: LARGE OUTBREAK OF BLOODY DIARRHEA AND HAEMOLYTIC URAEMIC SYNDROME BY SHIGA TOXIN–PRODUCING E. COLI VIA CONTAMINATED FOOD. *Improving Food Safety Through a One Health Approach: Workshop Summary.* Washington (DC): National Academies Press (US); 2012.

109. McLellan LK, Hunstad DA. Urinary Tract Infection: Pathogenesis and Outlook. *Trends Mol Med.* 2016;22: 946–957.

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

111. Laupland KB. Incidence of bloodstream infection: a review of population-based studies. *Clin Microbiol Infect*. 2013;19: 492–500.

112. Asadi Karam MR, Habibi M, Bouzari S. Urinary tract infection: Pathogenicity, antibiotic resistance and development of effective vaccines against Uropathogenic *Escherichia coli. Mol Immunol.* 2019;108: 56–67.

113. Vihta K-D, Stoesser N, Llewelyn MJ, Quan TP, Davies T, Fawcett NJ, et al. Trends over time in *Escherichia coli* bloodstream infections, urinary tract infections, and antibiotic susceptibilities in Oxfordshire, UK, 1998-2016: a study of electronic health records*. Lancet Infect Dis.* 2018;18: 1138–1149.

114. Pupo GM, Karaolis DK, Lan R, Reeves PR. Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and *mdh* sequence studies. *Infect Immun.* 1997;65: 2685–2692.

115. Hazen TH, Donnenberg MS, Panchalingam S, Antonio M, Hossain A, Mandomando I, et al. Genomic diversity of EPEC associated with clinical presentations of differing severity*. Nat Microbiol*. 2016;1: 15014.

116. Hazen TH, Sahl JW, Fraser CM, Donnenberg MS, Scheutz F, Rasko DA. Refining the pathovar paradigm via phylogenomics of the attaching and effacing *Escherichia coli. Proc Natl Acad Sci U S A*. 2013;110: 12810–12815.

117. von Mentzer A, Connor TR, Wieler LH, Semmler T, Iguchi A, Thomson NR, et al. Identification of enterotoxigenic *Escherichia coli* (ETEC) clades with long-term global distribution. *Nat Genet.* 2014;46: 1321–1326.

118. Rasko DA, Del Canto F, Luo Q, Fleckenstein JM, Vidal R, Hazen TH. Comparative genomic analysis and molecular examination of the diversity of enterotoxigenic *Escherichia coli* isolates from Chile. *PLoS Negl Trop Dis*. 2019;13: e0007828.

119. Ingle DJ, Tauschek M, Edwards DJ, Hocking DM, Pickard DJ, Azzopardi KI, et al. Evolution of atypical enteropathogenic *E. coli* by repeated acquisition of LEE pathogenicity island variants. *Nat Microbiol*. 2016;1: 15010.

120. Kallonen T, Brodrick HJ, Harris SR, Corander J, Brown NM, Martin V, et al. Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Res*. 2017 18;27(8):1437-1449.

121. Salipante SJ, Roach DJ, Kitzman JO, Snyder MW, Stackhouse B, Butler-Wu SM, et al. Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains. *Genome Res*. 2015;25: 119–128.

122. Riley LW. Pandemic lineages of extraintestinal pathogenic *Escherichia coli. Clin Microbiol Infect.* 2014;20: 380–390.

123. Morfin-Otero R, Noriega ER, Dowzicky MJ. Antimicrobial susceptibility trends among gram-positive and -negative clinical isolates collected between 2005 and 2012 in Mexico: results from the Tigecycline Evaluation and Surveillance Trial. *Ann Clin Microbiol Antimicrob.* 2015;14: 53.

124. Jones RN, Guzman-Blanco M, Gales AC, Gallegos B, Castro ALL, Martino MDV, et al. Susceptibility rates in Latin American nations: report from a regional resistance surveillance program (2011). *Braz J Infect Dis*. 2013;17: 672–681.

125. Sharma M. Prevalence and antibiogram of Extended Spectrum β-Lactamase (ESBL) producing Gram negative bacilli and further molecular characterization of ESBL producing *Escherichia coli* and *Klebsiella* spp. *J Clin Diagn Res*. 2013;7(10):2173-2177.

126. Cornejo-Juárez P, Vilar-Compte D, García-Horton A, López-Velázquez M, Ñamendys-Silva S, Volkow-Fernández P. Hospital-acquired infections at an oncological intensive care cancer unit: differences between solid and hematological cancer patients. *BMC Infect Dis*. 2016;16: 274.

127. Kelly AM, Mathema B, Larson EL. Carbapenem-resistant Enterobacteriaceae in the community: a scoping review*. Int J Antimicrob Agents*. 2017;50: 127–134.

128. Tang H-J, Hsieh C-F, Chang P-C, Chen J-J, Lin Y-H, Lai C-C, et al. Clinical Significance of Community- and Healthcare-Acquired Carbapenem-Resistant Enterobacteriaceae Isolates. *PLoS One*. 2016. p. e0151897.

129. Brennan BM, Coyle JR, Marchaim D, Pogue JM, Boehme M, Finks J, et al. Statewide surveillance of carbapenem-resistant Enterobacteriaceae in Michigan. *Infect Control Hosp Epidemiol.* 2014;35: 342–349.

130. Wang R, Liu Y, Zhang Q, Jin L, Wang Q, Zhang Y, et al. The prevalence of colistin resistance in *Escherichia coli* and *Klebsiella pneumoniae* isolated from food animals in China: coexistence of *mcr-1* and *bla*NDM with low fitness cost*. Int J Antimicrob Agents.* 2018;51: 739–744.

131. Yamamoto Y, Kawahara R, Fujiya Y, Sasaki T, Hirai I, Khong DT, et al. Wide dissemination of colistin-resistant *Escherichia coli* with the mobile resistance gene *mcr* in healthy residents in Vietnam. *J Antimicrob Chemother*. 2019;74: 523–524.

132. Blango MG, Mulvey MA. Persistence of uropathogenic *Escherichia coli* in the face of multiple antibiotics. *Antimicrob Agents Chemother*. 2010;54: 1855–1863.

133. Pitout JDD, DeVinney R. *Escherichia coli* ST131: a multidrug-resistant clone primed for global domination. *F1000Res*. 2017;6. doi:10.12688/f1000research.10609.1

134. Roer L, Overballe-Petersen S, Hansen F, Schønning K, Wang M, Røder BL, et al. *Escherichia coli* Sequence Type 410 Is Causing New International High-Risk Clones. *mSphere*. 2018;3.

135. Feng Y, Liu L, Lin J, Ma K, Long H, Wei L, et al. Key evolutionary events in the emergence of a globally disseminated, carbapenem resistant clone in *the Escherichia coli* ST410 lineage. *Commun Biol*. 2019;2: 322.

136. Bajaj P, Singh NS, Virdi JS. *Escherichia coli* β-Lactamases: What Really Matters. *Front Microbiol*. 2016;7: 417.

137. Scaletsky ICA, Souza TB, Aranda KRS, Okeke IN. Genetic elements associated with antimicrobial resistance in enteropathogenic *Escherichia coli* (EPEC) from Brazil. *BMC Microbiol*. 2010;10: 25.

138. Senerwa D, Mutanda LN, Gathuma JM, Olsvik O. Antimicrobial resistance of enteropathogenic *Escherichia coli* strains from a nosocomial outbreak in Kenya. *APMIS*. 1991;99: 728–734.

139. Xu Y, Sun H, Bai X, Fu S, Fan R, Xiong Y. Occurrence of multidrug-resistant and ESBL-producing atypical enteropathogenic *Escherichia coli* in China*. Gut Pathog.* 2018;10: 8.

140. Goyal D, Dean N, Neill S, Jones P, Dascomb K. Risk Factors for Community-Acquired Extended-Spectrum Beta-Lactamase-Producing Enterobacteriaceae Infections-A Retrospective Study of Symptomatic Urinary Tract Infections. *Open Forum Infect Dis.* 2019;6: ofy357.

141. Koksal E, Tulek N, Sonmezer MC, Temocin F, Bulut C, Hatipoglu C, et al. Investigation of risk factors for community-acquired urinary tract infections caused by extended-spectrum beta-lactamase *Escherichia coli* and *Klebsiella* species. *Investig Clin Urol.* 2019;60: 46–53.

142. Malvi S, Kumar Y, Appannanavar S, Gautam N, Taneja N, Kaur H, et al. Comparative analysis of virulence determinants, antibiotic susceptibility patterns and serogrouping of atypical enteropathogenic *Escherichia coli* versus typical enteropathogenic *E. coli* in India. *Journal of Medical Microbiology*. 2015. pp. 1208–1215.

143. Zhou Y, Zhu X, Hou H, Lu Y, Yu J, Mao L, et al. Characteristics of diarrheagenic *Escherichia coli* among children under 5 years of age with acute diarrhea: a hospital based study. *BMC Infect Dis.* 2018;18: 63.

144. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, et al. The complete genome sequence of *Escherichia coli* K-12. *Science*. 1997;277: 1453–1462.

145. Hayashi T. Complete Genome Sequence of Enterohemorrhagic *Escherichia coli* O157:H7 and Genomic Comparison with a Laboratory Strain K-12 (Supplement). *DNA Research.* 2001. pp. 47–52.

146. Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, Rasko D, et al. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci U S A*. 2002;99: 17020–17024.

147. Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, et al. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli*

commensal and pathogenic isolates. *J Bacteriol*. 2008;190: 6881–6893.

148. Gordienko EN, Kazanov MD, Gelfand MS. Evolution of pan-genomes of *Escherichia coli, Shigell*a spp., and *Salmonella enterica. J Bacteriol*. 2013;195: 2786–2792.

149. Robins-Browne RM, Holt KE, Ingle DJ, Hocking DM, Yang J, Tauschek M. Are *Escherichia coli* Pathotypes Still Relevant in the Era of Whole-Genome Sequencing? *Front Cell Infect Microbiol*. 2016;6: 141.

150. McDaniel TK, Kaper JB. A cloned pathogenicity island from enteropathogenic *Escherichia coli* confers the attaching and effacing phenotype on *E. coli* K-12. *Mol Microbiol*. 1997;23: 399–407.

151. Kenny B, DeVinney R, Stein M, Reinscheid DJ, Frey EA, Finlay BB. Enteropathogenic *E. coli* (EPEC) transfers its receptor for intimate adherence into mammalian cells. *Cell.* 1997;91: 511–520.

152. Tacket CO, Sztein MB, Losonsky G, Abe A. Role of EspB in Experimental Human Enteropathogenic *Escherichia coli* Infection. *Infection and immunity*. 2000; 68(6): 3689–3695.

153. Girón JA, Ho AS, Schoolnik GK. An inducible bundle-forming pilus of enteropathogenic *Escherichia coli. Science*. 1991;254: 710–713.

154. Tennant SM, Tauschek M, Azzopardi K, Bigham A, Bennett-Wood V, Hartland EL, et al. Characterisation of atypical enteropathogenic *E. coli* strains of clinical origin. BMC *Microbiol*. 2009;9: 117.

155. Lee M-S, Cherla RP, Tesh VL. Shiga toxins: intracellular trafficking to the ER leading to activation of host cell stress responses. *Toxins* . 2010;2: 1515–1535.

156. McWilliams BD, Torres AG. Enterohemorrhagic *Escherichia coli* Adhesins. Microbiology Spectrum. *Microbiol Spectr*. 2014;2(3):10.1128/microbiolspec.EHEC-0003-2013.

157. Sears CL, Kaper JB. Enteric bacterial toxins: mechanisms of action and linkage to intestinal secretion. *Microbiol Rev.* 1996;60(1):167-215.

158. Nataro JP, Kaper JB, Robins-Browne R, Prado V, Vial P, Levine MM. Patterns of adherence of diarrheagenic *Escherichia coli* to HEp-2 cells. *Pediatr Infect Dis J.* 1987;6: 829–831.

159. Panchalingam S, Antonio M, Hossain A, Mandomando I, Ochieng B, Oundo J, et al. Diagnostic microbiologic methods in the GEMS-1 case/control study. *Clin Infect Dis.* 2012;55 Suppl 4: S294–302.

160. Scaletsky IC, Silva ML, Trabulsi LR. Distinctive patterns of adherence of enteropathogenic *Escherichia coli* to HeLa cells. *Infect Immun.* 1984;45: 534–536.

161. Servin AL. Pathogenesis of Human Diffusely Adhering *Escherichia coli* Expressing Afa/Dr Adhesins (Afa/Dr DAEC): Current Insights and Future Challenges. *Clin Microbiol Rev.* 2014;27(4):823-869.

162. Schroeder GN, Hilbi H. Molecular pathogenesis of *Shigella* spp.: controlling host cell signaling, invasion, and death by type III secretion. *Clin Microbiol Rev*. 2008;21: 134–

156.

163. Pupo GM, Lan R, Reeves PR. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A*. 2000;97: 10567–10572.

164. Köhler C-D, Dobrindt U. What defines extraintestinal pathogenic *Escherichia coli*? *Int J Med Microbiol*. 2011;301: 642–647.

165. Pitout JDD. Extraintestinal Pathogenic *Escherichia coli*: A Combination of Virulence with Antibiotic Resistance. *Front Microbiol.* 2012;3:9.

166. Galardini M, Clermont O, Baron A, Busby B, Dion S, Schubert S, 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. *bioRxiv*. 2020; 712034. doi:10.1101/712034

167. Brzuszkiewicz E, Thürmer A, Schuldes J, Leimbach A, Liesegang H, Meyer F-D, et al. Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC). *Arch Microbiol*. 2011;193: 883–891.

168. Carattoli A. Plasmids and the spread of resistance. *Int J Med Microbiol*. 2013;303: 298–304.

169. Morgan-Linnell SK, Becnel Boyd L, Steffen D, Zechiedrich L. Mechanisms accounting for fluoroquinolone resistance in *Escherichia coli* clinical isolates. *Antimicrob Agents Chemother.* 2009;53: 235–241.

170. Olaitan AO, Morand S, Rolain J-M. Mechanisms of polymyxin resistance: acquired and intrinsic resistance in bacteria. *Front Microbiol.* 2014;5: 643.

171. Jacoby GA. AmpC β-Lactamases. *Clinical Microbiology Reviews*. 2009. pp. 161–182.

172. Livermore DM, Canton R, Gniadkowski M, Nordmann P, Rossolini GM, Arlet G, et al. CTX-M: changing the face of ESBLs in Europe. *J Antimicrob Chemother*. 2007;59: 165–174.

173. Nicolas-Chanoine M-H, Bertrand X, Madec J-Y. *Escherichia coli* ST131, an intriguing clonal group. *Clin Microbiol Rev*. 2014;27: 543–574.

174. Brodrick HJ, Raven KE, Kallonen T, Jamrozy D, Blane B, Brown NM, et al. Longitudinal genomic surveillance of multidrug-resistant *Escherichia coli* carriage in a long-term care facility in the United Kingdom. *Genome Med*. 2017;9: 70.

175. Khan AU, Maryam L, Zarrilli R. Structure, Genetics and Worldwide Spread of New Delhi Metallo-β-lactamase (NDM): a threat to public health. *BMC Microbiol*. 2017;17: 101.

176. Hammerl JA, Borowiak M, Schmoger S, Shamoun D, Grobbel M, Malorny B, et al. *mcr-5* and a novel *mcr-5.2* variant in *Escherichia coli* isolates from food and food-producing animals, Germany, 2010 to 2017. *J Antimicrob Chemother*. 2018;73: 1433–1435.

177. Xavier BB, Lammens C, Ruhal R, Kumar-Singh S, Butaye P, Goossens H, et al. Identification of a novel plasmid-mediated colistin-resistance gene, *mcr-2*, in *Escherichia coli,* Belgium, June 2016. *Euro Surveill.* 2016; 7;21(27). doi: 10.2807/1560-

7917.ES.2016.21.27.30280. Erratum in: *Euro Surveill. 2016 14;21(28): PMID: 27416987.*

178. Yin W, Li H, Shen Y, Liu Z, Wang S, Shen Z, et al. Novel Plasmid-Mediated Colistin Resistance Gene *mcr-3* in *Escherichia coli. MBio*. 2017;8.

179. Carattoli A, Villa L, Feudi C, Curcio L, Orsini S, Luppi A, et al. Novel plasmid-mediated colistin resistance *mcr-4* gene in *Salmonella* and *Escherichia coli*, Italy 2013, Spain and Belgium, 2015 to 2016. *Euro Surveill*. 2017;22. doi:10.2807/1560-7917.ES.2017.22.31.30589

180. Sun J, Zhang H, Liu Y-H, Feng Y. Towards Understanding MCR-like Colistin Resistance. *Trends Microbiol*. 2018;26: 794–808.

181. Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web of life. *Nat Rev Genet*. 2015;16: 472–482.

182. Treangen TJ, Rocha EPC. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes*. PLoS Genet*. 2011;7: e1001284.

183. Delavat F, Miyazaki R, Carraro N, Pradervand N, van der Meer JR. The hidden life of integrative and conjugative elements. *FEMS Microbiol Rev*. 2017;41: 512–537.

184. von Wintersdorff CJH, Penders J, van Niekerk JM, Mills ND, Majumder S, van Alphen LB, et al. Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer. *Front Microbiol*. 2016;7: 173.

185. Thomas CM, Nielsen KM. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol*. 2005;3: 711–721.

186. Christie PJ. Type IV secretion: intercellular transfer of macromolecules by systems ancestrally related to conjugation machines. *Mol Microbiol*. 2001;40: 294–305.

187. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, de la Cruz F. Mobility of plasmids. *Microbiol Mol Biol Rev*. 2010;74: 434–452.

188. Wong JJW, Lu J, Glover JNM. Relaxosome function and conjugation regulation in F-like plasmids - a structural biology perspective. *Mol Microbiol.* 2012;85: 602–617.

189. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clin Microbiol Rev*. 2018;31.

190. Mell JC, Redfield RJ. Natural competence and the evolution of DNA uptake specificity. *J Bacteriol*. 2014;196: 1471–1483.

191. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol*. 2005;3: 722–732.

192. Didelot X, Maiden MCJ. Impact of recombination on bacterial evolution. *Trends Microbiol*. 2010;18: 315–322.

193. Chen Z, Yang H, Pavletich NP. Mechanism of homologous recombination from the RecA-ssDNA/dsDNA structures. *Nature*. 2008;453: 489–484.

194. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* 2009;3: 199–208.

195. Siguier P, Gourbeyre E, Varani A, Ton-Hoang B, Chandler M. Everyman's Guide to Bacterial Insertion Sequences. *Microbiol Spectr*. 2015;3: MDNA3–0030–2014.

196. Gillings MR. Integrons: past, present, and future. *Microbiol Mol Biol Rev*. 2014;78: 257–277.

197. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*. 2011;13: 36–46.

198. Watt VM, Ingles CJ, Urdea MS, Rutter WJ. Homology requirements for recombination in *Escherichia coli. Proc Natl Acad Sci U S A.* 1985;82: 4768–4772.

199. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Site-Specific Recombination. *Molecular Biology of the Cell. 4th edition.* Garland Science; 2002.

200. Couturier M, Bex F, Bergquist PL, Maas WK. Identification and classification of bacterial plasmids. *Microbiol Rev.* 1988;52: 375–395.

201. Rozwandowicz M, Brouwer MSM, Fischer J, Wagenaar JA, Gonzalez-Zorn B, Guerra B, et al. Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. *J Antimicrob Chemother.* 2018;73: 1121–1137.

202. Moller AG, Lindsay JA, Read TD. Determinants of Phage Host Range in *Staphylococcus* Species. *Appl Environ Microbiol*. 2019;85.

203. Abedon ST. Bacterial "immunity" against bacteriophages. *Bacteriophage*. 2012; 50–54.

204. Hampton HG, Watson BNJ, Fineran PC. The arms race between bacteria and their phage foes. *Nature*. 2020;577: 327–336.

205. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science*. 2008;322: 1843–1845.

206. Popa O, Dagan T. Trends and barriers to lateral gene transfer in prokaryotes. *Curr Opin Microbiol.* 2011;14: 615–623.

207. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et al. Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet.* 2014;46: 305–309.

208. Rendueles O, de Sousa JAM, Bernheim A, Touchon M, Rocha EPC. Genetic exchanges are more frequent in bacteria encoding capsules. *PLoS Genet*. 2018;14: e1007862.

209. Jang J, Hur H-G, Sadowsky MJ, Byappanahalli MN, Yan T, Ishii S. Environmental *Escherichia coli*: ecology and public health implications-a review. *J Appl Microbiol.* 2017;123: 570–581.

210. Iranzo J, Wolf YI, Koonin EV, Sela I. Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence. *Nat Commun*. 2019;10: 5376.

211. Vos M, Hesselman MC, Te Beek TA, van Passel MWJ, Eyre-Walker A. Rates of Lateral Gene Transfer in Prokaryotes: High but Why? *Trends Microbiol*. 2015;23: 598–605.

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

213. McNally A, Cheng L, Harris SR, Corander J. The evolutionary path to extraintestinal pathogenic, drug-resistant *Escherichia coli* is marked by drastic reduction in detectable recombination within the core genome. *Genome Biol Evol.* 2013;5: 699–710.

214. Chen L, Mathema B, Pitout JDD, DeLeo FR, Kreiswirth BN. Epidemic *Klebsiella pneumoniae* ST258 is a hybrid strain. *MBio*. 2014;5: e01355–14.

215. Comandatore F, Sassera D, Bayliss SC, Scaltriti E, Gaiarsa S, Cao X, et al. Gene Composition as a Potential Barrier to Large Recombinations in the Bacterial Pathogen *Klebsiella pneumoniae. Genome Biol Evol*. 2019;11: 3240–3251.

216. Bergthorsson U, Ochman H. Distribution of chromosome length variation in natural isolates of *Escherichia coli. Mol Biol Evol.* 1998;15: 6–16.

217. Bergthorsson U, Ochman H. Heterogeneity of genome sizes among natural isolates of *Escherichia coli. J Bacteriol*. 1995;177: 5784–5789.

218. Marsh JW, Mustapha MM, Griffith MP, Evans DR, Ezeonwuka C, Pasculle AW, et al. Evolution of Outbreak-Causing Carbapenem-Resistant *Klebsiella pneumoniae* ST258 at a Tertiary Care Hospital over 8 Years. *MBio*. 2019;10.

219. Lin Y-C, Lu M-C, Tang H-L, Liu H-C, Chen C-H, Liu K-S, et al. Assessment of hypermucoviscosity as a virulence factor for experimental *Klebsiella pneumoniae* infections: comparative virulence analysis with hypermucoviscosity-negative strain. *BMC Microbiol*. 2011;11: 50.

220. Lin T-L, Lee C-Z, Hsieh P-F, Tsai S-F, Wang J-T. Characterization of integrative and conjugative element ICEK*p1*-associated genomic heterogeneity in *a Klebsiella pneumoniae* strain isolated from a primary liver abscess. *J Bacteriol.* 2008;190: 515–526.

221. Acheson DW, Reidl J, Zhang X, Keusch GT, Mekalanos JJ, Waldor MK. *In vivo* transduction with shiga toxin 1-encoding phage. *Infect Immun*. 1998;66: 4496–4498.

222. Datz M, Janetzki-Mittmann C, Franke S, Gunzer F, Schmidt H, Karch H. Analysis of the enterohemorrhagic *Escherichia coli* O157 DNA region containing lambdoid phage gene p and Shiga-like toxin structural genes. *Appl Environ Microbiol*. 1996;62: 791–797.

223. Neely MN, Friedman DI. Functional and genetic analysis of regulatory regions of coliphage H-19B: location of shiga-like toxin and lysis genes suggest a role for phage functions in toxin release. *Molecular Microbiology*. 1998. pp. 1255–1267.

224. Dudley EG, Thomson NR, Parkhill J, Morin NP, Nataro JP. Proteomic and microarray characterization of the AggR regulon identifies a pheU pathogenicity island in enteroaggregative *Escherichia coli. Mol Microbiol.* 2006;61: 1267–1282.

225. Nataro JP, Yikang D, Yingkang D, Walker K. AggR, a transcriptional activator of aggregative adherence fimbria I expression in enteroaggregative *Escherichia coli. J Bacteriol. 1*994;176: 4691–4699.

226. Pilla G, Tang CM. Going around in circles: virulence plasmids in enteric pathogens. *Nat Rev Microbiol*. 2018;16: 484–495.

227. Lobato-Márquez D, Díaz-Orejas R, García-Del Portillo F. Toxin-antitoxins and bacterial virulence. *FEMS Microbiol Rev.* 2016;40: 592–609.

228. Brzozowska I, Zielenkiewicz U. Regulation of toxin–antitoxin systems by proteolysis. *Plasmid*. 2013;70: 33–41.

229. Ogura T, Hiraga S. Mini-F plasmid genes that couple host cell division to plasmid proliferation. *Proc Natl Acad Sci U S A.* 1983;80: 4784–4788.

230. Gerdes K, Rasmussen PB, Molin S. Unique type of plasmid maintenance function: postsegregational killing of plasmid-free cells. *Proc Natl Acad Sci U S A.* 1986;83: 3116–3120.

231. Song S, Wood TK. Post-segregational Killing and Phage Inhibition Are Not Mediated by Cell Death Through Toxin/Antitoxin Systems. *Front Microbiol*. 2018;9: 814.

232. Yang QE, Walsh TR. Toxin-antitoxin systems and their role in disseminating and maintaining antimicrobial resistance. *FEMS Microbiol Rev.* 2017;41: 343–353.

233. Short FL, Pei XY, Blower TR, Ong S-L, Fineran PC, Luisi BF, et al. Selectivity and self-assembly in the control of a bacterial toxin by an antitoxic noncoding RNA pseudoknot. *Proc Natl Acad Sci U S A.* 2013;110: E241–9.

234. Masuda H, Tan Q, Awano N, Yamaguchi Y, Inouye M. A novel membrane-bound toxin for cell division, CptA (YgfX), inhibits polymerization of cytoskeleton proteins, FtsZ and MreB, in *Escherichia coli. FEMS Microbiol Lett.* 2012;328: 174–181.

235. Wang X, Lord DM, Cheng H-Y, Osbourne DO, Hong SH, Sanchez-Torres V, et al. A new type V toxin-antitoxin system where mRNA for toxin GhoT is cleaved by antitoxin GhoS. *Nat Chem Biol.* 2012;8: 855–861.

236. Aakre CD, Phung TN, Huang D, Laub MT. A bacterial toxin inhibits DNA replication elongation through a direct interaction with the β sliding clamp. *Mol Cell*. 2013;52: 617–628.

237. Marimon O, Teixeira JMC, Cordeiro TN, Soo VWC, Wood TL, Mayzel M, et al. An oxygen-sensitive toxin–antitoxin system. *Nat Commun*. 2016;7: 13634.

238. Page R, Peti W. Toxin-antitoxin systems in bacterial growth arrest and persistence. *Nat Chem Biol.* 2016;12: 208–214.

239. Fozo EM, Hemm MR, Storz G. Small toxic proteins and the antisense RNAs that repress them. *Microbiol Mol Biol Rev*. 2008;72: 579–89.

240. Dao-Thi M-H, Van Melderen L, De Genst E, Afif H, Buts L, Wyns L, et al. Molecular basis of gyrase poisoning by the addiction toxin CcdB. *J Mol Biol.* 2005;348: 1091–1102.

241. Jiang Y, Pogliano J, Helinski DR, Konieczny I. ParE toxin encoded by the broad-host-range plasmid RK2 is an inhibitor of *Escherichia coli* gyrase. *Mol Microbiol*. 2002;44: 971–979.

242. Mutschler H, Meinhart A. ε/ζ systems: their role in resistance, virulence, and their potential for antibiotic development. *J Mol Med .* 2011;89: 1183–1194.

243. Winther KS, Gerdes K. Enteric virulence associated protein VapC inhibits translation by cleavage of initiator tRNA. *Proc Natl Acad Sci U S A*. 2011;108: 7403–7407.

244. Kaspy I, Rotem E, Weiss N, Ronin I, Balaban NQ, Glaser G. HipA-mediated antibiotic persistence via phosphorylation of the glutamyl-tRNA-synthetase. *Nat Commun*. 2013;4: 3001.

245. Castro-Roa D, Garcia-Pino A, De Gieter S, van Nuland NAJ, Loris R, Zenkin N. The Fic protein Doc uses an inverted substrate to phosphorylate and inactivate EF-Tu. *Nat Chem Biol.* 2013;9: 811–817.

246. Christensen SK, Gerdes K. RelE toxins from bacteria and Archaea cleave mRNAs on translating ribosomes, which are rescued by tmRNA. Mol Microbiol. 2003;48: 1389–1400.

247. Zhang Y, Zhang J, Hoeflich KP, Ikura M, Qing G, Inouye M. MazF cleaves cellular mRNAs specifically at ACA to block protein synthesis in *Escherichia coli. Mol Cell.* 2003;12: 913–923.

248. Blower TR, Pei XY, Short FL, Fineran PC, Humphreys DP, Luisi BF, et al. A processed noncoding RNA regulates an altruistic bacterial antiviral system. *Nat Struct Mol Biol.* 2011;18: 185–190.

249. Masuda H, Tan Q, Awano N, Wu K-P, Inouye M. YeeU enhances the bundling of cytoskeletal polymers of MreB and FtsZ, antagonizing the CbtA (YeeV) toxicity in *Escherichia coli. Mol Microbiol.* 2012;84: 979–989.

250. McVicker G, Tang CM. Deletion of toxin–antitoxin systems in the evolution of *Shigella sonnei* as a host-adapted pathogen. *Nat Microbiol*. 2016;2:16204.

251. Makarova KS, Wolf YI, Koonin EV. Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. *Biol Direct*. 2009;4: 19.

252. Harms A, Maisonneuve E, Gerdes K. Mechanisms of bacterial persistence during stress and antibiotic exposure. *Science*. 2016;354.

253. Helaine S, Cheverton AM, Watson KG, Faure LM, Matthews SA, Holden DW. Internalization of *Salmonella* by macrophages induces formation of nonreplicating persisters. *Science*. 2014;343: 204–208.

254. Cataudella I, Trusina A, Sneppen K, Gerdes K, Mitarai N. Conditional cooperativity in toxin-antitoxin regulation prevents random toxin activation and promotes fast translational recovery. *Nucleic Acids Res*. 2012;40: 6424–6434.

255. Ramisetty BCM, Santhosh RS. Horizontal gene transfer of chromosomal Type II toxin–antitoxin systems of *Escherichia coli. FEMS Microbiol Lett.* 2016;363.

256. Hazan R, Engelberg-Kulka H. *Escherichia coli* mazEF-mediated cell death as a defense mechanism that inhibits the spread of phage P1. *Molecular Genetics and Genomics.* 2004;227–234.

257. Fineran PC, Blower TR, Foulds IJ, Humphreys DP, Lilley KS, Salmond GPC. The phage abortive infection system, ToxIN, functions as a protein–RNA toxin–antitoxin pair. *Proc Natl Acad Sci U S A*. 2009;106: 894–899.

258. Dy RL, Przybilski R, Semeijn K, Salmond GPC, Fineran PC. A widespread bacteriophage abortive infection system functions through a Type IV toxin–antitoxin mechanism. *Nucleic Acids Res*. 2014;42: 4590–4605.

259. Wang Y, Wang H, Hay AJ, Zhong Z, Zhu J, Kan B. Functional RelBE-Family Toxin-Antitoxin Pairs Affect Biofilm Maturation and Intestine Colonization in Vibrio cholerae. *PLoS One*. 2015;10: e0135696.

260. Barrios AFG, Zuo R, Hashimoto Y, Yang L, Bentley WE, Wood TK. Autoinducer 2 controls biofilm formation in *Escherichia coli* through a novel motility quorum-sensing regulator (MqsR, B3022). *J Bacteriol*. 2006;188: 305–316.

261. Soo VWC, Wood TK. Antitoxin MqsA represses curli formation through the master biofilm regulator CsgD. *Sci Rep.* 2013;3: 3186.

262. Wood TL, Wood TK. The HigB/HigA toxin/antitoxin system of *Pseudomonas aeruginosa* influences the virulence factors pyochelin, pyocyanin, and biofilm formation. *MicrobiologyOpen*. 2016;5: 499–511.

263. Yan J, Bassler BL. Surviving as a Community: Antibiotic Tolerance and Persistence in Bacterial Biofilms. *Cell Host Microbe*. 2019;26: 15–21.

264. Stewart PS. Antimicrobial Tolerance in Biofilms. *Microbiol Spectr*. 2015;3.

265. Decano AG, Downing T. An *Escherichia coli* ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. *Sci Rep.* 2019;9: 17394.

266. Chung The H, Karkey A, Pham Thanh D, Boinett CJ, Cain AK, Ellington M, et al. A high-resolution genomic analysis of multidrug-resistant hospital outbreaks of *Klebsiella pneumoniae. EMBO Mol Med*. 2015;7: 227–239.

267. Domman D, Quilici M-L, Dorman MJ, Njamkepo E, Mutreja A, Mather AE, et al. Integrated view of Vibrio cholerae in the Americas. *Science*. 2017. pp. 789–793.

268. Teng JLL, Yeung M-Y, Yue G, Au-Yeung RKH, Yeung EYH, Fung AMY, et al. *In silico* analysis of 16S rRNA gene sequencing based methods for identification of medically important aerobic Gram-negative bacteria. *J Med Microbiol*. 2011;60: 1281–1286.

269. Joensen KG, Tetzschner AMM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and Easy *In Silico* Serotyping of *Escherichia coli* Isolates by Use of Whole-Genome Sequencing Data. *J Clin Microbiol*. 2015;53: 2410–2426.

270. Clermont O, Bonacorsi S, Bingen E. Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Appl Environ Microbiol*. 2000;66: 4555–4558.

271. Clermont O, Christenson JK, Denamur E, Gordon DM. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ Microbiol Rep*. 2013;5: 58–65.

272. Page AJ, Keane JA. Rapid multi-locus sequence typing direct from uncorrected long reads using Krocus. *PeerJ.* 2018;6: e5233.

273. Zhou H, Liu W, Qin T, Liu C, Ren H. Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Whole-Genome Sequence-Based Typing of *Klebsiella pneumoniae. Front Microbiol*. 2017;8: 371.

274. Kingry LC, Rowe LA, Respicio-Kingry LB, Beard CB, Schriefer ME, Petersen JM. Whole genome multilocus sequence typing as an epidemiologic tool for *Yersinia pestis. Diagn Microbiol Infect Dis*. 2016;84: 275–280.

275. De Been M, Pinholt M, Top J, Bletz S. Core genome multilocus sequence typing scheme for high-resolution typing of *Enterococcus faecium*. *J Clin Microbiol*. 2015;53(12):3788-3797

276. Corander J, Waldmann P, Marttinen P, Sillanpaa MJ. BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*. 2004;20(15):2363-2369.

277. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* 2019;29(2):304-316.

278. Katz L, Griswold T, Morrison S, Caravas J, Zhang S, Bakker H, et al. Mashtree: a rapid comparison of whole genome sequence files. *Journal of Open Source Software*. 2019;4: 1762.

279. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*. 2016;17: 132.

280. Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. *Nat Rev Genet*. 2020;21**:** 428–444

281. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 2015;32: 268–274.

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

283. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genom.* 2017;3: e000131.

284. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015;12: 59–60.

285. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215: 403–410.

286. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes*. J Antimicrob Chemother*. 2012;67: 2640–2644.

287. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, et al. *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother*. 2014;58: 3895–3903.

288. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT (Antibiotic Resistance Gene-ANNOTation), a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob Agents Chemother*. 2014;58(1):212-220.

289. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, et al. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli. J Clin Microbiol*. 2014;52: 1501–1510.

290. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med*. 2014;6: 90.

291. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. *Compuational Biology.* Cambridge University Press; 1998;4.

292. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics.* 2010;11: 119.

293. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30: 2068–2069.

294. Tanizawa Y, Fujisawa T, Nakamura Y. DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. *Bioinformatics.* 2018;34: 1037–1039.

295. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST: architecture and applications. *BMC Bioinformatics*. *BMC Bioinformatics* 10, 421 (2009).

296. Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput Biol*. 2011;7: e1002195.

297. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* 2003;31: 371–373.

298. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*. 1997;28: 405–420.

299. Fitch WM. Homology a personal view on some of the problems. *Trends Genet*. 2000;16: 227–231.

300. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*. 2000;28: 33–36.

301. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 2003;13: 2178–2189.

302. Fouts DE, Brinkac L, Beck E, Inman J, Sutton G. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Res. 2012;40: e172.

303. Altenhoff AM, Glover NM, Dessimoz C. Inferring Orthology and Paralogy. *Evolutionary Genomics: Statistical and Computational Methods*. Springer New York; 2019; 149–175.

304. Kristensen DM, Kannan L, Coleman MK, Wolf YI, Sorokin A, Koonin EV, et al. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*. 2010;26: 1481–1487.

305. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31: 3691–3693.

306. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A. Producing Polished Prokaryotic Pangenomes with the Panaroo Pipeline. *bioRxiv.* 2020; 2020.01.28.92298.

doi: https://doi.org/10.1101/2020.01.28.92298

307. Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol.* 2009;5: e1000262.

308. Zekic T, Holley G, Stoye J. Pan-Genome Storage and Analysis Techniques. *Comparative Genomics: Methods and Protocols*. Springer New York; 2018; 29–53.

309. Ding W, Baumdicker F, Neher RA. panX: pan-genome analysis and exploration. *Nucleic Acids Res*. 2018;46: e5.

310. Lassalle F, Veber P, Jauneikaite E, Didelot X. Automated reconstruction of all gene histories in large bacterial pangenome datasets and search for co-evolved gene modules with Pantagruel. *bioRxiv.* 2019; 586495. doi:10.1101/586495

311. Horesh G, Harms A, Fino C, Parts L, Gerdes K, Heinz E, et al. SLING: a tool to search for linked genes in bacterial datasets. *Nucleic Acids Res*. 2018;46(21):e128.

312. Rocha EPC. The organization of the bacterial genome. *Annu Rev Genet.* 2008;42: 211–233.

313. Itoh T, Takemoto K, Mori H, Gojobori T. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol Biol Evol.* 1999;16: 332–346.

314. Price MN, Arkin AP, Alm EJ. The life-cycle of operons. *PLoS Genet*. 2006;2: e96.

315. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, et al. An updated evolutionary classification of CRISPR-Cas systems*. Nat Rev Microbiol.* 2015;13: 722–736.

316. Pandey DP, Gerdes K. Toxin–antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Res.* 2005;33: 966–976.

317. Leplae R, Geeraerts D, Hallez R, Guglielmini J, Drèze P, Van Melderen L. Diversity of bacterial type II toxin-antitoxin systems: a comprehensive search and functional analysis of novel families*. Nucleic Acids Res*. 2011;39: 5513–5525.

318. Xie Y, Wei Y, Shen Y, Li X, Zhou H, Tai C, et al. TADB 2.0: an updated database of bacterial type II toxin-antitoxin loci. *Nucleic Acids Res*. 2018;46(D1):D749-D753.

319. Shao Y, Harrison EM, Bi D, Tai C, He X, Ou H-Y, et al. TADB: a web-based resource for Type 2 toxin-antitoxin loci in bacteria and archaea. *Nucleic Acids Res*. 2011;39: D606–D611.

320. Akarsu H, Bordes P, Mansour M, Bigot D-J, Genevaux P, Falquet L. TASmania: A bacterial Toxin-Antitoxin Systems database. *PLoS Comput Biol*. 2019;15: e1006946.

321. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications*. BMC Bioinformatics.* 2009;10: 421.

322. Sevin EW, Barloy-Hubler F. RASTA-Bacteria: a web-based tool for identifying toxin-antitoxin loci in prokaryotes. *Genome Biol.* 2007;8: R155.

323. Abby SS, Cury J, Guglielmini J, Néron B, Touchon M, Rocha EPC. Identification of protein secretion systems in bacterial genomes. *Sci Rep*. 2016;6: 23080.

324. Martínez-García PM, Ramos C, Rodríguez-Palenzuela P. T346Hunter: a novel web-based tool for the prediction of type III, type IV and type VI secretion systems in bacterial genomes. *PLoS One*. 2015;10: e0119317.

325. Costa TRD, Felisberto-Rodrigues C, Meir A, Prevost MS, Redzej A, Trokter M, et al. Secretion systems in Gram-negative bacteria: structural and mechanistic insights. *Nat Rev Microbiol.* 2015;13: 343–359.

326. Rath D, Amlinger L, Rath A, Lundgren M. The CRISPR-Cas immune system: biology, mechanisms and applications. *Biochimie*. 2015;117: 119–128.

327. Anes J, McCusker MP, Fanning S, Martins M. The ins and outs of RND efflux pumps in *Escherichia coli. Front Microbiol*. 2015;6: 587.

328. Zhang Q, Ye Y. Not all predicted CRISPR--Cas systems are equal: isolated cas genes and classes of CRISPR like elements*. BMC Bioinformatics*. 2017;18: 92.

329. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, et al. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature.* 2011;477: 462–465.

330. Reuter S, Connor TR, Barquist L, Walker D, Feltwell T, Harris SR, et al. Parallel independent evolution of pathogenicity within the genus *Yersinia*. *Proc Natl Acad Sci U S A.* 2014;111: 6768–6773.

331. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25: 1422–1423.

332. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom*. 2016;2: e000056.

333. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res*. 2016;44: W242–5.

334. Wen Z, Wang P, Sun C, Guo Y, Wang X. Interaction of Type IV Toxin/Antitoxin Systems in Cryptic Prophages of *Escherichia coli* K-12. *Toxins*. 2017;9(3):77.

335. Wei Y-Q, Bi D-X, Wei D-Q, Ou H-Y. Prediction of Type II Toxin-Antitoxin Loci in Klebsiella pneumoniae Genome Sequences*. Interdiscip Sci.* 2016;8: 143–149.

336. Brown JM, Shaw KJ. A novel family of *Escherichia coli* toxin-antitoxin gene pairs. *J Bacteriol.* 2003;185: 6600–6608.

337. Wei Y, Zhan L, Gao Z, Privé GG, Dong Y. Crystal structure of GnsA from *Escherichia coli. Biochem Biophys Res Commun*. 2015;462: 1–7.

338. Zhang Y, Inouye M. RatA (YfjG), an *Escherichia coli* toxin, inhibits 70S ribosome association to block translation initiation. *Mol Microbiol*. 2011;79: 1418–1429.

339. Yamaguchi Y, Inouye M. Regulation of growth and death in *Escherichia coli* by toxin–antitoxin systems*. Nat Rev Microbiol.* 2011;9: 779–790.

340. Sun J, Deng Z, Yan A. Bacterial multidrug efflux pumps: mechanisms, physiology and pharmacological exploitations. *Biochem Biophys Res Commun*. 2014;453: 254–267.

341. Blair JMA, Piddock LJV. Structure, function and inhibition of RND efflux pumps in Gram-negative bacteria: an update. *Curr Opin Microbiol*. 2009;12: 512–519.

342. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2017;45: D158–D169.

343. Godoy P, Molina-Henares AJ, de la Torre J, Duque E, Ramos JL. Characterization of the RND family of multidrug efflux pumps: *in silico* to *in vivo* confirmation of four functionally distinct subgroups. *Microb Biotechnol*. 2010;3: 691–700.

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

345. Horesh G, Fino C, Harms A, Dorman MJ, Parts L, Gerdes K, et al. Type II and type IV toxin–antitoxin systems show different evolutionary patterns in the global *Klebsiella pneumoniae* population. *Nucleic Acids Res. 2020;48(8):4357-4370.*

346. Harms A, Brodersen DE, Mitarai N, Gerdes K. Toxins, Targets, and Triggers: An Overview of Toxin-Antitoxin Biology. *Mol Cell.* 2018;70: 768–784.

347. Balaban NQ, Merrin J, Chait R, Kowalik L, Leibler S. Bacterial Persistence as a Phenotypic Switch. *Science*. 2004; 305(5690):1622-1625.

348. Moyed HS, Bertrand KP. *hipA*, a newly recognized gene of *Escherichia coli* K-12 that affects frequency of persistence after inhibition of murein synthesis. *J Bacteriol.* 1983;155: 768–775.

349. Norton JP, Mulvey MA. Toxin-antitoxin systems are important for niche-specific colonization and stress resistance of uropathogenic *Escherichia coli. PLoS Pathog*. 2012;8: e1002954.

350. Fiedoruk K, Daniluk T, Swiecicka I, Sciepuk M, Leszczynska K. Type II toxin–antitoxin systems are unevenly distributed among *Escherichia coli* phylogroups. *Microbiology*. 2015;161: 158–167.

351. Fernández-García L, Blasco L, Lopez M, Bou G, García-Contreras R, Wood T, et al. Toxin-Antitoxin Systems in Clinical Pathogens. *Toxins*. 2016;8(7):227.

352. Lee K-Y, Lee B-J. Structure, Biology, and Therapeutic Application of Toxin–Antitoxin Systems in Pathogenic Bacteria. *Toxins*. 2016;8: 305.

353. Coray DS, Wheeler NE, Heinemann JA, Gardner PP. Why so narrow: distribution of anti-sense regulated, type I toxin-antitoxin systems compared to type II and type III systems. *RNA Biol*. 2017;14(3):275-280.

354. Wen Y, Behiels E, Devreese B. Toxin-Antitoxin systems: their role in persistence, biofilm formation, and pathogenicity. *Pathog Dis*. 2014;70: 240–249.

355. Lobato-Márquez D, Moreno-Córdoba I, Figueroa V, Díaz-Orejas R, García-del Portillo F. Distinct type I and type II toxin-antitoxin modules control *Salmonella* lifestyle inside eukaryotic cells. *Sci Rep*. 2015;5: 9374.

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

357. Page AJ, De Silva N, Hunt M, Quail MA, Parkhill J, Harris SR, et al. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data.

Microb Genom. 2016;2: e000083.

358. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32: 1792–1797.

359. Dixon P. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science.* 2003 Dec;14(6):927-930

360. Wickham H. ggplot2: Elegant Graphics for Data Analysis. *Use R!* Springer; 2016.

361. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014;30: 1236–1240.

362. Gotfredsen M, Gerdes K. The *Escherichia coli* relBE genes belong to a new toxin-antitoxin gene family. *Mol Microbiol*. 1998;29(4):1065-1076.

363. Guzman LM, Belin D, Carson MJ, Beckwith J. Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J Bacteriol.* 1995;177: 4121–4130.

364. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013;30: 772–780.

365. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics.* 2009;25: 1189–1191.

366. Jurėnas D, Garcia-Pino A, Van Melderen L. Novel toxins from type II toxin-antitoxin systems with acetyltransferase activity. *Plasmid*. 2017;93: 30–35.

367. Qian H, Yao Q, Tai C, Deng Z, Gan J, Ou H-Y. Identification and characterization of acetyltransferase-type toxin-antitoxin locus in *Klebsiella pneumoniae. Mol Microbiol.* 2018;108: 336–349.

368. Moyed HS, Broderick SH. Molecular cloning and expression of *hipA*, a gene of *Escherichia coli* K-12 that affects frequency of persistence after inhibition of murein synthesis. *J Bacteriol*. 1986;166: 399–403.

369. Germain E, Castro-Roa D, Zenkin N, Gerdes K. Molecular mechanism of bacterial persistence by HipA. *Mol Cell.* 2013;52: 248–254.

370. Heller DM, Tavag M, Hochschild A. CbtA toxin of *Escherichia coli* inhibits cell division and cell elongation via direct and independent interactions with FtsZ and MreB. *PLoS Genet.* 2017;13: e1007007.

371. Qian H, Yu H, Li P, Zhu E, Yao Q, Tai C, et al. Toxin–antitoxin operon kacAT of Klebsiella pneumoniae is regulated by conditional cooperativity via a W-shaped KacA–KacT complex. *Nucleic Acids Res*. 2019;47(14):7690-7702.

372. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol*. 2015;16: 18–29.

373. Villa L, Capone A, Fortini D, Dolejska M, Rodríguez I, Taglietti F, et al. Reversion to susceptibility of a carbapenem-resistant clinical isolate of *Klebsiella pneumoniae* producing KPC-3. *J Antimicrob Chemother*. 2013;68: 2482–2486.

374. Harmer CJ, Hall RM. pRMH760, a precursor of A/C_2 plasmids carrying *bla*CMY and

*bla*NDM genes. *Microb Drug Resist*. 2014;20: 416–423.

375. Ho P-L, Lo W-U, Chan J, Cheung Y-Y, Chow K-H, Yam W-C, et al. pIMP-PH114 carrying bla IMP-4 in a Klebsiella pneumoniae strain is closely related to other multidrug-resistant IncA/C2 plasmids. Curr Microbiol. 2014;68: 227–232.

376. Carattoli A, Villa L, Poirel L, Bonnin RA, Nordmann P. Evolution of IncA/C *bla*CMY-$_2$-carrying plasmids by acquisition of the *bla*NDM-$_1$ carbapenemase gene. *Antimicrob Agents Chemother.* 2012;56: 783–786.

377. Doublet B, Boyd D, Douard G, Praud K, Cloeckaert A, Mulvey MR. Complete nucleotide sequence of the multidrug resistance IncA/C plasmid *p*R55 from *Klebsiella pneumoniae* isolated in 1969. *J Antimicrob Chemother*. 2012;67: 2354–2360.

378. Wu K-M, Li L-H, Yan J-J, Tsao N, Liao T-L, Tsai H-C, et al. Genome sequencing and comparative analysis of *Klebsiella pneumoniae* NTUH-K2044, a strain causing liver abscess and meningitis. *J Bacteriol*. 2009;191: 4492–4501.

379. Chen Y-T, Chang H-Y, Lai Y-C, Pan C-C, Tsai S-F, Peng H-L. Sequencing and analysis of the large virulence plasmid *p*LVPK of *Klebsiella pneumoniae* CG43. *Gene*. 2004;337: 189–198.

380. Chan WT, Espinosa M, Yeo CC. Keeping the Wolves at Bay: Antitoxins of Prokaryotic Type II Toxin-Antitoxin Systems. *Front Mol Biosci*. 2016;3: 9.

381. Muthuramalingam M, White JC, Bourne CR. Toxin-Antitoxin Modules Are Pliable Switches Activated by Multiple Protease Pathways. *Toxins*. 2016;8(7):214.

382. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res*. 2019;47: D351–D360.

383. Saavedra De Bast M, Mine N, Van Melderen L. Chromosomal toxin-antitoxin systems may act as antiaddiction modules. *J Bacteriol*. 2008;190: 4603–4609.

384. Lin C-Y, Awano N, Masuda H, Park J-H, Inouye M. Transcriptional repressor HipB regulates the multiple promoters in *Escherichia coli. J Mol Microbiol Biotechnol*. 2013;23: 440–447.

385. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res*. 2012;40: D54–6.

386. Swetha RG, Sekar DKK, Devi ED, Ahmed ZZ, Ramaiah S, Anbarasu A, et al. *Streptococcus pneumoniae* Genome Database (SPGDB): A database for strain specific comparative analysis of *Streptococcus pneumoniae* genes and proteins. *Genomics*. 2014. pp. 582–586.

387. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19: 455–477.

388. Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ. PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *Gigascience*. 2019;8(10):giz119.

389. Bolger A, Giorgi F. Trimmomatic: a flexible read trimming tool for illumina NGS data.

Bioinformatics. 2014;30(15):2114-2120.

390. Page AJ, Taylor B, Keane JA. Multilocus sequence typing by blast from de novo assemblies against PubMLST. *J Open Source Softw*. 2016;1: 118.

391. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*. 2010;5: e9490.

392. Menardo F, Loiseau C, Brites D, Coscolla M, Gygli SM, Rutaihwa LK, et al. Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics*. 2018;19: 164.

393. Waters NR, Abram F, Brennan F, Holmes A, Pritchard L. Easily phylotyping E. coli via the EzClermont web app and command-line tool. *bioRxiv*. 2018; 317610. doi:10.1101/317610

394. Pedregosa F, Alexandre Gramfort N, Michel V, Thirion B, Grisel O, Blondel M, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12: 2825–2830.

395. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*. 2004;20: 289–290.

396. Yu G, Smith DK, Zhu H, Guan Y, Lam TT. ggtree : an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol*. 2017;8: 28–36.

397. Kallonen T, Brodrick HJ, Harris SR, Corander J, Brown NM, Martin V, et al. Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Res*. 2017;27(8):1437-1449.

398. Chen SL, Wu M, Henderson JP, Hooton TM, Hibbing ME, Hultgren SJ, et al. Genomic diversity and fitness of *E. coli* strains recovered from the intestinal and urinary tracts of women with recurrent urinary tract infection. *Sci Transl Med.* 2013;5: 184ra60.

399. Baker KS, Burnett E, McGregor H, Deheer-Graham A, Boinett C, Langridge GC, et al. The Murray collection of pre-antibiotic era Enterobacteriacae: a unique research resource. *Genome Med*. 2015;7: 97.

400. Alikhan N-F, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the population structure of *Salmonella*. *PLoS Genet*. 2018;14: e1007261.

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

402. Kim M, Oh H-S, Park S-C, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol.* 2014;64: 346–351.

403. Day MJ, Doumith M, Abernethy J, Hope R, Reynolds R, Wain J, et al. Population structure of *Escherichia coli* causing bacteraemia in the UK and Ireland between 2001 and 2010. *J Antimicrob Chemother*. 2016;71: 2139–2142.

404. Bortolaia V, Larsen J, Damborg P, Guardabassi L. Potential pathogenicity and host range of extended-spectrum beta-lactamase-producing *Escherichia coli* isolates from healthy poultry. *Appl Environ Microbiol*. 2011;77: 5830–5833.

405. Edgar R, Bibi E. MdfA, an *Escherichia coli* multidrug resistance protein with an extraordinarily broad spectrum of drug recognition. *J Bacteriol*. 1997;179: 2274–2280.

406. Magiorakos A-P, Srinivasan A, Carey RB, Carmeli Y, Falagas ME, Giske CG, et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clin Microbiol Infect.* 2012;18: 268–281.

407. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, et al. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N Engl J Med.* 2011;365: 709–717.

408. Gladstone RA, Lo SW, Lees JA, Croucher NJ, van Tonder AJ, Corander J, et al. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine*. 2019;43: 338–346.

409. Bidet P, Bonacorsi S, Clermont O, De Montille C, Brahimi N, Bingen E. Multiple insertional events, restricted by the genetic background, have led to acquisition of pathogenicity island IIJ96-like domains among *Escherichia coli* strains of different clinical origins. *Infect Immun*. 2005;73: 4081–4087.

410. Escobar-Páramo P, Clermont O, Blanc-Potard A-B, Bui H, Le Bouguénec C, Denamur E. A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli. Mol Biol Evol*. 2004;21: 1085–1094.

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

412. Farris JS. Methods for Computing Wagner Trees. *Syst Biol*. 1970;19: 83–92.

413. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol*. 2017;34: 2115–2122.

414. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* 2015;43: D261–9.

415. Dallman T, Cross L, Bishop C, Perry N, Olesen B, Grant KA, et al. Whole genome sequencing of an unusual serotype of Shiga toxin-producing *Escherichia coli*. *Emerg Infect Dis*. 2013;19: 1302–1304.

416. Ewers C, Bethe A, Stamm I, Grobbel M, Kopp PA, Guerra B, et al. CTX-M-15-D-ST648 *Escherichia coli* from companion animals and horses: another pandemic clone combining multiresistance and extraintestinal virulence? *J Antimicrob Chemother*. 2014;69: 1224–1230.

417. Paulshus E, Thorell K, Guzman-Otazo J, Joffre E, Colque P, Kühn I, et al. Repeated Isolation of Extended-Spectrum-β-Lactamase-Positive *Escherichia coli* Sequence Types 648 and 131 from Community Wastewater Indicates that Sewage Systems Are Important Sources of Emerging Clones of Antibiotic-Resistant Bacteria. *Antimicrob Agents Chemother.* 2019;63(9):e00823-19

418. Oteo J, Diestra K, Juan C, Bautista V, Novais A, Pérez-Vázquez M, et al. Extended-spectrum beta-lactamase-producing *Escherichia coli* in Spain belong to a large variety

of multilocus sequence typing types, including ST10 complex/A, ST23 complex/A and ST131/B2. *Int J Antimicrob Agents*. 2009;34: 173–176.

419. Matamoros S, van Hattem JM, Arcilla MS, Willemse N, Melles DC, Penders J, et al. Global phylogenetic analysis of *Escherichia coli* and plasmids carrying the *mcr-1* gene indicates bacterial diversity but plasmid restriction. *Sci Rep*. 2017;7: 15364.

420. Yang F, Yang J, Zhang X, Chen L, Jiang Y, Yan Y, et al. Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res.* 2005;33: 6445–6458.

421. Yong D, Toleman MA, Giske CG, Cho HS, Sundman K, Lee K, et al. Characterization of a New Metallo-β-Lactamase Gene, blaNDM-1, and a Novel Erythromycin Esterase Gene Carried on a Unique Genetic Structure in *Klebsiella pneumoniae* Sequence Type 14 from India. *Antimicrob Agents Chemother.* 2009;53(12):5046-5054.

422. Stoesser N, Sheppard AE, Pankhurst L, De Maio N, Moore CE, Sebra R, et al. Evolutionary History of the Global Emergence of the *Escherichia coli* Epidemic Clone ST131. *MBio*. 2016;7: e02162.

423. Bradley P, den Bakker HC, Rocha EPC, McVean G, Iqbal Z. Ultrafast search of all deposited bacterial and viral genomic data. *Nat Biotechnol*. 2019;37: 152–159.

424. Chang W, Cheng J, Allaire J, Xie Y, McPherson J et al. Shiny: web application framework for R. *RStudio, Inc*. 2017;1.

425. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic blueprint of the human gut microbiota. *Nature*. 2019;568: 499–504.

A Strains, plasmids and oligonucleotides used in this study

AmpR: ampicillin resistant.
CmR: chloramphenicol resistant.

Internal strain ID	Strain name	Genotype/Details	Source/ Reference
Escherichia coli			
MJD841	NEB® 5-alpha	*fhuA2* Δ(*argF- lacZ*)U169 *phoA glnV44* Φ80 Δ(*lacZ*)M15 *gyrA96 recA1 relA1 endA1 thi-* 1 *hsdR17*	New England Biolabs
CF323	MG1655	*Escherichia coli* K-12 MG1655: F– λ– *ilvG– rfb-* 50 *rph-* 1	wild type strain; Gerdes laboratory collection

Plasmid name	Genotype/ Details	Source/Reference
pNDM220	mini- R1 ori; *bla*; *lacI*^{q}; Plac promoter; AmpR	doi.org/ 10.1046/ j.1365-2958.1998.00993.x
pBAD33	p15A ori; *cat*; *araC*; ParaB promoter; CmR	doi: 10.1128/ jb.177.14.4121- 4130.1995
pAH154_doc_v2	Derivative of pNDM220 encoding the doc toxin of the bacteriophage P1 Doc- Phd toxin- antitoxin module. Amp 30 µg/ ml	this study; the ORF of *doc* was amplified from bacteriophage P1vir with a weak RBS (ATTCCTCCaacaattttATG) using primers prAH1542 / prAH1541 and ligated into pNDM220 downstream Plac after digestion of backbone and insert with KpnI / XhoI. IPTG induction of *doc* expression

pAH153_phd_CPH_v1	Derivative of pBAD33 encoding the *phd* antitoxin of the bacteriophage P1 Doc-Phd toxin- antitoxin module. Cam 25 µg/ ml	this study; the ORF of *phd* was amplified from bacteriophage P1vir with a strong RBS (TCAGGAGGatctctATG) using primers prAH1810 / prAH1811 and ligated into pBAD33 downstream ParaB after digestion of backbone and insert with SacI / PstI. L-arabinose induction of phd expression
pMJD119	pNDM220 Plac- doc; AmpR	This study
pMJD120	pNDM220 Plac- 7H; AmpR	This study
pMJD121	pNDM220 Plac- 8H; AmpR	This study
pMJD122	pNDM220 Plac- 12H; AmpR	This study
pMJD123	pNDM220 Plac- 18H; AmpR	This study
pMJD124	pNDM220 Plac- 22H; AmpR	This study
pMJD125	pNDM220 Plac- 31H; AmpR	This study
pMJD126	pNDM220 Plac- 27H; AmpR	This study
pMJD127	pNDM220 Plac- 37H; AmpR	This study
pMJD128	pNDM220 Plac- 51H- 39; AmpR	This study
pMJD129	pNDM220 Plac- 72H; AmpR	This study
pMJD130	pNDM220 Plac- 61H; AmpR	This study
pMJD131	pNDM220 Plac- 51H- 147; AmpR	This study
pMJD132	pNDM220 Plac- 87H; AmpR	This study
pMJD138	pNDM220 Plac- 44H; AmpR	This study
pMJD140	pNDM220 Plac- 54H; AmpR	This study
pMJD142	pBAD33 ParaB- 39P; CmR	This study
pMJD143	pBAD33 ParaB- 44P; CmR	This study
pMJD144	pBAD33 ParaB- 45P; CmR	This study
pMJD145	pBAD33 ParaB- 197P; CmR	This study

pMJD146	pBAD33 ParaB- phd; CmR	This study
pMJD147	pBAD33 ParaB- 3P; CmR	This study
pMJD148	pBAD33 ParaB- 24P; CmR	This study
pMJD149	pBAD33 ParaB- 26P; CmR	This study
pMJD150	pBAD33 ParaB- 27P; CmR	This study
pMJD151	pBAD33 ParaB- 48P; CmR	This study
pMJD152	pBAD33 ParaB- 52P- 31; CmR	This study
pMJD153	pBAD33 ParaB- 62P; CmR	This study
pMJD154	pBAD33 ParaB- 67P; CmR	This study
pMJD155	pBAD33 ParaB- 147P; CmR	This study
pMJD156	pBAD33 ParaB- 168P; CmR	This study

Primer ID	Description	Sequence 5'- 3'
prAH1541	rv. Amplification of *doc* from phage P1vir. XhoI RS	GCCTTCCCTCGAGCTACTCCGCAGAACCATACAA
prAH1542	fw. Amplification of *doc* from phage P1vir. KpnI RS	CGAGTGGGTACCATTCCTCCAACAATTTTATGAGGCATATATCACCGGA
prAH1810	fw. Amplification of *phd* from phage P1vir. SacI RS	GTTGTCGAGCTCTCAGGAGGATCTCTATGCAATCCATTAACTTCCGT
prAH1811	rv. Amplification of *phd* from phage P1vir. PstI RS	CTGGGGTCTGCAGTTATCGGTTAACCAGTTCCTTG
prAH_pNDM220	fw. Screening for cloning in pNDM220	AAAACAGGAAGGCAAAATGC
prAH500	rv. Screening for cloning in pNDM220 and pBAD33	CTGTTTTATCAGACCGCTTC
prAH501	fw. Screening for cloning in pBAD33	CGTCACACTTTGCTATGCC

B Identified toxin groups

Name	Pfam Profile	Type	Antitoxins	Count	Category	local aa identity (min-max)	range alignment length (min-max)	global aa identity (min-max)
11H	Cpta_toxin	IV	35P	258	ubiq	99.79 (97.06-100.0)	99.56 (43.0-100.0)	99.3 (43.0-100.0)
5H	Polyketide_cyc2	II	15P	258	ubiq	99.27 (92.16-100.0)	99.51 (91.0-100.0)	98.9 (87.0-100.0)
34H	Fic	II	173P/ 34P	255	ubiq	98.7 (93.0-100.0)	100.0 (100.0-100.0)	98.52 (93.0-100.0)
10H	HD	II	81P/ 112P/ 32P/ 65P/ 236P	251	ubiq	98.27 (90.68-100.0)	99.37 (52.0-100.0)	97.76 (51.0-100.0)
22H	GNAT_acetyltran	II	8P/ 98P	250	ubiq	99.43 (95.93-100.0)	100.0 (100.0-100.0)	99.39 (96.0-100.0)
27H	HipA_C	II	59P/ 244P/ 24P	242	ubiq	98.16 (63.33-100.0)	97.59 (9.0-100.0)	97.09 (25.0-100.0)
8H	DUF3749	II	12P	236	ubiq	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
17H	RelE	II	73P/ 36P/ 9P	282	species associated	96.97 (53.75-100.0)	76.76 (14.0-100.0)	75.08 (10.0-100.0)
7H	Fic	II	168P/ 3P	240	species associated	99.29 (96.72-100.0)	100.0 (100.0-100.0)	99.16 (97.0-100.0)
21H	PIN	II	212P/ 31P/ 5P	238	species associated	99.5 (95.31-100.0)	98.33 (41.0-100.0)	97.74 (4.0-100.0)
16H	ParE_toxin	II	4P/ 7P	235	species associated	96.15 (77.66-100.0)	99.98 (77.0-100.0)	96.27 (78.0-100.0)
26H	ParE_toxin	II	127P/ 1P	230	species associated	98.02 (91.43-100.0)	96.1 (58.0-100.0)	94.17 (41.0-100.0)
13H	Gp49	II	30P	156	species associated	97.75 (90.16-100.0)	94.4 (50.0-100.0)	92.16 (44.0-100.0)
9H	Zeta_toxin	II	63P/ 11P	95	species associated	99.23 (95.97-100.0)	98.94 (95.0-100.0)	98.26 (93.0-100.0)

2H	PemK_toxin	II	115P/ 171P/ 101P/ 146P/ 22P	65	species associated	87.91 (72.73-100.0)	100.0 (100.0-100.0)	87.93 (73.0-100.0)
42H	YdaT_toxin	II	119P/ 133P/ 111P/ 54P/ 28P/ 210P/ 47P	63	species associated	91.39 (78.26-100.0)	100.0 (100.0-100.0)	91.33 (78.0-100.0)
37H	AntA	II	26P/ 62P/ 42P	45	species associated	86.91 (36.71-100.0)	74.74 (6.0-100.0)	66.75 (14.0-100.0)
25H	Gp49	II	70P/ 77P	40	species associated	97.94 (94.92-100.0)	98.75 (75.0-100.0)	96.76 (74.0-100.0)
33H	RelE	II	155P/ 19P	26	species associated	86.54 (70.91-100.0)	97.76 (70.0-100.0)	84.76 (52.0-100.0)
87H	HicA_toxin	II	45P	8	species associated	98.08 (95.6-100.0)	97.0 (93.0-100.0)	95.25 (89.0-100.0)
51H	HigB_toxin	II	147P/ 39P	7	species associated	88.72 (71.84-100.0)	76.67 (51.0-100.0)	67.14 (43.0-100.0)
64H	Fic	II	124P/ 201P	5	species associated	98.16 (95.76-100.0)	83.2 (58.0-100.0)	82.0 (55.0-100.0)
18H	CcdB	II	143P/ 21P/ 58P	199	sporadic	96.65 (77.78-100.0)	87.27 (16.0-100.0)	86.09 (12.0-100.0)
4H	PIN	II	204P/ 6P/ 238P/ 240P/ 51P/ 103P	156	sporadic	92.87 (77.78-100.0)	94.84 (36.0-100.0)	90.48 (30.0-100.0)
6H	Gp49	II	175P/ 17P	120	sporadic	99.04 (93.33-100.0)	99.53 (77.0-100.0)	98.53 (73.0-100.0)
14H	Gp49	II	27P	105	sporadic	99.73 (96.0-100.0)	96.28 (43.0-100.0)	96.17 (42.0-100.0)
23H	YdaT_toxin	II	93P/ 232P/ 123P/ 117P/ 76P/ 121P/ 33P/ 192P/ 221P/ 176P/ 80P/ 61P	98	sporadic	85.29 (71.91-100.0)	98.17 (50.0-100.0)	83.97 (41.0-100.0)
15H	HigB_toxin	II	16P	96	sporadic	99.9 (99.0-100.0)	100.0 (100.0-100.0)	99.9 (99.0-100.0)
3H	RES	II	92P/ 106P/ 69P/ 13P/ 231P	70	sporadic	98.97 (94.87-100.0)	99.77 (92.0-100.0)	98.63 (88.0-100.0)

35H	HigB_toxin	II	14P	59	sporadic	99.97 (99.03-100.0)	100.0 (100.0-100.0)	99.97 (99.0-100.0)
61H	CcdB	II	48P	57	sporadic	98.19 (94.06-100.0)	98.07 (45.0-100.0)	96.25 (44.0-100.0)
45H	CbtA_toxin	IV	74P/ 170P/ 132P/ 75P/ 234P/ 179P/ 38P/ 157P	46	sporadic	77.84 (53.33-100.0)	87.27 (25.0-100.0)	70.49 (22.0-100.0)
62H	ParE_toxin	II	95P/ 41P/ 79P	42	sporadic	96.39 (71.11-100.0)	85.09 (67.0-100.0)	81.91 (62.0-100.0)
20H	NTP_transf_2	II	223P/ 207P/ 66P/ 205P/ 40P	30	sporadic	96.74 (82.81-100.0)	84.87 (20.0-100.0)	84.38 (17.0-100.0)
57H	NTP_transf_2	II	29P	27	sporadic	99.92 (98.96-100.0)	100.0 (100.0-100.0)	99.93 (99.0-100.0)
28H	Bro- N	II	149P/ 109P/ 120P	26	sporadic	68.95 (41.85-100.0)	98.18 (91.0-100.0)	68.25 (40.0-100.0)
12H	RES	II	105P/ 46P	26	sporadic	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
52H	ParE_toxin	II	2P	22	sporadic	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
44H	ParE_toxin	II	44P/ 23P	19	sporadic	96.7 (77.55-100.0)	94.36 (49.0-100.0)	91.49 (38.0-100.0)
24H	Gp49	II	67P	12	sporadic	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
72H	HD_3	II	50P	10	sporadic	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
71H	DUF955	II	60P	8	sporadic	99.35 (98.86-100.0)	100.0 (100.0-100.0)	99.43 (99.0-100.0)
60H	HipA_C	II	233P/ 20P	8	sporadic	94.75 (82.77-100.0)	100.0 (100.0-100.0)	94.75 (83.0-100.0)
41H	HigB-like_toxin	II	209P/ 154P	7	sporadic	98.34 (94.17-100.0)	100.0 (100.0-100.0)	98.29 (94.0-100.0)
96H	AntA	II	91P	6	sporadic	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)

1H	Bro- N	II	118P/ 136P	15	rare	98.83 (97.31-100.0)	91.62 (80.0-100.0)	90.75 (78.0-100.0)
49H	Gp49	II	78P	15	rare	99.63 (97.25-100.0)	99.75 (99.0-100.0)	99.35 (97.0-100.0)
70H	CbtA_toxin	IV	185P/ 237P/ 88P/ 49P/ 174P	13	rare	87.47 (60.48-100.0)	59.33 (28.0-100.0)	46.37 (12.0-100.0)
32H	AbiEii	IV	102P	10	rare	99.86 (99.3-100.0)	100.0 (100.0-100.0)	99.8 (99.0-100.0)
40H	PemK_toxin	II	87P	9	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
19H	DUF488	II	181P/ 128P	9	rare	99.24 (98.31-100.0)	61.78 (33.0-100.0)	71.64 (34.0-99.0)
29H	YdaT_toxin	II	126P/ 104P/ 68P/ 182P	8	rare	88.83 (74.56-100.0)	92.71 (83.0-100.0)	84.5 (65.0-100.0)
56H	YdaT_toxin	II	218P/ 99P	8	rare	95.67 (91.16-100.0)	85.07 (41.0-100.0)	81.11 (40.0-100.0)
73H	YdaT_toxin	II	76P/ 80P/ 123P	7	rare	99.39 (98.4-100.0)	93.71 (78.0-100.0)	92.95 (78.0-100.0)
67H	Gp49	II	152P	7	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
30H	YdaT_toxin	II	68P	6	rare	96.22 (92.91-100.0)	100.0 (100.0-100.0)	96.27 (93.0-100.0)
31H	HicA_toxin	II	52P	6	rare	96.72 (93.85-100.0)	100.0 (100.0-100.0)	96.8 (94.0-100.0)
53H	Bro- N	II	144P	5	rare	85.94 (77.41-100.0)	100.0 (100.0-100.0)	86.1 (78.0-100.0)
107H	Bro- N	II	107P	5	rare	99.16 (97.89-100.0)	100.0 (100.0-100.0)	99.2 (98.0-100.0)
115H	ParE_toxin	II	113P/ 196P	5	rare	95.11 (92.05-100.0)	100.0 (100.0-100.0)	95.1 (92.0-100.0)
106H	AbiEii	IV	184P/ 227P/ 166P	5	rare	96.24 (90.37-100.0)	73.9 (61.0-100.0)	66.6 (40.0-100.0)
50H	DUF955	II	55P	5	rare	99.5 (98.75-100.0)	100.0 (100.0-100.0)	99.6 (99.0-100.0)

86H	Fic	II	135P/ 43P	5	rare	99.8 (99.49-100.0)	100.0 (100.0-100.0)	99.6 (99.0-100.0)
74H	AntA	II	199P	4	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
109H	CcdB	II	64P	4	rare	97.67 (97.2-98.13)	100.0 (100.0-100.0)	97.5 (97.0-98.0)
131H	PemK_toxin	II	178P/ 158P/ 110P	4	rare	99.53 (99.07-100.0)	100.0 (100.0-100.0)	99.5 (99.0-100.0)
54H	Bro- N	II	52P/ 197P	4	rare	94.1 (88.21-100.0)	92.5 (85.0-100.0)	87.17 (75.0-100.0)
48H	Gp49	II	134P	3	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
68H	Bro- N	II	186P/ 139P	3	rare	100.0 (100.0-100.0)	94.67 (92.0-100.0)	94.67 (92.0-100.0)
77H	HigB-like_toxin	II	164P	3	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
100H	PemK_toxin	II	211P/ 245P	3	rare	86.96 (81.82-96.51)	85.33 (78.0-100.0)	74.0 (65.0-82.0)
94H	ParE_toxin	II	225P/ 214P	3	rare	85.29 (77.94-100.0)	76.0 (64.0-100.0)	66.0 (49.0-100.0)
82H	ParE_toxin	II	189P/ 96P	3	rare	91.93 (90.53-93.68)	100.0 (100.0-100.0)	92.33 (91.0-94.0)
39H	HicA_toxin	II	193P	3	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
88H	AbiEii	IV	188P	3	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
118H	CcdB	II	163P	2	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
63H	HD_3	II	202P	2	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
130H	Fic	II	220P/ 129P	2	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
122H	ParE_toxin	II	160P	2	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)

113H	YafQ_toxin	II	138P	2	rare	97.56 (97.56-97.56)	100.0 (100.0-100.0)	98.0 (98.0-98.0)
123H	Fic	II	116P	2	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
66H	Peptidase_M78	II	222P	2	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
139H	YafO_toxin	II	190P	2	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
83H	PIN	II	100P	2	rare	89.55 (89.55-89.55)	100.0 (100.0-100.0)	90.0 (90.0-90.0)
90H	HigB_toxin	II	183P	2	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
84H	DUF955	II	215P	2	rare	99.47 (99.47-99.47)	100.0 (100.0-100.0)	99.0 (99.0-99.0)
93H	YafO_toxin	II	108P	2	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
65H	PemK_toxin	II	142P	2	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
133H	CbtA_toxin	IV	198P	2	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
125H	Bro- N	II	26P/ 107P	2	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
114H	AntA	II	150P	2	rare	98.44 (98.44-98.44)	100.0 (100.0-100.0)	98.0 (98.0-98.0)
78H	HD	II	195P	2	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
126H	CbtA_toxin	IV	145P	2	rare	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
38H	HipA_C	II	137P	1	rare	NA	NA	NA
129H	RelE	II	131P	1	rare	NA	NA	NA
141H	ParE_toxin	II	141P	1	rare	NA	NA	NA
81H	DUF3749	II	167P	1	rare	NA	NA	NA

121H	HicA_toxin	II	206P	1	rare	NA	NA	NA
89H	HD	II	140P	1	rare	NA	NA	NA
111H	Zeta_toxin	II	156P	1	rare	NA	NA	NA
92H	HigB-like_toxin	II	208P	1	rare	NA	NA	NA
127H	YafQ_toxin	II	53P	1	rare	NA	NA	NA
59H	ANT	II	97P/ 180P	1	rare	NA	NA	NA
135H	DUF4111	II	247P	1	rare	NA	NA	NA
97H	Fic	II	82P	1	rare	NA	NA	NA
58H	Gp49	II	56P	1	rare	NA	NA	NA
85H	HicA_toxin	II	172P	1	rare	NA	NA	NA
128H	DUF488	II	219P	1	rare	NA	NA	NA
120H	HD	II	194P	1	rare	NA	NA	NA
112H	PIN	II	10P	1	rare	NA	NA	NA
105H	PemK_toxin	II	187P	1	rare	NA	NA	NA
148H	PemK_toxin	II	239P	1	rare	NA	NA	NA
146H	DUF4258	II	162P	1	rare	NA	NA	NA
137H	HipA_C	II	89P	1	rare	NA	NA	NA
104H	HicA_toxin	II	241P/ 246P	1	rare	NA	NA	NA
43H	PIN	II	229P	1	rare	NA	NA	NA
117H	Fic	II	230P	1	rare	NA	NA	NA
145H	ParE_toxin	II	228P	1	rare	NA	NA	NA
98H	ANT	II	161P	1	rare	NA	NA	NA
102H	DUF488	II	191P	1	rare	NA	NA	NA
99H	ParE_toxin	II	130P	1	rare	NA	NA	NA
76H	HigB-like_toxin	II	235P	1	rare	NA	NA	NA

143H	AHSA1	II	85P	1	rare	NA	NA	NA
91H	HipA_C	II	25P	1	rare	NA	NA	NA
110H	YdaT_toxin	II	151P	1	rare	NA	NA	NA
47H	CbtA_toxin	IV	84P	1	rare	NA	NA	NA
142H	HD	II	243P	1	rare	NA	NA	NA
108H	HigB_toxin	II	125P	1	rare	NA	NA	NA
95H	YafO_toxin	II	90P	1	rare	NA	NA	NA
101H	RES	II	169P	1	rare	NA	NA	NA
147H	CcdB	II	242P	1	rare	NA	NA	NA
136H	ParE_toxin	II	224P	1	rare	NA	NA	NA
69H	CbtA_toxin	IV	38P/ 217P	1	rare	NA	NA	NA
132H	Peptidase_M78	II	153P	1	rare	NA	NA	NA
103H	CcdB	II	159P	1	rare	NA	NA	NA
36H	YdaT_toxin	II	213P	1	rare	NA	NA	NA
124H	YdaT_toxin	II	114P	1	rare	NA	NA	NA
138H	YafO_toxin	II	165P	1	rare	NA	NA	NA
119H	PIN	II	6P	1	rare	NA	NA	NA
134H	YafO_toxin	II	71P	1	rare	NA	NA	NA

C Identified antitoxin groups

Name	Pfam Profile	Toxin	Type	Up-stream count	Down stream count	In TADB?	Interpro	Mean local aa identity (min-max)	Mean alignment length (min-max)	Mean global aa identity (min-max)
15P	Polyketide_cyc2	5H	II	0	258	No	Ubiquitin	99.29 (94.79-100.0)	100.0 (100.0-100.0)	99.32 (95.0-100.0)
35P	Cpta_toxin	11H	IV	258	0	Yes	In TADB	99.96 (97.73-100.0)	100.0 (100.0-100.0)	99.97 (98.0-100.0)
34P	Fic	34H	II	255	0	Yes	In TADB	98.78 (89.09-100.0)	99.78 (93.0-100.0)	98.57 (85.0-100.0)
8P	GNAT_acetyltran	22H	II	0	250	No	Inner membrane transporter	98.66 (90.6-100.0)	99.44 (66.0-100.0)	98.07 (62.0-100.0)
36P	RelE	17H	II	0	243	Yes	In TADB	99.51 (81.03-100.0)	99.03 (52.0-100.0)	98.53 (43.0-100.0)
24P	HipA_C	27H	II	242	0	Yes	In TADB	98.04 (88.17-100.0)	99.41 (93.0-100.0)	97.55 (85.0-100.0)
3P	Fic	7H	II	240	0	Yes	In TADB	99.69 (97.26-100.0)	99.64 (61.0-100.0)	99.45 (61.0-100.0)
9P	RelE	17H	II	238	0	No	None	95.43 (85.71-100.0)	100.0 (100.0-100.0)	95.39 (86.0-100.0)
12P	DUF3749	8H	II	0	236	No	DUF	98.84 (94.44-100.0)	98.03 (48.0-100.0)	96.96 (47.0-100.0)
5P	PIN	21H	II	236	0	Yes	In TADB	99.87 (97.56-100.0)	100.0 (100.0-100.0)	99.89 (98.0-100.0)
7P	ParE_toxin	16H	II	235	0	Yes	In TADB	99.55 (94.5-100.0)	95.75 (80.0-100.0)	95.37 (77.0-100.0)
1P	ParE_toxin	26H	II	0	229	No	None	97.23 (85.45-100.0)	100.0 (100.0-100.0)	97.15 (85.0-100.0)
32P	HD	10H	II	208	0	No	None	98.29 (88.04-100.0)	99.84 (99.0-100.0)	98.1 (87.0-100.0)
6P	PIN	119H/4H	II	157	0	Yes	In TADB	94.45 (67.74-100.0)	95.9 (36.0-100.0)	93.91 (32.0-100.0)

30P	Gp49	13H	II	0	156	No	DNA binding	98.75 (94.95-100.0)	98.59 (58.0-100.0)	97.97 (59.0-100.0)
4P	ParE_toxin	16H	II	0	153	No	None	91.48 (80.3-100.0)	100.0 (100.0-100.0)	91.31 (80.0-100.0)
21P	CcdB	18H	II	0	142	No	None	98.37 (86.57-100.0)	98.07 (88.0-100.0)	96.92 (77.0-100.0)
58P	CcdB	18H	II	126	0	No	None	98.3 (91.94-100.0)	99.91 (98.0-100.0)	98.16 (92.0-100.0)
17P	Gp49	6H	II	0	120	No	DNA binding	99.15 (94.74-100.0)	100.0 (100.0-100.0)	99.19 (95.0-100.0)
27P	Gp49	14H	II	0	105	No	DNA binding	99.59 (95.65-100.0)	99.5 (74.0-100.0)	99.14 (72.0-100.0)
16P	HigB_toxin	15H	II	0	96	Yes	In TADB	99.27 (95.0-100.0)	100.0 (100.0-100.0)	99.24 (95.0-100.0)
11P	Zeta_toxin	9H	II	0	85	No	ABC transporter	96.52 (64.0-100.0)	72.69 (14.0-100.0)	78.73 (38.0-100.0)
13P	RES	3H	II	70	0	Yes	In TADB	98.86 (95.92-100.0)	100.0 (100.0-100.0)	98.84 (96.0-100.0)
14P	HigB_toxin	35H	II	0	59	Yes	In TADB	99.98 (99.24-100.0)	99.8 (94.0-100.0)	99.76 (94.0-100.0)
48P	CcdB	61H	II	57	0	Yes	In TADB	98.55 (93.06-100.0)	98.14 (73.0-100.0)	97.1 (69.0-100.0)
80P	YdaT_toxin	73H/ 23H	II	54	0	No	antitoxin	87.14 (71.05-100.0)	96.81 (86.0-100.0)	85.48 (67.0-100.0)
28P	YdaT_toxin	42H	II	0	50	No	None	99.08 (93.33-100.0)	100.0 (100.0-100.0)	99.01 (93.0-100.0)
22P	PemK_toxin	2H	II	45	0	Yes	In TADB	98.15 (93.58-100.0)	90.87 (47.0-100.0)	91.31 (46.0-100.0)
62P	AntA	37H	II	42	0	No	consensus disorder prediction	76.99 (37.5-100.0)	72.79 (21.0-100.0)	69.99 (22.0-100.0)
95P	ParE_toxin	62H	II	42	0	Yes	In TADB	96.86 (89.02-100.0)	100.0 (100.0-100.0)	96.79 (89.0-100.0)
70P	Gp49	25H	II	40	0	No	None	94.13 (86.96-100.0)	79.86 (58.0-100.0)	74.98 (47.0-100.0)
74P	CbtA_toxin	45H	IV	39	0	No	antitoxin	82.66 (70.37-100.0)	93.19 (74.0-100.0)	77.99 (58.0-100.0)

77P	Gp49	25H	II	0	39	No	None	98.89 (96.0-100.0)	98.26 (67.0-100.0)	97.03 (64.0-100.0)
47P	YdaT_toxin	42H	II	37	0	No	None	99.85 (97.18-100.0)	100.0 (100.0-100.0)	99.89 (97.0-100.0)
73P	RelE	17H	II	32	0	No	None	93.8 (57.32-100.0)	90.54 (54.0-100.0)	86.26 (47.0-100.0)
26P	AntA	37H/ 125H	II	0	30	No	DUF	92.9 (76.27-100.0)	100.0 (100.0-100.0)	92.91 (76.0-100.0)
29P	NTP_transf_2	57H	II	0	27	Yes	In TADB	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
105P	RES	12H	II	0	26	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
46P	RES	12H	II	26	0	No	None	99.7 (96.08-100.0)	100.0 (100.0-100.0)	99.69 (96.0-100.0)
19P	RelE	33H	II	0	25	Yes	In TADB	90.64 (78.72-100.0)	98.48 (81.0-100.0)	89.43 (65.0-100.0)
76P	YdaT_toxin	73H/ 23H	II	0	25	No	consensus disorder prediction	74.64 (44.3-100.0)	79.09 (24.0-100.0)	59.88 (16.0-100.0)
81P	HD	10H	II	24	0	No	None	98.87 (96.36-100.0)	100.0 (100.0-100.0)	98.76 (96.0-100.0)
109P	Bro- N	28H	II	22	0	No	regulation of transcriptio n	94.63 (85.45-100.0)	95.92 (92.0-100.0)	91.18 (78.0-100.0)
2P	ParE_toxin	52H	II	22	0	Yes	In TADB	99.34 (98.73-100.0)	100.0 (100.0-100.0)	99.48 (99.0-100.0)
40P	NTP_transf_2	20H	II	21	0	No	DUF	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
101P	PemK_toxi n	2H	II	20	0	Yes	In TADB	98.04 (95.29-100.0)	100.0 (100.0-100.0)	97.82 (95.0-100.0)
44P	ParE_toxin	44H	II	19	0	No	None	97.59 (80.46-100.0)	99.05 (91.0-100.0)	96.85 (73.0-100.0)
65P	HD	10H	II	18	0	No	None	98.47 (95.51-100.0)	100.0 (100.0-100.0)	98.64 (96.0-100.0)
23P	ParE_toxin	44H	II	0	16	No	None	96.56 (84.75-100.0)	98.52 (88.0-100.0)	95.28 (76.0-100.0)

41P	ParE_toxin	62H	II	0	15	No	None	99.01 (96.67-100.0)	95.33 (65.0-100.0)	94.36 (62.0-100.0)
78P	Gp49	49H	II	0	15	Yes	In TADB	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
67P	Gp49	24H	II	0	12	Yes	In TADB	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
68P	YdaT_toxin	30H/ 29H	II	0	12	No	None	90.35 (80.95-100.0)	99.39 (98.0-100.0)	89.61 (80.0-100.0)
136P	Bro- N	1H	II	0	11	No	None	99.77 (98.72-100.0)	100.0 (100.0-100.0)	99.82 (99.0-100.0)
54P	YdaT_toxin	42H	II	11	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
61P	YdaT_toxin	23H	II	11	0	No	antitoxin	84.86 (73.61-100.0)	98.73 (93.0-100.0)	84.45 (70.0-100.0)
93P	YdaT_toxin	23H	II	11	0	No	None	99.37 (98.85-100.0)	100.0 (100.0-100.0)	99.45 (99.0-100.0)
102P	AbiEii	32H	IV	0	10	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
50P	HD_3	72H	II	0	10	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
51P	PIN	4H	II	0	10	No	None	97.94 (93.83-100.0)	100.0 (100.0-100.0)	97.96 (94.0-100.0)
63P	Zeta_toxin	9H	II	10	0	No	None	96.3 (90.74-100.0)	100.0 (100.0-100.0)	96.24 (91.0-100.0)
33P	YdaT_toxin	23H	II	9	0	No	antitoxin	99.12 (97.37-100.0)	100.0 (100.0-100.0)	99.25 (97.0-100.0)
49P	CbtA_toxin	70H	IV	9	0	No	antitoxin	94.38 (88.23-100.0)	100.0 (100.0-100.0)	94.28 (88.0-100.0)
87P	PemK_toxin	40H	II	9	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
92P	RES	3H	II	0	9	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
123P	YdaT_toxin	73H/ 23H	II	8	0	No	antitoxin	93.76 (80.0-100.0)	100.0 (100.0-100.0)	93.79 (80.0-100.0)

132P	CbtA_toxin	45H	IV	0	8	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
146P	PemK_toxin	2H	II	0	8	No	None	100.0 (100.0-100.0)	86.5 (46.0-100.0)	86.5 (46.0-100.0)
168P	Fic	7H	II	0	8	No	None	96.12 (87.72-100.0)	100.0 (100.0-100.0)	95.54 (86.0-100.0)
42P	AntA	37H	II	0	8	No	None	89.29 (73.68-100.0)	97.86 (96.0-100.0)	87.71 (72.0-100.0)
45P	HicA_toxin	87H	II	0	8	Yes	In TADB	99.79 (99.16-100.0)	100.0 (100.0-100.0)	99.75 (99.0-100.0)
60P	DUF955	71H	II	8	0	No	DNA binding	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
107P	Bro- N	125H/ 107H	II	7	0	No	prophage	91.16 (78.57-100.0)	100.0 (100.0-100.0)	91.14 (79.0-100.0)
152P	Gp49	67H	II	0	7	Yes	In TADB	99.7 (98.95-100.0)	100.0 (100.0-100.0)	99.71 (99.0-100.0)
181P	DUF488	19H	II	0	7	No	None	100.0 (100.0-100.0)	93.71 (78.0-100.0)	93.71 (78.0-100.0)
20P	HipA_C	60H	II	7	0	No	DNA binding	98.25 (95.92-100.0)	100.0 (100.0-100.0)	98.29 (96.0-100.0)
52P	HicA_toxin	54H/ 31H	II	6	1	No	DUF	98.77 (97.65-100.0)	92.38 (84.0-100.0)	91.33 (82.0-100.0)
99P	YdaT_toxin	56H	II	0	7	No	None	98.83 (95.89-100.0)	100.0 (100.0-100.0)	98.86 (96.0-100.0)
104P	YdaT_toxin	29H	II	6	0	No	antitoxin	98.58 (97.59-100.0)	95.2 (92.0-100.0)	94.0 (90.0-100.0)
111P	YdaT_toxin	42H	II	6	0	No	DNA binding	98.85 (97.33-100.0)	100.0 (100.0-100.0)	99.0 (97.0-100.0)
154P	HigB-like_toxin	41H	II	0	6	No	None	95.36 (88.57-100.0)	100.0 (100.0-100.0)	95.4 (89.0-100.0)
91P	AntA	96H	II	6	0	No	prophage	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
121P	YdaT_toxin	23H	II	5	0	No	antitoxin	100.0 (100.0-100.0)	99.6 (99.0-100.0)	99.6 (99.0-100.0)

124P	Fic	64H	II	5	0	No	consensus disorder prediction	92.36 (82.02-100.0)	100.0 (100.0-100.0)	92.4 (82.0-100.0)
135P	Fic	86H	II	0	5	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
144P	Bro- N	53H	II	5	0	No	None	98.42 (96.05-100.0)	100.0 (100.0-100.0)	98.4 (96.0-100.0)
39P	HigB_toxin	51H	II	0	5	No	DNA binding	99.65 (98.65-100.0)	81.6 (54.0-100.0)	81.2 (53.0-100.0)
43P	Fic	86H	II	5	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
55P	DUF955	50H	II	5	0	No	DNA binding	99.32 (98.29-100.0)	99.2 (98.0-100.0)	98.8 (97.0-100.0)
66P	DUF4111	20H	II	5	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
103P	PIN	4H	II	0	4	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
110P	PemK_toxin	131H	II	4	0	No	None	92.94 (88.23-100.0)	100.0 (100.0-100.0)	92.67 (88.0-100.0)
118P	Bro- N	1H	II	0	4	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
133P	YdaT_toxin	42H	II	4	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
157P	CbtA_toxin	45H	IV	0	4	No	None	93.33 (87.69-100.0)	100.0 (100.0-100.0)	93.5 (88.0-100.0)
173P	Fic	34H	II	0	4	No	Glutamine amidotransferase	99.18 (98.36-100.0)	94.5 (91.0-99.0)	94.0 (90.0-99.0)
199P	AntA	74H	II	4	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
38P	CbtA_toxin	45H/69H	IV	4	0	No	antitoxin	79.12 (73.0-84.61)	91.17 (84.0-100.0)	71.83 (60.0-84.0)
64P	CcdB	109H	II	4	0	No	consensus disorder prediction	97.58 (95.16-100.0)	100.0 (100.0-100.0)	97.5 (95.0-100.0)
98P	GNAT_acetyltran	22H	II	4	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)

113P	ParE_toxin	115H	II	3	0	Yes	In TADB	88.55 (87.84-89.19)	86.33 (80.0-93.0)	77.33 (72.0-84.0)
117P	YdaT_toxin	23H	II	3	0	No	antitoxin	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
119P	YdaT_toxin	42H	II	3	0	No	None	99.1 (98.65-100.0)	100.0 (100.0-100.0)	99.33 (99.0-100.0)
134P	Gp49	48H	II	0	3	Yes	In TADB	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
164P	HigB-like_toxin	77H	II	3	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
185P	CbtA_toxin	70H	IV	3	0	No	toxin domain	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
188P	AbiEii	88H	IV	3	0	No	regulation of transcription	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
193P	HicA_toxin	39H	II	3	0	No	DUF	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
197P	Bro- N	54H	II	3	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
88P	CbtA_toxin	70H	IV	0	3	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
96P	ParE_toxin	82H	II	3	0	Yes	In TADB	90.36 (86.75-92.77)	100.0 (100.0-100.0)	90.67 (87.0-93.0)
100P	PIN	83H	II	2	0	No	antitoxin	86.91 (86.91-86.91)	99.0 (99.0-99.0)	86.0 (86.0-86.0)
108P	YafO_toxin	93H	II	2	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
116P	Fic	123H	II	2	0	No	antitoxin	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
120P	Bro- N	28H	II	2	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
126P	YdaT_toxin	29H	II	2	0	No	None	95.51 (95.51-95.51)	100.0 (100.0-100.0)	96.0 (96.0-96.0)
128P	DUF488	19H	II	2	0	No	DUF	97.89 (97.89-97.89)	92.0 (92.0-92.0)	90.0 (90.0-90.0)

138P	YafQ_toxin	113H	II	0	2	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
142P	PemK_toxin	65H	II	2	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
145P	CbtA_toxin	126H	IV	0	2	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
147P	HigB_toxin	51H	II	0	2	No	DNA binding	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
149P	Bro- N	28H	II	2	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
150P	AntA	114H	II	2	0	No	Endodeoxyribonuclease	97.73 (97.73-97.73)	100.0 (100.0-100.0)	98.0 (98.0-98.0)
160P	ParE_toxin	122H	II	2	0	No	DNA integration	98.65 (98.65-98.65)	61.0 (61.0-61.0)	60.0 (60.0-60.0)
163P	CcdB	118H	II	2	0	No	antitoxin	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
166P	AbiEii	106H	IV	0	2	No	None	96.49 (96.49-96.49)	64.0 (64.0-64.0)	63.0 (63.0-63.0)
179P	CbtA_toxin	45H	IV	2	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
183P	HigB_toxin	90H	II	2	0	No	None	98.44 (98.44-98.44)	100.0 (100.0-100.0)	98.0 (98.0-98.0)
184P	AbiEii	106H	IV	2	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
186P	Bro- N	68H	II	2	0	No	None	96.3 (96.3-96.3)	100.0 (100.0-100.0)	96.0 (96.0-96.0)
190P	YafO_toxin	139H	II	2	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
192P	YdaT_toxin	23H	II	2	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
195P	HD	78H	II	0	2	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
196P	ParE_toxin	115H	II	2	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)

198P	CbtA_toxin	133H	IV	2	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
202P	HD_3	63H	II	2	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
207P	DUF4111	20H	II	2	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
210P	YdaT_toxin	42H	II	2	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
211P	PemK_toxin	100H	II	2	0	No	antitoxin	87.65 (87.65-87.65)	100.0 (100.0-100.0)	88.0 (88.0-88.0)
214P	ParE_toxin	94H	II	0	2	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
215P	DUF955	84H	II	2	0	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
222P	Peptidase_M78	66H	II	0	2	No	None	100.0 (100.0-100.0)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
232P	YdaT_toxin	23H	II	0	2	No	None	98.84 (98.84-98.84)	100.0 (100.0-100.0)	99.0 (99.0-99.0)
31P	PIN	21H	II	2	0	No	toxin domain	98.08 (98.08-98.08)	98.0 (98.0-98.0)	96.0 (96.0-96.0)
69P	RES	3H	II	0	2	No	None	100.0 (100.0-100.0)	96.0 (96.0-96.0)	96.0 (96.0-96.0)
106P	RES	3H	II	0	1	No	None	100	100	100
10P	PIN	112H	II	0	1	No	consensus disorder prediction	100	100	100
112P	HD	10H	II	1	0	No	None	100	100	100
114P	YdaT_toxin	124H	II	1	0	No	DNA binding	100	100	100
115P	PemK_toxin	2H	II	0	1	No	None	100	100	100
125P	HigB_toxin	108H	II	0	1	No	None	100	100	100
127P	ParE_toxin	26H	II	0	1	Yes	In TADB	100	100	100
129P	Fic	130H	II	0	1	No	None	100	100	100

130P	ParE_toxin	99H	II	1	0	Yes	In TADB	100	100	100
131P	RelE	129H	II	0	1	Yes	In TADB	100	100	100
137P	HipA_C	38H	II	1	0	Yes	In TADB	100	100	100
139P	Bro- N	68H	II	0	1	No	None	100	100	100
140P	HD	89H	II	1	0	No	None	100	100	100
141P	ParE_toxin	141H	II	1	0	Yes	In TADB	100	100	100
143P	CcdB	18H	II	0	1	No	toxin domain	100	100	100
151P	YdaT_toxin	110H	II	1	0	No	antitoxin	100	100	100
153P	Peptidase_M78	132H	II	1	0	No	DNA binding	100	100	100
155P	RelE	33H	II	1	0	No	consensus disorder prediction	100	100	100
156P	Zeta_toxin	111H	II	1	0	No	ABC transporter	100	100	100
158P	PemK_toxin	131H	II	0	1	No	None	100	100	100
159P	CcdB	103H	II	1	0	No	antitoxin	100	100	100
161P	ANT	98H	II	1	0	No	Endodeoxyribonuclease	100	100	100
162P	DUF4258	146H	II	0	1	No	None	100	100	100
165P	YafO_toxin	138H	II	1	0	No	None	100	100	100
167P	DUF3749	81H	II	0	1	No	consensus disorder prediction	100	100	100
169P	RES	101H	II	0	1	No	None	100	100	100
170P	CbtA_toxin	45H	IV	0	1	No	None	100	100	100
171P	PemK_toxin	2H	II	0	1	No	None	100	100	100
172P	HicA_toxin	85H	II	0	1	No	antitoxin	100	100	100
174P	CbtA_toxin	70H	IV	1	0	No	antitoxin	100	100	100
175P	Gp49	6H	II	1	0	No	consensus disorder prediction	100	100	100

176P	YdaT_toxin	23H	II	1	0	No	None	100	100	100
178P	PemK_toxin	131H	II	0	1	No	None	100	100	100
180P	ANT	59H	II	1	0	No	prophage	100	100	100
182P	YdaT_toxin	29H	II	0	1	No	None	100	100	100
187P	PemK_toxin	105H	II	1	0	Yes	In TADB	100	100	100
189P	ParE_toxin	82H	II	0	1	No	None	100	100	100
191P	DUF488	102H	II	1	0	No	consensus disorder prediction	100	100	100
194P	HD	120H	II	1	0	No	None	100	100	100
201P	Fic	64H	II	0	1	No	None	100	100	100
204P	PIN	4H	II	0	1	No	None	100	100	100
205P	DUF4111	20H	II	1	0	No	None	100	100	100
206P	HicA_toxin	121H	II	1	0	No	DUF	100	100	100
208P	HigB-like_toxin	92H	II	0	1	No	None	100	100	100
209P	HigB-like_toxin	41H	II	0	1	No	None	100	100	100
212P	PIN	21H	II	0	1	No	toxin domain	100	100	100
213P	YdaT_toxin	36H	II	0	1	No	None	100	100	100
217P	CbtA_toxin	69H	IV	0	1	No	None	100	100	100
218P	YdaT_toxin	56H	II	0	1	No	None	100	100	100
219P	DUF488	128H	II	0	1	No	None	100	100	100
220P	Fic	130H	II	0	1	No	None	100	100	100
221P	YdaT_toxin	23H	II	1	0	No	None	100	100	100
223P	DUF4111	20H	II	0	1	No	None	100	100	100
224P	ParE_toxin	136H	II	1	0	Yes	In TADB	100	100	100
225P	ParE_toxin	94H	II	1	0	No	None	100	100	100

227P	AbiEii	106H	IV	1	0	No	None	100	100	100
228P	ParE_toxin	145H	II	1	0	No	None	100	100	100
229P	PIN	43H	II	1	0	Yes	In TADB	100	100	100
230P	Fic	117H	II	1	0	No	None	100	100	100
231P	RES	3H	II	0	1	No	None	100	100	100
233P	HipA_C	60H	II	1	0	No	None	100	100	100
234P	CbtA_toxin	45H	IV	0	1	No	None	100	100	100
235P	HigB-like_toxin	76H	II	0	1	No	antitoxin	100	100	100
236P	HD	10H	II	0	1	No	None	100	100	100
237P	CbtA_toxin	70H	IV	0	1	No	None	100	100	100
238P	PIN	4H	II	0	1	No	None	100	100	100
239P	PemK_toxin	148H	II	0	1	No	None	100	100	100
240P	PIN	4H	II	0	1	No	None	100	100	100
241P	HicA_toxin	104H	II	1	0	No	None	100	100	100
242P	CcdB	147H	II	1	0	Yes	In TADB	100	100	100
243P	HD	142H	II	0	1	No	toxin domain	100	100	100
244P	Couple_hipA	27H	II	0	1	No	None	100	100	100
245P	PemK_toxin	100H	II	1	0	No	None	100	100	100
246P	HicA_toxin	104H	II	0	1	Yes	In TADB	100	100	100
247P	DUF4111	135H	II	1	0	No	consensus disorder prediction	100	100	100
25P	HipA_C	91H	II	1	0	No	None	100	100	100
53P	YafQ_toxin	127H	II	1	0	Yes	In TADB	100	100	100
56P	Gp49	58H	II	0	1	Yes	In TADB	100	100	100
59P	HipA_C	27H	II	0	1	No	None	100	100	100

71P	YafO_toxin	134H	II	1	0	No	None	100	100	100
75P	CbtA_toxin	45H	IV	1	0	No	None	100	100	100
79P	ParE_toxin	62H	II	0	1	No	None	100	100	100
82P	Fic	97H	II	1	0	Yes	In TADB	100	100	100
84P	CbtA_toxin	47H	IV	0	1	No	None	100	100	100
85P	AHSA1	143H	II	1	0	No	consensus disorder prediction	100	100	100
89P	HipA_C	137H	II	0	1	No	DNA binding	100	100	100
90P	YafO_toxin	95H	II	1	0	No	consensus disorder prediction	100	100	100
97P	ANT	59H	II	0	1	No	None	100	100	100

D Identified orphan antitoxin groups

ID	Total Orphans	*K. pneumoniae sensu stricto*	*K. quasipneumoniae*	*K. variicola*	Predicted function	Original Toxin Domain	Toxin type
104P	1	0	1	0	antitoxin	YdaT_toxin	II
106P	59	58	1	0	None	RES	II
107P	34	30	3	1	prophage	Bro-N	II
109P	3	3	0	0	regulation of transcription	Bro-N	II
10P	2	2	0	0	consensus disorder prediction	PIN	II
114P	10	5	5	0	DNA binding	YdaT_toxin	II
115P	36	0	19	17	None	PemK_toxin	II
118P	135	135	0	0	None	Bro-N	II
11P	1	1	0	0	ABC transporter	Zeta_toxin	II
120P	2	1	1	0	None	Bro-N	II
124P	248	219	15	14	consensus disorder prediction	Fic	II
126P	1	0	1	0	None	YdaT_toxin	II
127P	28	8	19	1	In TADB	ParE_toxin	II
128P	8	6	0	2	DUF	DUF488	II
129P	2	0	2	0	None	Fic	II
12P	11	10	0	1	DUF	DUF3749	II
132P	5	5	0	0	None	CbtA_toxin	IV
134P	6	6	0	0	In TADB	Gp49	II
136P	47	47	0	0	None	Bro-N	II
137P	2	2	0	0	In TADB	HipA_C	II

139P	8	4	4	0	None	Bro-N	II
13P	25	21	4	0	In TADB	RES	II
142P	2	2	0	0	None	PemK_toxin	II
143P	110	109	0	1	toxin domain	CcdB	II
144P	2	1	1	0	None	Bro-N	II
146P	3	3	0	0	None	PemK_toxin	II
14P	2	2	0	0	In TADB	HigB_toxin	II
150P	13	12	0	1	Endodeoxyribonuclease	AntA	II
156P	30	27	0	3	ABC transporter	Zeta_toxin	II
15P	1	1	0	0	Ubiquitin	Polyketide_cyc2	II
160P	3	3	0	0	DNA integration	ParE_toxin	II
161P	77	69	2	6	Endodeoxyribonuclease	ANT	II
162P	4	4	0	0	None	DUF4258	II
166P	8	8	0	0	None	AbiEii	IV
167P	37	32	0	5	consensus disorder prediction	DUF3749	II
16P	1	1	0	0	In TADB	HigB_toxin	II
172P	1	1	0	0	antitoxin	HicA_toxin	II
173P	3	3	0	0	Glutamine amidotransferase	Fic	II
181P	21	19	2	0	None	DUF488	II
184P	195	179	0	16	None	AbiEii	IV
185P	6	2	2	2	toxin domain	CbtA_toxin	IV
188P	1	1	0	0	regulation of transcription	AbiEii	IV
191P	1	1	0	0	consensus disorder prediction	DUF488	II
194P	4	4	0	0	None	HD	II
195P	6	5	0	1	None	HD	II

19P	1	0	0	1	In TADB	RelE	II
1P	10	9	0	1	None	ParE_toxin	II
202P	12	12	0	0	None	HD_3	II
204P	4	2	0	2	None	PIN	II
205P	2	2	0	0	None	DUF4111	II
206P	1	1	0	0	DUF	HicA_toxin	II
207P	3	3	0	0	None	DUF4111	II
212P	1	0	0	1	toxin domain	PIN	II
214P	1	1	0	0	None	ParE_toxin	II
21P	42	42	0	0	None	CcdB	II
224P	2	2	0	0	In TADB	ParE_toxin	II
225P	2	2	0	0	None	ParE_toxin	II
22P	8	7	0	1	In TADB	PemK_toxin	II
231P	123	114	6	3	None	RES	II
233P	1	1	0	0	None	HipA_C	II
236P	1	0	0	1	None	HD	II
238P	6	5	1	0	None	PIN	II
23P	1	1	0	0	None	ParE_toxin	II
243P	2	2	0	0	toxin domain	HD	II
244P	1	1	0	0	None	Couple_hipA	II
246P	5	5	0	0	In TADB	HicA_toxin	II
24P	17	16	0	1	In TADB	HipA_C	II
26P	2	1	1	0	DUF	AntA	II
29P	1	1	0	0	In TADB	NTP_transf_2	II
30P	2	2	0	0	DNA binding	Gp49	II

32P	7	6	1	0	None	HD	II
34P	2	2	0	0	In TADB	Fic	II
35P	1	1	0	0	In TADB	Cpta_toxin	II
36P	6	5	0	1	In TADB	RelE	II
38P	2	2	0	0	antitoxin	CbtA_toxin	IV
39P	92	80	9	3	DNA binding	HigB_toxin	II
3P	1	0	0	1	In TADB	Fic	II
40P	4	4	0	0	DUF	DUF4111	II
43P	1	1	0	0	None	Fic	II
45P	45	45	0	0	In TADB	HicA_toxin	II
49P	2	0	2	0	antitoxin	CbtA_toxin	IV
4P	10	10	0	0	None	ParE_toxin	II
51P	2	2	0	0	None	PIN	II
58P	123	123	0	0	None	CcdB	II
59P	251	214	19	18	None	HipA_C	II
60P	1	1	0	0	DNA binding	DUF955	II
62P	1	1	0	0	consensus disorder prediction	AntA	II
63P	8	0	0	8	None	Zeta_toxin	II
66P	80	72	8	0	None	DUF4111	II
68P	2	0	2	0	None	YdaT_toxin	II
6P	4	4	0	0	In TADB	PIN	II
70P	1	0	1	0	None	Gp49	II
73P	16	13	3	0	None	RelE	II
74P	1	1	0	0	antitoxin	CbtA_toxin	IV
75P	12	7	4	1	None	CbtA_toxin	IV

76P	18	18	0	0	consensus disorder prediction	YdaT_toxin	II
77P	2	0	1	1	None	Gp49	II
78P	1	1	0	0	In TADB	Gp49	II
79P	60	52	6	2	None	ParE_toxin	II
7P	11	9	2	0	In TADB	ParE_toxin	II
80P	1	0	1	0	antitoxin	YdaT_toxin	II
89P	18	0	0	18	DNA binding	HipA_C	II
91P	4	3	1	0	prophage	AntA	II
92P	5	5	0	0	None	RES	II
9P	2	1	1	0	None	RelE	II

E Summary of *E. coli* PopPUNK Clusters

PopPUNK Cluster	Genomes	Median Length	Median Genes	Phylo-group	Isolation	Continents	STs	Pathotypes	MDR
1	3266	5.37	5064	E	unknown/ other (0.97); faeces (0.03)	Europe (0.94); North America (0.05)	11 (0.93); 11~ (0.03)	aEPEC/ EPEC (0.05); EHEC (0.95)	No
2	781	5.16	4844	B2	faeces (0.36); blood (0.31); urine (0.24); unknown/ other (0.1)	Europe (0.73); nd (0.12); North America (0.11); Oceania (0.03)	131 (0.99)	ExPEC (0.54); ND (0.46)	Yes
3	463	5.13	4738	B2	blood (0.7); unknown/ other (0.17); urine (0.12)	Europe (0.85); North America (0.11); nd (0.03)	73 (0.93)	ExPEC (0.83); ND (0.16); STEC (0.01)	No
4	363	5.13	4814	B2	blood (0.61); unknown/ other (0.17); urine (0.13); faeces (0.08)	Europe (0.75); North America (0.22)	95 (0.79); 416 (0.07); 421 (0.03); 95~ (0.03)	ExPEC (0.75); EXPEC (0.01); ND (0.24)	No
5	237	5.42	5195	B1	unknown/ other (0.69); faeces (0.31)	North America (0.63); Europe (0.33)	17 (0.78); 1967 (0.11); 386 (0.04); 17~ (0.03)	aEPEC/ EPEC (0.04); EHEC (0.95); STEC (0.01)	No
6	239	5.55	5347	B1	unknown/ other (0.59); faeces (0.41)	North America (0.67); Europe (0.31)	21 (0.87); 29 (0.08)	aEPEC/ EPEC (0.04); EHEC (0.94); STEC (0.01)	No
7	174	5.26	4924.5	D	blood (0.72);	Europe (0.83);	69 (0.94);	EAEC (0.02); ExPEC	Yes

					unknown/ other (0.2); urine (0.07)	North America (0.12); nd (0.05)	106 (0.03)	(0.79); EXPEC (0.01); ND (0.19)	
8	146	5.47	5178.5	B1	unknown/ other (0.79); faeces (0.21)	Europe (0.75); North America (0.25)	442 (0.85); 442~ (0.06); 7714 (0.05)	ND (0.06); STEC (0.94)	No
9	124	5.38	5118	B1	unknown/ other (0.99)	Europe (0.98); North America (0.02)	33 (0.84); 33~ (0.13)	ND (0.02); STEC (0.98)	No
10	104	5.29	4987	E	unknown/ other (0.89); faeces (0.11)	Europe (0.74); North America (0.24)	32 (0.87); 137 (0.08); 32~ (0.04)	aEPEC/ EPEC (0.1); EHEC (0.9)	No
11	106	5.12	4740	B2	blood (0.48); urine (0.34); unknown/ other (0.16)	Europe (0.71); North America (0.27)	127 (0.91); 5337 (0.03)	EAEC (0.01); ExPEC (0.83); EXPEC (0.02); ND (0.14)	No
12	92	4.87	4540	A	unknown/ other (0.7); faeces (0.21); blood (0.05); urine (0.04)	Europe (0.73); Africa (0.11); North America (0.07); South America (0.05); Asia (0.02); nd (0.02)	10 (0.76); 43 (0.08); 4305 (0.05); 5353 (0.03); 10~ (0.02)	aEPEC/ EPEC (0.24); EAEC (0.29); EAEC+STEC (0.02); ETEC (0.01); ExPEC (0.11); ND (0.24); STEC (0.1)	No
13	110	5.16	4801.5	B2	blood (0.65); unknown/ other (0.18); urine (0.15)	Europe (0.78); North America (0.16); nd (0.04)	12 (0.92)	ExPEC (0.81); ND (0.19)	No
14	109	5.30	5090	B1	unknown/ other (0.7); faeces (0.3)	North America (0.72); Europe (0.27)	16 (0.93)	EHEC (0.98); STEC (0.02)	No

15	89	5.46	5148	B1	unknown/ other (0.99)	Europe (0.99)	25 (0.65); 25~ (0.1); 4748 (0.09); 811 (0.09); 6265~ (0.02)	ND (0.33); STEC (0.67)	No
16	62	5.34	4995.5	E	unknown/ other (0.79); faeces (0.21)	Europe (0.87); North America (0.06); Asia (0.05)	335 (0.95); 7444 (0.03)	aEPEC/ EPEC (0.26); EHEC (0.62); STEC (0.11)	Yes
17	79	5.19	4884	B2	blood (0.58); unknown/ other (0.29); urine (0.11)	Europe (0.77); North America (0.18); nd (0.05)	14 (0.29); 404 (0.29); 1193 (0.18); 550 (0.15); 1057 (0.06); 5669 (0.03)	ExPEC (0.7); ND (0.29); STEC (0.01)	Yes
18	55	5.05	4889	U	unknown/ other (1)	Europe (1)	504 (0.53); 5292 (0.24); 6880 (0.13); 504~ (0.07)	STEC (1)	Yes
19	66	5.18	4767.5	D	blood (0.76); unknown/ other (0.12); urine (0.09); faeces (0.03)	Europe (0.68); North America (0.24); nd (0.05); Africa (0.03)	393 (0.8); 31 (0.11); 1394 (0.03)	EAEC (0.12); ExPEC (0.85); ND (0.03)	Yes
20	48	5.00	4657	B2	blood (0.48); unknown/ other (0.27); urine (0.23); faeces (0.02)	Europe (0.75); North America (0.12); nd (0.1); Oceania (0.02)	131 (0.94); 131~ (0.02); 5432 (0.02); 5494 (0.02)	ExPEC (0.71); ND (0.29)	Yes

21	70	5.39	5078	B1	faeces (1)	North America (1)	392 (0.67); 5738 (0.33)	ND (1)	No
22	46	5.21	5017	B1	unknown/ other (0.87); faeces (0.13)	Europe (0.76); North America (0.24)	300 (0.57); 343 (0.2); 4942 (0.09); 300~ (0.07); 5343 (0.04); 343~ (0.02); 6668 (0.02)	aEPEC/ EPEC (0.2); EHEC (0.8)	No
23	42	5.03	4767	A	unknown/ other (0.98); faeces (0.02)	Europe (0.93); nd (0.05); Africa (0.02)	6 (0.86); 6~ (0.12); 8300 (0.02)	EIEC (1)	No
24	45	5.23	4895	B1	unknown/ other (0.84); faeces (0.16)	Europe (0.89); North America (0.07); Africa (0.02); Asia (0.02)	678 (0.84); 678~ (0.16)	EAEC (0.62); EAEC+STEC (0.38)	Yes
25	49	5.09	4728	B2	blood (0.51); unknown/ other (0.43); urine (0.06)	Europe (0.76); North America (0.14); nd (0.1)	141 (0.71); 998 (0.2); 8290 (0.04); 141~ (0.02); 998~ (0.02)	ExPEC (0.58); ND (0.4); STEC (0.02)	No
26	44	5.06	4751.5	B1	faeces (0.66); unknown/ other (0.34)	Europe (0.39); Asia (0.32); Africa (0.3)	517 (0.8); 5241 (0.11); 517~ (0.07); 5485 (0.02)	aEPEC/ EPEC (0.93); EHEC (0.05); EPEC/ ETEC (0.02)	Yes
27	27	4.95	4738	B2	unknown/ other (0.89);	Europe (0.93); Africa (0.04);	583 (0.89); 122 (0.07);	aEPEC/ EPEC (0.33); EHEC (0.67)	No

					faeces (0.11)	North America (0.04)	7703 (0.04)		
28	42	5.19	4852.5	A	unknown/ other (1)	Europe (0.95); North America (0.05)	10 (0.86); 10~ (0.14)	ND (0.1); STEC (0.9)	No
29	40	5.19	4865	B2	blood (0.7); unknown/ other (0.2); faeces (0.05); urine (0.05)	Europe (0.85); North America (0.12); South America (0.02)	144 (0.95); 5346 (0.02); 703 (0.02)	ExPEC (0.75); ND (0.25)	No
30	41	4.28	4231	*S. sonnei*	unknown/ other (0.88); faeces (0.12)	nd (0.98); Europe (0.02)	245 (0.73); 1024 (0.17); 631 (0.05); 1753 (0.02); 5233 (0.02)	EIEC (0.22); ND (0.78)	No
31	36	5.28	4987.5	F	blood (0.67); unknown/ other (0.22); faeces (0.08); urine (0.03)	Europe (0.69); North America (0.28); Oceania (0.03)	62 (0.97); 1810 (0.03)	ExPEC (0.74); EXPEC (0.03); ND (0.24)	No
32	26	5.25	4931.5	B1	unknown/ other (0.88); faeces (0.08); urine (0.04)	Europe (0.88); Asia (0.04); North America (0.04); South America (0.04)	200 (1)	EAEC (0.92); ETEC / EAEC (0.04); ExPEC (0.04)	Yes
33	35	5.19	4948	F	blood (0.49); unknown/ other (0.46); urine (0.06)	Europe (0.89); North America (0.09); South America (0.03)	59 (0.94); 415 (0.03); 415~ (0.03)	ExPEC (0.56); ND (0.41); STEC (0.03)	No

34	34	5.13	4888	B1	faeces (0.88); unknown/ other (0.12)	Africa (0.59); Asia (0.21); Europe (0.15); South America (0.06)	328 (0.79); 328~ (0.18); 5618 (0.03)	aEPEC/ EPEC (0.85); EPEC/ ETEC (0.15)	Yes
35	22	4.97	4653	A	unknown/ other (0.95); faeces (0.05)	Europe (0.95); South America (0.05)	34 (0.91); 34~ (0.05); 8053 (0.05)	EAEC (1)	Yes
36	24	5.34	4987	B1	unknown/ other (1)	Europe (1)	675 (0.79); 675~ (0.12); 180~ (0.04); 7953 (0.04)	STEC (1)	No
37	27	5.33	4960	D	blood (0.74); unknown/ other (0.15); urine (0.07); faeces (0.04)	Europe (0.63); North America (0.33); Asia (0.04)	405 (0.96); 964 (0.04)	ETEC (0.04); ExPEC (0.85); ND (0.08); STEC (0.04)	Yes
38	27	5.26	4973	F	unknown/ other (0.81); blood (0.07); urine (0.07); faeces (0.04)	Europe (0.52); nd (0.41); North America (0.07)	59 (0.93); 2618 (0.04); 59~ (0.04)	ExPEC (0.16); ND (0.84)	No
39	29	5.26	4994	B1	unknown/ other (0.76); faeces (0.24)	North America (0.59); Europe (0.31); Asia (0.07); South America (0.03)	655 (1)	aEPEC/ EPEC (0.03); EHEC (0.97)	No
40	28	4.85	4551	C	blood (0.57); faeces (0.21); unknown/	Europe (0.64); North America (0.25); Asia	23 (0.39); 410 (0.32); 2230	ETEC (0.18); ExPEC (0.57); ND (0.04); STEC (0.21)	Yes

					other (0.21)	(0.07); South America (0.04)	(0.07); 369 (0.07); 5491 (0.07); 23~ (0.04); 5286 (0.04)		
41	26	5.20	4825	B2	blood (0.73); urine (0.15); unknown/ other (0.12)	Europe (0.96); North America (0.04)	80 (0.88); 5351 (0.04); 5384 (0.04); 5609 (0.04)	ExPEC (0.88); ND (0.12)	No
42	27	4.89	4564	B1	faeces (0.81); unknown/ other (0.15); blood (0.04)	North America (0.85); Europe (0.15)	297 (0.96); 297~ (0.04)	ExPEC (0.04); ND (0.96)	No
43	27	4.63	4274	B1	faeces (1)	North America (1)	906 (1)	ND (1)	No
44	25	5.37	4995	F	blood (0.48); unknown/ other (0.28); urine (0.24)	Europe (0.4); nd (0.4); North America (0.2)	648 (1)	ExPEC (0.72); ND (0.28)	Yes
45	26	4.69	4532.5	*S. flexneri*	unknown/ other (0.96); faeces (0.04)	nd (0.85); Europe (0.15)	152 (0.96); 1502 (0.04)	EIEC (0.35); ND (0.65)	No
46	26	5.40	5040	D	blood (0.5); unknown/ other (0.27); faeces (0.19); urine (0.04)	Europe (0.69); nd (0.12); South America (0.12); North America (0.08)	405 (0.88); 38~ (0.04); 402~ (0.04); 5377 (0.04)	ExPEC (0.54); ND (0.46)	No
47	23	4.94	4614	B2	blood (0.78); unknown/ other (0.22)	Europe (0.96); nd (0.04)	357 (1)	ExPEC (0.78); ND (0.22)	No

48	21	5.32	5180	B1	faeces (0.81); unknown/ other (0.19)	Africa (0.38); Asia (0.29); Europe (0.19); nd (0.05); North America (0.05); South America (0.05)	3 (0.9); 3~ (0.05); 5326 (0.05)	aEPEC/ EPEC (1)	Yes
49	23	5.45	5213	B1	faeces (1)	North America (1)	154 (1)	ND (1)	No
51	22	4.95	4575	D	faeces (1)	North America (0.91); South America (0.09)	501 (1)	EAEC (0.05); ND (0.95)	No