Elucidating the Unknown Ecology of Bacterial Pathogens from Genomic Data

Tristan Kishan Seecharran

A thesis submitted in partial fulfilment of the requirements of

Nottingham Trent University for the degree of

Doctor of Philosophy

June 2018

Copyright Statement

Acknowledgements

I would like to express my sincere gratitude and thanks to my external advisor Alan McNally and director of studies Ben Dickins for their continued support, guidance and encouragement, and without whom, the completion of this thesis would not have been possible. Many thanks also go to the members of the Pathogen Research Group at Nottingham Trent University. I would like to thank Gina Manning and Jody Winter in particular for their invaluable advice and contributions during lab meetings. I would also like to thank our collaborators, Mikael Skurnik and colleagues from the University of Helsinki and Jukka Corander from the University of Oslo, for their much-appreciated support and assistance in this project and the published work on *Yersinia pseudotuberculosis*. I also express my gratitude to the Vice Chancellor of Nottingham Trent University for funding the duration of my PhD study.

Special thanks go to all my friends and colleagues in the microbiology laboratory at Nottingham Trent University for being a source of motivation and encouragement and providing an enjoyable atmosphere, in which I had the pleasure of carrying out my PhD research. This thesis is a dedication to my loving parents Royston and Seeta, to whom I am eternally grateful for their unconditional love, endurance, and encouragement. My wonderful sister Trisha, grandparents, aunts, and uncles also deserve my heartfelt and utmost thanks and appreciation for their unequivocal and unwavering support along this journey.

Abstract

Our knowledge and understanding of how bacterial pathogens have evolved has been limited by inadequate information on the full ecology of these organisms. Large-scale population genomic analyses have enabled a high-resolution view of variation at the core and accessory genome level, within and between bacterial populations, revealing previously hidden patterns of variation among microorganisms. This sheds light on the evolution and maintenance of ecologically distinct populations of bacteria and has raised the question of whether the same approach can uncover novel information on the ecology of established, clinically important pathogens. *Yersinia pseudotuberculosis* and *Escherichia coli* represent 'model' organisms in the study of microbial evolution, but given the high degree of niche overlap in both species, their ecology is largely unknown.

In this study, genomic analyses of *Y. pseudotuberculosis* strains, obtained from various habitats worldwide, revealed a phylogeographic split within the population, with an Asian ancestry and subsequent dispersal of successful clonal lineages across the rest of the world. These lineages were differentiated by CRISPR arrays and we demonstrated that genetic exchange between lineages is restricted. Despite the coexistence of these lineages for thousands of years, the discrete lineage structure of the population is maintained due to the restriction of inter-lineage genetic exchange. The analyses did not identify a role for ecological barriers in defining the distinct lineage structure of the species, suggesting that *Y. pseudotuberculosis* is a host generalist able to succeed in multiple habitats.

The relative abundance of multidrug-resistant extraintestinal pathogenic *E. coli* (ExPEC) among *E. coli* inhabiting non-human niches is undetermined, owing to many studies selectively isolating resistant bacteria. To compare the population structure of *E. coli* from non-human environments with the well-defined population structure of human-clinical *E. coli*, unbiased sampling of *E. coli* isolates from river water and retail poultry samples was undertaken. Genomic analysis of isolates revealed a low prevalence of clinically-associated sequence types and potential ExPEC strains among non-human *E. coli* when contrasted with human-clinical *E. coli*, suggesting two distinct populations. Comparative genomic analyses further supported this, revealing a noticeable difference in accessory genome content between the two populations and low levels of genetic exchange between closely related strains. This suggests ecological barriers, resulting in gradual genetic isolation, may have contributed to the divergence of these niche-associated populations of *E. coli*. The investigation concluded that the non-human population of *E. coli* is unlikely to contribute significantly to the weight of hospital- and community-acquired extraintestinal infections in humans.

Table of Contents

Copyright Statement ..... **i**

Acknowledgements ..... **ii**

Abstract ..... **iii**

Table of Contents ..... **iv**

List of Tables ..... **ix**

List of Figures ..... **x**

List of Abbreviations ..... **xiii**

List of Publications ..... **xvi**

1. Introduction **1**

1.1. Unravelling microbial ecology 2

1.1.1. 16S ribosomal RNA sequencing 3

1.1.2. Metagenomics 3

1.1.3. Combining a culture-based approach with whole-genome sequencing 4

1.1.4. Advancements in whole-genome sequencing technologies 5

1.1.5. Using population genomics to uncover patterns of microbial ecology 6

1.2. *Yersinia pseudotuberculosis* 9

1.2.1. The pathogenic *Yersiniae* 9

1.2.2. Hidden ecological patterns in the pathogenic *Yersiniae* 10

1.2.3. An introduction to the *Y. pseudotuberculosis* species 11

1.2.4. Pathogenesis of *Y. pseudotuberculosis* 12

1.2.4.1. Transmission of the bacteria to humans 12

1.2.4.2. Epidemiology of *Y. pseudotuberculosis*-associated yersiniosis 13

1.2.4.3. Virulence factors of *Y. pseudotuberculosis* 14

1.2.4.4. Pathophysiology of *Y. pseudotuberculosis* infection 15

1.2.5. Clinical manifestations of *Y. pseudotuberculosis* infection 18

1.2.6. Treatment and prevention 18

1.2.7. Identification and typing of *Y. pseudotuberculosis* 19

1.2.7.1. Serological characterisation of *Y. pseudotuberculosis* 19
1.2.7.2. *Y. pseudotuberculosis* multilocus sequence typing (MLST) 20
1.3. *Escherichia coli* 21
1.3.1. An introduction to the *E. coli* species 21
1.3.2. Commensal *E. coli* 22
1.3.3. Intestinal pathogenic *E. coli* (IPEC) 23
1.3.4. Extraintestinal pathogenic *E. coli* (ExPEC) 25
1.3.4.1. Avian pathogenic *E. coli* (APEC) 26
1.3.4.2. Neonatal meningitis *E. coli* (NMEC) 26
1.3.4.3. Uropathogenic *E. coli* (UPEC) 27
1.3.5. Antibiotic resistance in extraintestinal pathogenic *E. coli* 30
1.3.6. Source attribution 33
1.3.6.1. Surface waters as an environmental reservoir for ExPEC 35
1.3.6.2. Retail poultry meat as a reservoir for ExPEC 36
1.3.7. Population structure of extraintestinal pathogenic *E. coli* 37
1.3.7.1. *E. coli* phylo-typing 37
1.3.7.2. ExPEC genotyping 38
1.4. Aims and objectives 41
2. Materials and methods 43
2.1. Media and reagents 44
2.1.1. Growth and storage media 44
2.1.2. API identification kit 44
2.1.3. Molecular microbiology reagents 45
2.1.4. NGS reagents and sequencing kits 45
2.1.5. Buffers and reagents 46
2.2. *Y. pseudotuberculosis* phylogenomics 46
2.2.1. Determining phylogenetic relationships 46
2.2.2. Analysis of CRISPR loci 47
2.2.3. Pan-genome and accessory genome analyses 47

2.2.4. Detection of core genome recombination events 48
2.2.5. Dating analysis 48
2.3. Non-human *E. coli* strain collection 49
2.3.1 River water sampling 49
2.3.2. Retail poultry sampling 52
2.3.3. Isolation and identification of *E. coli* from non-human samples 53
2.4. Molecular characterisation of non-human *E. coli* 56
2.4.1. Preparation of genomic DNA 56
2.4.2. Detection of β-lactamase genes 56
2.4.3. Agarose gel electrophoresis 57
2.5. Whole-genome sequencing 59
2.5.1. Quality assessment of genomic DNA 59
2.5.2. Illumina Nextera XT library preparation 59
2.5.3. Sequencing on the MiSeq 61
2.6. Analysis of *E. coli* sequence data 64
2.6.1. Genome assembly, annotation, and quality assessment 64
2.6.2. Sequence-typing and clonal complex assignment 64
2.6.3. Detecting antibiotic resistance genes and ExPEC virulence determinants 64
2.6.4. Reconstructing phylogenetic trees 65
2.6.5. Pan-genome analysis 65
2.6.6. Detection of recombinant genomic regions 66
2.6.7. Statistical analyses 66
3. Ecology and evolution of a global population of *Yersinia pseudotuberculosis* 67
3.1. Introduction 68
3.1.1. Aim and objectives 70
3.2. Materials and Methods 71
3.3. Results and Discussion 78
3.3.1. Phylogeographic structure of *Y. pseudotuberculosis* indicates an Asian ancestry for the species 78

3.3.2. Phylogenetic clusters within *Y. pseudotuberculosis* associate with discrete CRISPR cassette patterns 80

3.3.3. Cryptic ecology suggests that *Y. pseudotuberculosis* is a host generalist 85

3.3.4. Phylogenetic dating suggests recent geographical divergence 88

3.3.5. CRISPR-associated phylogenetic clusters are associated with patterns of accessory gene conservation and core genome recombination 90

3.4. Conclusions 96

4. Defining the population structure of *Escherichia coli* from non-human sources 100

4.1. Introduction 101

4.1.1. Aim and objectives 103

4.2. Materials and Methods 105

4.3. Results and Discussion 106

4.3.1. Prevalence of *E. coli* isolated from river water and retail chicken meat 106

4.3.2. Molecular detection of β-lactamase genes in non-human *E. coli* 109

4.3.3. Whole-genome sequencing of non-human *E. coli* isolates 112

4.3.4. Whole-genome-based multilocus sequence typing (MLST) analysis of the non-human population of *E. coli* 114

4.3.5. Defining the phylogeny of the non-human population of *E. coli* 120

4.3.6. Distribution of antimicrobial resistance genes among non-human *E. coli* 124

4.3.7. Determining the prevalence of ExPEC strains in the non-human population of *E. coli* 130

4.4. Conclusions 133

5. Comparative population genomics of *Escherichia coli* from human-clinical and non-human sources 136

5.1. Introduction 137

5.1.1. Objectives 139

5.2. Materials and Methods 141

5.3. Results and Discussion 153

5.3.1. Comparing the prevalence of STs between the human-clinical and non-human populations of *E. coli* 153

5.3.2. Comparative phylogenomic analysis of the human-clinical and non-human populations of *E. coli* 160

5.3.3. Determining the prevalence of ExPEC strains in the human-clinical population of *E. coli* 163

5.3.4. Distribution of antimicrobial resistance genes among human-clinical *E. coli* .. 165

5.3.5. Phylogenetic analyses of the *E. coli* lineages ST131 and ST648..... 168

5.3.5.1. *E. coli* ST131..... 168

5.3.5.2. *E. coli* ST648..... 170

5.3.6. Comparative genomic analyses 173

5.3.6.1. Pan-genome approach to compare human-clinical and non-human *E. coli*... 173

5.3.6.2. Comparisons of the accessory genomes of human-clinical and non-human *E. coli* 175

5.3.7. Pan-genome and recombination analysis of *E. coli* lineages ST69 and ST10.... 184

5.3.7.1. *E. coli* ST69 184

5.3.7.2. *E. coli* ST10 188

5.4. Conclusions 192

6. Conclusions and future directions 198

References 207

Appendix 240

List of Tables

Table **Page**

Table 2.1: Numbers of whole chickens obtained from 6 major supermarket chains and a snapshot of chicken processing companies in the UK..... 52

Table 2.2: Identification characteristics used to presumptively identify *E. coli* from other bacterial species present in environmental samples..... 55

Table 2.3: Primer sequences used in this study..... 58

Table 2.4: Illumina adapter sequences – index 1 (i7) adapters..... 62

Table 2.5: Illumina adapter sequences – index 2 (i5) adapters..... 63

Table 3.1: *Yersinia pseudotuberculosis* genomes used in this study..... 72

Table 3.2: CRISPR cluster-type *Y. pseudotuberculosis* strains with known isolation dates..... 84

Table 4.1: *E. coli* reference genomes used for phylogrouping of non-human *E. coli* strains..... 105

Table 4.2: Total number of *E. coli* isolates identified from river water samples..... 107

Table 4.3: Total number of *E. coli* isolates identified from retail chicken samples..... 108

Table 4.4: Sequence type designations for the 128 sequenced non-human *E. coli* genomes..... 116

Table 5.1: List of sequenced human-clinical *E. coli* genomes used for comparative genomic analyses in this study..... 141

Table 5.2: List of sequenced ST131 *E. coli* genomes used for comparative phylogenetic analysis in this study..... 145

Table 5.3: List of sequenced ST648 *E. coli* genomes used for comparative phylogenetic analysis in this study..... 151

Table 5.4: Quantified recombinations between non-human and human-clinical genomes..... 191

List of Figures

Figure **Page**

Figure 1.1: Routes of transmission and mechanisms of pathogenesis of the human pathogenic *Yersinia* species..... 17

Figure 1.2: Sites of colonisation by pathogenic *E. coli*..... 29

Figure 1.3: Potential ecological habitat and routes of transmission of *E. coli* in a global ecosystem..... 34

Figure 2.1: Map of Giltbrook sample site..... 50

Figure 2.2: Map of Erewash Pinxton sample site..... 50

Figure 2.3: Map of East Leake sample site..... 51

Figure 2.4: Map of Keyworth sample site..... 51

Figure 3.1: Maximum-likelihood phylogenetic tree of 134 *Y. pseudotuberculosis* isolates annotated with continents of origin..... 79

Figure 3.2: Maximum-likelihood phylogenetic tree of 134 *Y. pseudotuberculosis* isolates annotated with the 33 identified CRISPR clusters..... 82

Figure 3.3: Global map showing sources of isolation of strains belonging to each of the 33 identified CRISPR clusters..... 83

Figure 3.4: Maximum-likelihood phylogenetic tree of 134 *Y. pseudotuberculosis* isolates annotated with all available metadata..... 87

Figure 3.5A: Dated maximum clade credibility tree of 46 *Y. pseudotuberculosis* strains for which isolation dates were available..... 89

Figure 3.5B: Bayesian Skyline reconstruction from BEAST 2 analysis..... 89

Figure 3.6: Distribution of accessory gene profiles for 134 *Y. pseudotuberculosis* isolates..... 94

Figure 3.7: BratNextGen analysis of core genome recombination events for 134 *Y. pseudotuberculosis* isolates..... 95

Figure 4.1: Electrophoresis gel showing PCR amplicons of β-lactamase genes detected in *E. coli* isolated from non-human samples..... 110

Figure 4.2A: Percentage prevalence of β-lactamase genes bla_{TEM}, bla_{SHV}, $bla_{CTX\text{-}M}$, and bla_{OXA} in *E. coli* isolates from human-clinical samples..... 111

Figure 4.2B: Percentage prevalence of β-lactamase genes *bla*$_{TEM}$, *bla*$_{SHV}$, *bla*$_{CTX-M}$, and *bla*$_{OXA}$ in *E. coli* isolates from non-human samples..... 111

Figure 4.3: Workflow indicating the numbers of *E. coli* isolates from river water and retail chicken samples consolidated at each stage of the investigation..... 113

Figure 4.4: Minimum spanning tree illustrating STs of the non-human *E. coli* population isolated from river water and retail chicken samples..... 119

Figure 4.5: Maximum-likelihood phylogenetic tree of 128 *E. coli* strains isolated from river water and retail chicken samples in Nottingham and 18 reference strains..... 123

Figure 4.6: Distribution of antibiotic resistance gene profiles across 128 *E. coli* strains isolated from river water and retail chicken samples in Nottingham..... 129

Figure 4.7: Distribution of ExPEC VAGs among 128 *E. coli* strains isolated from river water and retail chicken samples in Nottingham..... 132

Figure 5.1A: Prevalence of *E. coli* STs among *E. coli* strains isolated from human-clinical samples collected in Nottingham..... 156

Figure 5.1B: Prevalence of *E. coli* STs among *E. coli* strains isolated from non-human samples collected in Nottingham..... 156

Figure 5.2: Prevalence of the most predominant *E. coli* STs associated with human extraintestinal infection in the human-clinical and non-human populations of *E. coli* in Nottingham..... 157

Figure 5.3: Minimum spanning tree illustrating the *E. coli* STs isolated from human-clinical and non-human samples..... 159

Figure 5.4: Maximum-likelihood SNP-based phylogenetic tree of 264 *E. coli* strains isolated from human-clinical and non-human samples in Nottingham..... 162

Figure 5.5: Distribution of ExPEC VAGs among 136 *E. coli* strains isolated from human-clinical samples in Nottingham..... 164

Figure 5.6: Distribution of antimicrobial resistance genes among 136 *E. coli* strains isolated from human-clinical samples in Nottingham..... 167

Figure 5.7: Maximum likelihood SNP-based phylogenetic tree of *E. coli* ST131 isolates..... 169

Figure 5.8: Maximum likelihood SNP-based phylogenetic tree of *E. coli* ST648 isolates..... 172

Figure 5.9: Core and pan-genome frequency plots of all 264 *E. coli* strains isolated from human-clinical and non-human samples in Nottingham..... 174

Figure 5.10: Functional categories of genes present in ≥ 85% of all 264 *E. coli* strains isolated from human-clinical and non-human samples in Nottingham..... 176

Figure 5.11: Comparison of population-unique accessory genes and shared accessory genes between the human-clinical and non-human populations of *E. coli* in Nottingham..... 178

Figure 5.12: Frequency of gene occurrence plots for population-unique and shared accessory genes for the human-clinical and non-human populations of *E. coli* in Nottingham..... 179

Figure 5.13: Distribution of accessory genes in the human-clinical and non-human populations of *E. coli* in Nottingham..... 183

Figure 5.14: Distribution of gene profiles for ST69 *E. coli* strains isolated from human-clinical and non-human samples in Nottingham..... 187

Figure 5.15: Distribution of core genome recombination events for ST69 *E. coli* strains isolated from human-clinical and non-human samples in Nottingham..... 187

Figure 5.16: Distribution of gene profiles for ST10 *E. coli* strains isolated from human-clinical and non-human samples in Nottingham..... 190

Figure 5.17: Distribution of core genome recombination events for ST10 *E. coli* strains isolated from human-clinical and non-human samples in Nottingham..... 190

List of Abbreviations

ABU	Asymptomatic bacteriuria
A/E	Attaching and effacing
APEC	Avian pathogenic *Escherichia coli*
API	Analytical profile index
ATM	Amplicon Tagment Mix
BBB	Blood-brain barrier
BEAST	Bayesian Evolutionary Analysis by Sampling Trees
BLAST	Basic Local Alignment Search Tool
bp	Base pair
BTI	Biliary tract infection
BSAC	British Society for Antimicrobial Chemotherapy
Cas	CRISPR-associated system (proteins)
CDS	Coding sequence
CLED	Cystine-, lactose-, and electrolyte-deficient
CNS	Central nervous system
CRISPR	Clustered regularly interspaced short palindromic repeat
crRNA	CRISPR ribonucleic acid
CSV	Comma-separated values (file)
CTX-M	Cefotaxime preferential extended-spectrum β-lactamase
DA	Domesticated animal
DAEC	Diffusely adherent *Escherichia coli*
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
DOI	Digital object identifier
DR	Direct repeats
EAEC	Enteroaggregative *Escherichia coli*
EAHEC	Entero-aggregative-haemorrhagic *Escherichia coli*
ECOR	*Escherichia coli* Reference (collection)
EHEC	Enterohaemorrhagic *Escherichia coli*
EIEC	Enteroinvasive *Escherichia coli*
EPEC	Enteropathogenic *Escherichia coli*
ESBL	Extended-spectrum beta-lactamase
ESS	Effective sample size
ETEC	Enterotoxigenic *Escherichia coli*

ExPEC	Extraintestinal pathogenic *Escherichia coli*
FIMM	Institute for Molecular Medicine Finland
GC	Guanine and cytosine
gDNA	Genomic deoxyribonucleic acid
GFF	General feature format (file)
HC	Haemorrhagic colitis
HCl	Hydrochloric acid
HPI	High-pathogenicity island
HT1	Hybridization Buffer
HUS	Haemolytic uraemic syndrome
IPEC	Intestinal pathogenic *Escherichia coli*
iTOL	Interactive Tree of Life
Kbp	Kilobase pairs
LPS	lipopolysaccharide
LS-BSR	Large-Scale Blast Score Ratio
Mbp	Megabase pairs
MCC	Maximum clade credibility (tree)
MDR	Multidrug resistance/multidrug-resistant
MLEE	Multilocus enzyme electrophoresis
MLS	Macrolide-lincosamide-streptogramins
MLST	Multilocus sequence typing
MST	Minimum spanning tree
MSU	Midstream sample of urine
NaOH	Sodium hydroxide
NCBI	National Centre for Biotechnology Information
NGS	Next-generation sequencing
NMEC	Neonatal meningitis *Escherichia coli*
NPM	Nextera PCR Master Mix
NT	Neutralize Tagment (buffer)
OD	Optical density
OH	Hydroxyl (group)
ONT	Oxford Nanopore Technologies
OPS	O-specific polysaccharide (O-antigen)
ORF	Open reading frame
OXA	Oxacillin preferential β-lactamase

PAI	Pathogenicity island
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PFGE	Pulsed field gel electrophoresis
PG	Phylogroup
pH	Potential hydrogen
PSA	Proportion of shared ancestry (tree)
QUAST	Quality Assessment Tool
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
rpm	Revolutions per minute
RTI	Respiratory tract infection
SBS	Sequencing by synthesis
SHV	Sulfhydryl-variable β-lactamase
SNP	Single nucleotide polymorphism
ST	Sequence type
STC	Sequence type complex
STEC	Shiga toxin-producing *Escherichia coli*
Stx	Shiga toxin
SWI	Surgical wound infection
T3SS	Type III secretion system
TAE	Tris-acetate-EDTA (buffer)
TD	Tagment DNA (buffer)
TEM	Temoneira β-lactamase
TMRCA	Time to most recent common ancestor
UPEC	Uropathogenic *Escherichia coli*
UTI	Urinary tract infection
UV	Ultraviolet
VAG	Virulence-associated gene
VFDB	Virulence Factors Database
WGS	Whole-genome sequencing
YAPI	*Yersinia* adhesion pathogenicity island

List of Publications

Conference presentations

Seecharran T, Dickins B, Skurnik M, Corander J, McNally A. Phylogeographic separation and formation of sexually discrete lineages in a global population of *Yersinia pseudotuberculosis*. 5th Molecular Microbiology and 4th M4 Meeting 2017, Birmingham (poster).

Seecharran T, Dickins B, Skurnik M, Corander J, McNally A. Phylogenetic and CRISPR analysis reveals ecological separation and restricted gene flow in a global population of *Yersinia pseudotuberculosis*. Microbiology Society Annual Conference 2017, Edinburgh (poster).

Seecharran T, Winter J, Skurnik M, Corander J, McNally A. Elucidating the unknown ecology of bacterial pathogens from phylogenetic data. STAR Conference 2016, NTU (poster/oral presentation). *Awarded first prize for best oral presentation*.

Seecharran T, Winter J, McNally A. Defining the population structure of *Escherichia coli* in the environment. Microbiology Society Annual Conference 2016, Liverpool (poster).

Journal articles

Seecharran T, Kalin-Manttari L, Koskela K, Nikkari S, Dickins B, Corander J, Skurnik M, McNally A. (2017) Phylogeographic separation and formation of sexually discrete lineages in a global population of *Yersinia pseudotuberculosis. Microbial Genomics*, 3 (10), e000133. http://doi.org/10.1099/mgen.0.000133.

McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, **Seecharran T** *et al*., (2016) Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations. *PLoS Genet* 12 (9), e1006280. https://doi.org/10.1371/journal.pgen.1006280.

CHAPTER 1

Introduction

1. Introduction

1.1. Unravelling microbial ecology

Microbial ecology examines the diversity of microorganisms and how they maintain diversity through the interaction with each other, and with their environment. Microbial ecological studies have traditionally concentrated on two areas: (i) microbial diversity, which encompasses the isolation, identification, and quantification of different microorganisms in various habitats; and (ii) microbial activity, which pertains to what microorganisms do in their habitats and how this contributes to the observed microbial diversity (Xu, 2006). The study of microorganisms began in the late 17th century, through the work of Robert Hooke and Antonie van Leeuwenhoek, who individually published the first original observations of single-celled organisms (Gest, 2004). The diversity of microorganisms in the environment can be measured at various levels such as phylogenetic diversity, species diversity, genotypic diversity, and gene diversity (Xu, 2006). Microbial populations were traditionally understood at the species level and above solely through culture-based techniques, of which less than 1% of microbes are culturable (Amann *et al*., 1990), and typically require numerous physiological and biochemical tests for full characterisation. Classical microbiology primarily consisted of isolating microbes, growing them as pure cultures, and identifying biochemical properties of these organisms, such as cell wall structure by Gram staining, oxygen tolerance, and carbon or nitrogen sources that supported their growth. These techniques can be very time-consuming, laborious, and require prior knowledge of the organisms of interest to selectively and successfully culture from a complex microbial community. There are several limitations to culture-dependent methods, most important of which is a bias towards readily cultivable organisms, thus ignoring non-culturable bacteria, which comprise the largest proportion of microbes (Xu, 2006). Furthermore, even among culturable microorganisms, the true diversity found in nature may not be accurately represented by the observed diversity on standard microbiological media. This is because even though many different variants of media and growth conditions have been developed over the years to culture microorganisms, it is impossible to explore all of the conditions required to culture every microorganism. A more accurate estimation of microbial diversity has been achieved over the past two decades due to the application of culture-independent genomics tools. Immunological methods, such as enzyme-linked immunosorbent assay (ELISA) have been used to characterise and identify bacteria from a variety of ecosystems. However, these techniques are still limited in assessing functionality and have only been used sparingly due to the need for specific antigens/antibodies. Consequently, the focus of many recent microbial

ecological studies has turned towards development of molecular approaches for understanding these complex communities which, in some cases, eliminates the need for culturing beforehand.

1.1.1. 16S ribosomal RNA sequencing

Molecular techniques have superseded traditional microbiological methods due to the ease of use, reproducibility, sensitivity and speed of working with nucleic acids that these techniques provide. Culture-independent approaches such as DNA cloning and polymerase chain reaction (PCR) have been used to detect specific members as well as functional genes within a microbial community. Functional genes such as the 16S ribosomal RNA (rRNA) gene and 5S rRNA gene have been used as molecular markers for studying phylogenetic relationships. The rRNA gene consists of different regions, some of which are highly conserved across all phylogenetic domains of life (*i.e.* bacteria, eukarya, and archaea) and are more resistant to mutations, while other regions, called 'hot spots', are variable between closely related bacterial species. This variability allows for inferring phylogenetic information from microorganisms inhabiting different ecosystems (Clarridge, 2004). Another molecular marker would include the *recA* gene, which is essential for the repair and maintenance of DNA in all species, and has been used to define the phylogeny of *Vibrio cholerae* (Stine *et al*., 2000). In addition, the *rpoA* gene encodes the alpha subunit of RNA polymerase and has also proved useful in determining phylogenetic relationships (Fox and Sorhannus, 2003). A new era began with the innovation of Next-Generation sequencing (NGS) technologies, which transformed the existing gold standard techniques of microbial community analysis. NGS allows large sets of sequence data to be generated in parallel, utilising phylogenetic markers such as the 16S rRNA gene, inexpensively and in considerably less time. The high-throughput (amount of DNA processed per unit time) results generated by 16S gene sequencing enables comprehensive analysis of community microbiota from various environments, as well as providing insights into the interactions with their ecosystems. In general, differentiation between organisms at the genus level across all major phyla of bacteria can be achieved through the comparison of 16S rRNA sequence data. Although 16S rRNA gene sequencing is highly useful with regards to bacterial classification, this technique does carry several limitations and sometimes exhibits low phylogenetic resolution at the species level, due to sequence similarities between species that exhibit different phenotypes, and in some cases, offers poor discriminatory power for some genera.

1.1.2. Metagenomics

A possible solution to these taxonomic problems is to utilise metagenomics, which is another high-throughput sequencing technology used to reveal microbial diversity in natural

environments. Metagenomics is the culture-independent analysis of the collective microbial genomes (termed the metagenome) in an environmental community, using an approach based either on expression or sequencing (Riesenfeld, Schloss and Handelsman, 2004). Such microbial communities may include a soil or water sample that contains substantially more genetic information than is available from the cultured subset. Metagenomic studies usually require the cloning of DNA fragments isolated directly from microbial samples, followed by sequencing and functional analysis of the cloned fragments, ultimately to make taxonomic assignments to characterise the microorganisms. Fully sequencing the DNA within a sample allows investigators to study the taxonomic diversity of all microbes within a sample and not just the culturable. However, one of the main issues presented by metagenomics is that functional genes are not usually analysed. Metagenomics is often limited to taxonomy – analysis of the 16S rRNA gene – which is used to identify organisms with the aim of determining the species present in an environmental sample. A specific function often cannot be reliably inferred from identification of the microorganism. Shotgun metagenomics studies are now starting to present opportunities for unravelling microbial community diversity, such as species sub-typing and strain-level profiling; however, the genomic resolution of single isolate sequencing is still higher than what can be achieved for single organisms in a metagenomics context (Quince *et al*., 2017). Increasing the profiling resolution to the level of single strains would be vital for in-depth population genomic analyses and the study of microbial ecology and epidemiology – something which currently can only be achieved at a high resolution when using a whole-genome sequencing-based approach.

1.1.3. Combining a culture-based approach with whole-genome sequencing

Before the development of molecular techniques, microbial ecological studies were largely reliant upon traditional methods of isolation and identification to probe the structures of microbial populations. Culture-based approaches have not disappeared and indeed are still used, together with molecular analyses, to fully understand bacterial population dynamics. Advances in DNA sequencing technology, starting with the widespread implementation of automated DNA sequencing techniques in the 1990s, have revolutionised our understanding of microbial processes, from the physiology of single cells to large-scale population biology. The last two decades have seen an increase in the use of high-throughput or next-generation sequencing technologies, enabling investigations of microbial communities with unprecedented resolution, underpinning important research in pathogen epidemiology and evolution. To comprehensively characterise the complexity of a population of a particular bacterial species to the strain-level of classification, culture-based techniques must be complemented by the vast

array of information obtained from culture-independent techniques, such as NGS technology. Whole-genome sequencing (WGS) provides more resolution through the generation of genome-wide information for a cultured microbial population. WGS has emerged as the gold standard in bacterial typing, enabling successful tracking of worldwide epidemics, regional outbreaks, and foodborne outbreaks, and demonstrating that the fine-scale resolution provided by WGS facilitates our understanding of the structure of microbial populations and the spread of infectious agents. There are several WGS platforms available that have revolutionised the field of microbial ecology, ranging from traditional Sanger sequencing biochemistry to second-generation (Illumina) and third-generation (single-molecule) sequencing technologies (Schadt, Turner and Kasarskis, 2010).

1.1.4. Advancements in whole-genome sequencing technologies

The first generation of whole-genome sequencing was initially developed by Sanger and colleagues (1975) and in parallel by Maxam and Gilbert (1977), which were termed the 'chain-termination method' and the 'chemical sequencing method', respectively. Of the two methods, Sanger sequencing ultimately prevailed given that it was less complex and more favourable to being scaled up than the chemical sequencing method (Schadt, Turner and Kasarskis, 2010). Sanger sequencing typically results in a read length of ~800 bases, which may be extended to above 1,000 bases (Schadt, Turner and Kasarskis, 2010). While fully automated implementations of this approach were the mainstay for original DNA sequencing technology, their main limitation was the low throughput results, as well as high cost, which resulted in a fundamental shift in methodology, leading to second-generation sequencing technologies, also generally known as next-generation sequencing (NGS) or high-throughput sequencing technologies.

Second-generation sequencing technologies are known for extremely high throughput, resulting in an overall low cost per identified base. However, the time-to-result for these methods is generally long and can take up to several days to complete, due to the requirement of many scanning and washing cycles (Schadt, Turner and Kasarskis, 2010). Illumina NGS workflows include four basic steps: library preparation, cluster generation, sequencing, and data analysis. A sequencing library is prepared by random fragmentation of the DNA sample into short fragments, followed by ligation of 5' and 3' adaptor sequences to the ends of each fragment. For cluster generation, the library is loaded into a flow cell where the fragments attach to a lawn of surface-bound oligonucleotides complementary to the library adapter sequences. Each fragment is then amplified, using a bridge amplification technique, where copies are generated *in situ*, resulting in distinct clonal clusters of ~1,000 amplicons of each fragment. Sequencing of the templates is carried out using Illumina's sequencing-by-synthesis (SBS) technology, where

all four nucleotides are added to the flow cell simultaneously, along with DNA polymerase for the incorporation of bases into the DNA template strands. Each nucleotide is fluorescently labelled with a different colour, and each incorporation is a unique event due to the chemical blockage of the 3'-OH group after incorporation (Illumina, 2017). The flow cell is imaged after each incorporation step and the emission from each cluster is recorded. Each base is identified using the emission wavelength and intensity (Illumina, 2017). This cycle is repeated a specific number of times to create a read length of *n* number of bases. Integrated software checks the quality of each read and removes any poor-quality reads from the genome construct. Bioinformatics software can then align the newly identified sequence reads to a reference genome. Following alignment, many different forms of analysis are possible, such as identification of single nucleotide polymorphisms (SNPs) or insertions and deletions, read counting for RNA methods, phylogenetic analysis, and more. One of the limitations of second-generation sequencing is the generation of short reads, which leads to highly-fragmented assemblies (Schatz, Delcher and Salzberg, 2010).

The demand for technologies that can operate at higher speeds and produce longer reads resulted in the third-generation of sequencing. These sequencing technologies directly target single DNA molecules, enabling real-time sequencing where reads are available for analysis as and when they have passed through the sequencer (Schadt, Turner and Kasarskis, 2010). In 2014, Oxford Nanopore Technologies (ONT) released a new third-generation sequencing platform, known as the MinION, which is a portable, real-time sequencing device. ONT platforms are capable of producing incredibly long reads, with a maximum read length up to a few hundred thousand base pairs (Laver *et al*., 2015). On the other hand, ONT reads have high error rates, with accuracy ranging from 65%−88% (Laver *et al*., 2015). However, due to its small size and low-costing equipment, the MinION sequencer has attracted considerable interest in the genomics community, particularly for pathogen surveillance and clinical diagnostic applications, as these areas can best exploit the real-time nature of this sequencing platform.

1.1.5. Using population genomics to uncover patterns of microbial ecology

With the development and application of genomic tools, microbial ecology is undergoing a resurgence. Genomic tools, which now offer high-quality reads at decreasing costs and high throughput, have given us unprecedented access to microbial diversity. Whole-genome sequencing is increasingly being utilised as the gold standard approach to study the population structure of closely related microbes in place of more traditional molecular techniques such as pulsed-field gel electrophoresis (PFGE) and 16S rRNA sequencing. In the past, only single 'type strains' were sequenced to reveal information on the genome content of the species in question

due to the high costs of WGS (Avasthi *et al*., 2011). Recent advances have enabled large-scale sequencing projects to investigate variation within and between bacterial populations. Evolutionary forces of recombination and selection are responsible for shaping microbial genomes and maintaining diversity, leaving signatures that can be identified using comparative population genomics. Recently, studies have analysed genome sequences – for example to identify genes under positive selection – in order to discover ecologically distinct populations of bacteria and how they adapt to different niches. Whole-genome sequence data, when analysed with the appropriate statistical and computational methods, can provide insights into the structure and function of closely-related populations of microbes in their natural habitats. Ultimately, these analyses reveal that microbial evolution is usually driven by a divergence in populations that are adapted to distinct ecological niches. Lineage distinctness is then maintained by barriers to gene flow, which in some cases, are a consequence of ecological specialisation (Shapiro *et al*., 2012).

The traditional niche theory of evolution would assume that organisms of different species occupy different ecological niches due to their species-specific properties and the niches are of limited similarities. In contrast to this, the neutral model of evolution assumes that microbial species are ecologically and functionally equivalent and that stochastic processes (random drift of mutant alleles that are selectively neutral or nearly neutral) are the main factors shaping species' distributions and community structure (Hubbell, 2001). There are at least some examples where neutral models are able to reconstruct and predict relative species abundance in many environmental microbial communities (Woodcock *et al*., 2007; Sloan *et al*., 2006). More recent studies however, particularly those on the human and animal microbiome, seem to provide more negative cases than positive cases with regards to the neutral model (Nemergut *et al*., 2013). It is perhaps most likely that both niche and neutral mechanisms are in effect in microbial communities, but niche effect is often more prevalent than the neutral effect (Jeraldo *et al*., 2012; Dumbrell *et al*., 2010).Many bacterial lineages have a history of frequent and continuous horizontal gene transfer and loss, as evidenced by vast differences in genome content, even among isolates that are closely related. The evolution and maintenance of ecologically distinct populations of bacteria under such conditions of rapid gene exchange has been an area of investigation for many researchers. A review of recent population genomic data, by Polz, Alm and Hanage (2013), demonstrated the importance of habitat and niche in directing horizontal gene transfer and evolution. This led to a model of ecological speciation through a mechanism of gradual genetic isolation as a result of divergent niche-associated populations, where subpopulations can evolve through gene exchange within the local gene pool (Polz, Alm and Hanage, 2013). This insight provides an explanation for why, despite the potential for free

gene flow, bacteria remain genotypically and phenotypically clustered. A population genomics study by Shapiro *et al*. (2012) provides an example of how ecological divergence may drive genotypic differentiation of bacteria. Two populations of *Vibrio cyclotrophicus*, that are almost genetically indistinguishable but are ecologically divergent, were subjected to whole-genome sequencing to investigate how the populations had diversified. These populations coexist in coastal oceans but exhibit differential proclivity for occurring as free-living bacteria or attached to zoo- and phytoplankton. Analysis of whole-genome sequence data revealed that in these two recently diverged populations, ecological differentiation had occurred through a mechanism of gradual genetic isolation in which a few genomic regions, rather than whole genomes, have swept through subpopulations in a niche-specific manner (Shapiro *et al*., 2012). Furthermore, analysis of homologous recombination events within and between the populations demonstrated that both populations had been actively recombining in the past, however the most recent recombination events had become population-specific, suggesting gradual separation of the gene pools and independent evolutionary trajectory of these populations (Shapiro *et al*., 2012). These inferences would further support a model explaining genotypic cluster formation, in which the ancestral, ecologically uniform, and recombining population of bacteria diverges into unique, ecologically distinct subpopulations (Shapiro *et al*., 2012).

Genome comparisons in another previous study have shown that several clades of *Escherichia* spp. isolated from predominantly non-host environments are more adapted to life outside of hosts, whilst gastrointestinal clades of *E. coli* possessed genomic features adapting them to the human gut environment (Luo *et al*., 2011). In an approach using the inventory of gene functions derived from the human microbiome project (Qin *et al*., 2010), the authors showed that genes that are common in other gut bacteria are also present in the gut-associated *E. coli*, but not in the environmental clades. Furthermore, genome-based analysis of recombination rates indicated that these environmental relatives of *E. coli* have historically not shared ecology with *E. coli* strains that classically inhabit the gastrointestinal tract of humans. The authors described frequent recombination among the environmental clades and among the intestinal *E. coli* clades, but not between the environmental and intestinal clades, providing evidence for an ecological barrier to recombination (Luo *et al*., 2011). Further phylogenetic analysis of the environmental clades by Cohan and Kopac (2011) confirmed each environmental clade to be distinguished as ecologically distinct from the human intestinal *E. coli*, and that finer divisions within each clade suggests that a greater extent of ecological diversity exists among these bacteria.

Given that evidence of population divergence associated with ecological specialisation has been described in bacteria previously, it raises the question of whether we can use the same approach

to uncover novel information on the ecology of clinically important and well-studied bacterial pathogens. The current study proposes a similar whole-genome sequencing-based approach of single isolates to conduct population genomic analyses of closely related genomes obtained from various environments. Two particular species, *Yersinia pseudotuberculosis* and *Escherichia coli*, represent 'model' organisms for the study of microbial population genomics due to their environmental ubiquity and culturability under laboratory conditions. Both *Yersinia pseudotuberculosis* and *Escherichia coli* will be investigated in the present study to determine whether population genomics can provide valuable further insights into the ecological and genetic structures of these important human pathogens.

1.2. *Yersinia pseudotuberculosis*

1.2.1. The pathogenic *Yersiniae*

The genus *Yersinia* belongs to the Enterobacteriaceae, a large and diverse group of Gram-negative bacteria that includes many harmless commensals, along with pathogenic organisms such as those belonging to the genera *Salmonella*, *Escherichia*, *Klebsiella*, and *Shigella*. Definition of the genus *Yersinia* is based on classical systematics and biochemical species-classification methods, which resulted in the description of *Yersinia* being a highly diverse genus comprising 18 distinct species (Savin *et al*., 2014; Hurst *et al*., 2011; Murros-Kontiainen *et al*., 2011a; 2011b; Merhej *et al*., 2008; Sprague *et al*., 2008; Sprague and Neubauer, 2005; Wren, 2003; Carniel, 2003). Of these, *Y. enterocolitica*, *Y. pseudotuberculosis*, and *Y. pestis* are the most studied and are the only species that cause disease in mammals, including humans. *Y. enterocolitica* and *Y. pseudotuberculosis* are both zoonotic pathogens that cause self-limiting enteric infections in humans, whilst *Y. pestis*, the causative agent of plague, is a pathogen in fleas and rodents that can occasionally be transmitted to humans (McNally *et al*., 2016b). The further 15 known species of the genus are commonly isolated from soil and aquatic environments and are typically non-pathogenic to mammals. Some *Yersinia* species however, are pathogenic in other hosts, such as *Yersinia ruckeri* (Sulakvelidze, 2000), which causes red mouth disease in Salmonidae, and *Yersinia entomophaga*, which has insecticidal activity (Hurst *et al*., 2011). With regards to the phylogenetic relation of the pathogenic *Yersiniae*, *Y. enterocolitica* represents a distant relative of *Y. pseudotuberculosis* and *Y. pestis*; their relatedness is often compared to that between *Escherichia coli* and *Salmonella* species (Achtman *et al*., 1999). Recent phylogenetic investigation based on whole-genome single-nucleotide-polymorphism (SNP)-based analysis of the entire *Yersinia* genus has allowed an accurate assessment of its population structure (Reuter *et al*., 2014). This robust sequence-based taxonomy revealed that the genus contains 14 distinct

species clusters, which differed from the existing taxonomic description of the genus that was largely constructed on the basis of biochemical differences and 16S rRNA gene phylogeny (McNally *et al*., 2016b). The analysis also showed that the mammalian pathogens *Y. enterocolitica* and the *Y. pseudotuberculosis*–*Y. pestis* species complex form separate branches at opposite ends of the *Yersinia* phylogenetic tree and do not cluster together as was previously believed, and are thus genetically distinct (Reuter *et al*., 2014). The chromosomal DNA of *Y. pestis* and *Y. pseudotuberculosis* is extremely similar and global phylogenomic studies have identified recent evolution of *Y. pestis* from a clone of *Y. pseudotuberculosis*, as a result of gene loss and subsequent global dissemination (Reuter *et al*., 2014; Morelli *et al*., 2010; Achtman *et al*., 1999). The dispensability of metabolic functions in *Y. pestis* can be explained by adoption of a lifestyle which bypasses the gut infection phase. Proteins necessary for transmission by the faecal-oral route would therefore no longer be needed, resulting in the lack of selective pressure against mutations in genes such as *ure* (urease against gastric acid), *inv*, *ail*, and *yadA* (all essential for translocation across the intestinal barrier; Achtman *et al*., 1999). One commonality between the three pathogenic species is that they all share a ~70 kb plasmid that encodes for various *Yersinia* outer proteins (Yops) that are key virulence factors for pathogenesis. The enteropathogenic *Yersiniae* are genetically more diverse than the more recently evolved plague bacterium *Y. pestis*. *Y. enterocolitica* demonstrates greater diversity and it is categorised into 6 biogroups and more than 50 different serotypes, whilst *Y. pseudotuberculosis* has been classified into 21 distinct serotypes (Gage, 2012).

1.2.2. Hidden ecological patterns in the pathogenic *Yersiniae*

Large-scale population genomic analyses have previously been carried out for the human pathogenic *Yersiniae*, *Y. pestis* and *Y. enterocolitica* (Reuter *et al*., 2015; Reuter *et al*., 2014; Morelli *et al*., 2010), enabling a high-resolution understanding of the ecology, evolution, and population structure of these organisms. A population genomic investigation of *Y. enterocolitica* was performed in a recent study by our group, which involved carrying out pan-genome analysis to examine patterns of recombination in both the core and accessory genomes of the species (Reuter *et al*., 2015). This study highlights a restriction in genetic flow between phylogroups of *Y. enterocolitica*; when gene flow does occur, it is largely unidirectional with one phylogroup acting primarily as a reservoir for recombination with the rest of the species. Furthermore, the data uncovered hidden ecological patterns suggesting that the genetically distinct phylogroups of *Y. enterocolitica* may be ecologically separated, with phylogroup 1 (PG1) being ubiquitous and most commonly isolated from non-human environments, whilst PG2–5 are more commonly associated with human disease cases (Reuter *et al*., 2015). This parallels with a microbial

ecological study of the model bacterial species *Escherichia coli*, where whole-genome sequencing revealed that core genome recombination occurs between environmental isolates or between human/animal isolates, but never between environmental and human/animal isolates (Luo *et al*., 2011). Within the important multidrug-resistant nosocomial pathogen *Enterococcus faecium*, it has been suggested that subpopulations colonise distinct hospital niches, and once adapted to these new environments, these populations become isolated and recombination with other populations decline (Willems *et al*., 2012). The observation of host-restricted lineages of *Campylobacter jejuni* in another study (Sheppard *et al*., 2014) provides further supportive evidence of ecological separation playing a major role in limiting genomic recombination, and thus shaping the evolution of an important bacterial species through the formation of distinct ecotypes. Given the observation of ecologically separated lineages in *Y. enterocolitica*, and considering that *Y. pseudotuberculosis* is also a member of the enteropathogenic *Yersiniae* and is heterogeneous and ubiquitous in nature, it would be reasonable to investigate whether the ecology of this species overlaps with genetic patterns.

1.2.3. An introduction to the *Y. pseudotuberculosis* species

First isolated in 1883 from tuberculosis-like lesions in guinea pigs (Paff, Triplett and Saari, 1976), *Y. pseudotuberculosis* are characterised by Gram-negative rods with rounded ends (coccobacilli) and are facultative anaerobes. As with the rest of the genus, they are catalase-positive but oxidase-negative, and are relatively slow growing in comparison to other members of the Enterobacteriaceae. Much like *Y. enterocolitica*, *Y. pseudotuberculosis* is a cold-tolerant species that is motile at temperatures below 30 °C but are non-motile at temperatures above 37 °C. One feature of the genus *Yersinia* is that its members are well-adapted to survive in the environment with an ability to grow in conditions of minimal nutrients and at temperatures ranging from 4–43 °C (Brubaker, 1991). *Y. pseudotuberculosis* and *Y. enterocolitica* are the most divergent species of the genus, and are now thought to have gained pathogenicity independently, although they cause very similar gastrointestinal diseases in humans and animals. The two organisms also share pathogenicity islands and other virulence-associated genes, which are suggested to have been gained independently, perhaps initially from other genera, and then later via transfer within the genus (Reuter *et al*., 2014). As whole-genome sequence data for the genus became increasingly available, many strains of bacteria first typed as *Y. pseudotuberculosis* were consequently reclassified into other species that are now included in the phylogenetic group known as the '*Y. pseudotuberculosis* complex' (Laukkanen-Ninios *et al*., 2011). This species complex includes *Y. pseudotuberculosis/Y. pestis*, *Y. similis* (Sprague *et al*., 2008), and the recently characterised *Y. wautersii* (previously referred to as the 'Korean group'),

which is proposed to have pathogenic potential (Savin *et al*., 2014). *Y. pseudotuberculosis* and *Y. pestis* share ≥ 97% nucleotide sequence identity for most of their chromosomal genes (Koskela *et al*., 2015). Due to this close evolutionary relationship with *Y. pestis, Y pseudotuberculosis* is believed to have been the progenitor of the plague bacillus and is considered a model species for bacterial evolution (McNally *et al*., 2016b). *Y. pestis* is effectively a clone of *Y. pseudotuberculosis*, estimated to have evolved from its ancestor approximately 1,500–6,400 years ago, in Asia (Achtman *et al*., 1999).

1.2.4. Pathogenesis of *Y. pseudotuberculosis*

1.2.4.1. Transmission of the bacteria to humans

Y. pseudotuberculosis causes zoonoses in a wide range of hosts, including both wild and domesticated animals and birds (McNally *et al*., 2016b). With *Y. pseudotuberculosis* being a zoonotic pathogen, it can therefore be transmitted to humans through various routes. *Y. pseudotuberculosis* infection, though less common than those caused by *Y. enterocolitica*, has also been implicated in foodborne disease in humans, which is known as yersiniosis. Transmission of the bacterium is usually via the faecal–oral route, and human infection can result from ingestion of contaminated food products or water, or possibly through cross-contamination during food preparation. Typical sources of infection include dairy products, inadequately cooked meat, and certain vegetables such as lettuce and raw carrots (Kangas *et al*., 2008; Nuorti *et al*., 2004). Both *Y. pseudotuberculosis* and *Y. enterocolitica* are cold-tolerant species that can survive and proliferate slowly at 4 °C, accounting for growth in cold-stored foodstuffs. *Y. pseudotuberculosis* is also found widely in the environment, including soil, and in animals it causes tuberculosis-like disease. Pigs, rodents, rabbits, sheep, goats, cattle, horses, dogs, cats, deer, and sometimes birds serve as reservoirs for *Y. pseudotuberculosis*. Person-to-person transmission has also been reported, though less frequently, as has transmission via blood transfusion (Chiles *et al*., 2002). Most humans can serve as asymptomatic carriers of *Y. pseudotuberculosis*; however, several cases have been linked to handling and close-contact with infected animals. Patients exhibiting clear symptoms of infection tend to shed significant amounts of bacteria for up to 2–3 weeks. Infected individuals who are left untreated can become carriers and shed bacteria for as long as 2–3 months (Gage, 2012). Successful infection of the host requires a reasonably large dose of bacteria (median infective dose of 10^8–10^9 bacteria), and the incubation period is believed to be 3–7 days after ingestion. Although, incubation periods of 2–20 days have been seen in sporadic outbreaks, and symptoms typically appear at an average time of 4 days after exposure (Gage, 2012).

1.2.4.2. Epidemiology of *Y. pseudotuberculosis*-associated yersiniosis

Human yersiniosis caused by *Y. pseudotuberculosis* is usually sporadic and such cases generally occur worldwide (Sunahara, Yamanaka and Yamanishi, 2000). The majority of nationwide gastrointestinal outbreaks of foodborne infection have been reported in countries of the Northern Hemisphere or countries of largely temperate climates. Most cases of *Y. pseudotuberculosis* infection occur in the winter and early spring, a trend which is likely related to the enhanced growth characteristics of this pathogen in cold temperatures that occurs during long-term storage of vegetables during the winter (Galindo *et al*., 2011). The first conclusive report of a community outbreak of *Y. pseudotuberculosis* was described in 1984 among schoolchildren in Kurashiki, Japan (Inoue *et al*., 1984). *Y. pseudotuberculosis* infection in Europe would appear to be highly prevalent in Germany, and there is some suggestion that higher meat consumption in this country, particularly pork, when compared to other European nations might correlate with the higher incidence of yersiniosis in Germany (Galindo *et al*., 2011). The prevalence of *Y. pseudotuberculosis* infection in the United States is currently unknown. Sporadic cases of *Y. pseudotuberculosis* infection are likely to be underreported because stool cultures are not routinely requested for patients presenting with mild, self-limiting clinical features, such as diarrhoea. Furthermore, the need for specific differential culture media for isolation has restricted the presence of active surveillance in many other countries, including those of Africa, Asia, the Middle East, Latin America, the Caribbean, and others.

A report by Nuorti *et al*. (2004) at the National Public Health Institute of Finland provided solid documentation that *Yersinia pseudotuberculosis* can be transmitted through food, following a nationwide outbreak in 1998 that was detected by routine surveillance for *Yersinia* species. In this widespread outbreak, contaminated iceberg lettuce was strongly implicated as the vehicle of *Y. pseudotuberculosis* serotype O:3 infections, with 71% of case patients reported having eaten iceberg lettuce prior to hospitalisation. Prior to this study, *Y. pseudotuberculosis* was presumed to be a possible foodborne pathogen, by virtue of its similarity to *Y. enterocolitica*, but the evidence for this assumption was limited to a few suggestive clusters and to a large Canadian outbreak in 1998, which was epidemiologically linked to pasteurised homogenised milk and was the first recognised outbreak of *Y. pseudotuberculosis* serotype O:1b (Nowgesic *et al*., 1999). The investigation in Finland is the first to trace an outbreak of human illness to a likely environmental reservoir via contaminated food, and suggests that some cases of yersiniosis, which appear to be sporadic, may be part of unrecognised outbreaks caused by contaminated fresh produce (Nuorti *et al*., 2004). In 2004, an outbreak of several cases of gastroenteritis in schoolchildren in northern Finland was reported, and at the same time, an increase in *Y. pseudotuberculosis* cases was reported from other parts of the country. Kangas and co-workers

(2008) carried out an investigation which provided microbiologic and epidemiologic evidence that traced the school outbreak to consumption of raw carrots contaminated at the production farm. The long-term storage of raw carrots at cold temperatures would have favoured the growth of *Y. pseudotuberculosis* and thus result in human infection (Kangas *et al*., 2008). More recently, in 2014, a sustained outbreak of yersiniosis due to *Y. pseudotuberculosis* was reported in all of the major cities of New Zealand (Williamson *et al*., 2016). This study presented one of the largest globally reported outbreaks of human *Y. pseudotuberculosis* infection to date, with a total of over 200 laboratory-confirmed cases of infection reported. Prior to our study, the New Zealand study had provided the most inclusive genome-scale analysis of a *Y. pseudotuberculosis* population. Genomic and epidemiological analyses indicated a single point-source contamination of the food chain, with subsequent nationwide distribution of contaminated produce. Furthermore, the analysis carried out in this study involved incorporation of publicly available reference genomes within the context of a globally and taxonomically diverse dataset. This revealed that *Y. pseudotuberculosis* is a highly diverse species and that the New Zealand strains represented a geographically isolated clade of *Y. pseudotuberculosis* (Williamson *et al*., 2016). This study serves as an example of the exploitation of pathogen genome sequence data and the contribution of population genomic analysis to understanding the epidemiology and spatiotemporal spread of clinically important bacterial pathogens.

1.2.4.3. Virulence factors of *Y. pseudotuberculosis*

The genomes of the pathogenic *Yersiniae*, *Y. pseudotuberculosis*, *Y. enterocolitica*, and *Y. pestis*, are 97% identical, however the three organisms cause different types of diseases in humans, despite sharing a tropism for lymph nodes (Fig. 1.1; Bergsbaken and Cookson, 2009). The different routes of infection, types of infections, and severity of disease in humans caused by the pathogenic *Yersiniae* are influenced by the distribution of shared and unique virulence-associated genes (VAGs). Both chromosomal and plasmid-derived virulence-associated genes contribute to the pathogenesis of *Yersinia* species and the establishment and progression of yersiniosis. Recent work by our group, which delineated the phylogeny of the genus *Yersinia*, including 31 isolates of *Y. pseudotuberculosis*, revealed important information on the population structure and collection of virulence genes which define the genus (Reuter *et al*., 2014). Much like the other human pathogenic *Yersinia*e, the pathogenicity of *Y. pseudotuberculosis* is dependent on the presence of a ~70 kb virulence plasmid associated with *Yersinia* virulence, pYV (Portnoy and Falkow, 1981). The pYV plasmid differentiates pathogenic strains from non-pathogenic strains, because it is required for virulence. Although the pYV plasmid is commonly thought of as a single entity, it is highly variable between species and strains of species,

containing different origins of replication as well as exhibiting variable genetic architecture (Reuter *et al*., 2014; Portnoy and Falkow 1981). However, all pathogenic *Yersinia* strains harbour the large genetic locus encoding the Ysc type III secretion system (T3SS), located on the pYV plasmid (Reuter *et al*., 2014; Portnoy and Falkow, 1981). The Ysc T3SS, which was the first T3SS to be fully characterised, functions by mediating the targeted delivery of *Yersinia* outer protein (Yop) effector proteins into host cells. Yops are the key virulence factors in all pathogenic *Yersiniae*; injection of Yops and contact with host macrophages results in the inhibition of the pro-inflammatory cytokine response and ultimately, apoptotic death of the infected macrophages (Cornelis and Wolf-Watz, 1997). Additional virulence determinants that are variably present in *Y. pseudotuberculosis* include the chromosomal high-pathogenicity island (HPI), which is present in almost all European strains of *Y. pseudotuberculosis* serotype O:1 (Carniel, 1999). HPI encodes proteins that are involved in the biosynthesis, regulation, and transport of the iron uptake system Yersiniabactin. *Y. pseudotuberculosis* also harbour the *Yersinia* adhesion pathogenicity island (YAPI), which includes a pilin gene cluster, the *pil* operon, encoding a type IV pilus that contributes to pathogenicity (Collyn *et al*., 2004). The vast majority of all *Y. pseudotuberculosis* strains originating from the Far East of Asia additionally produce one of three variants of a chromosomally-encoded novel superantigenic toxin YPM (the *Y. pseudotuberculosis*-derived mitogen), encoded by the *ypm* gene. The YPMa variant is encoded by the *ypmA* gene (Ramamurthy *et al*., 1997) and plays a more crucial role in systemic infections than in gastroenteritis. YPMb and YPMc are the other two variants of the superantigen, which are encoded by the *ypmB* and *ypmC* genes, respectively (Ramamurthy *et al*., 1997). A small conserved RNA chaperone protein, known as Hfq, is essential for the full spectrum of virulence in a variety of pathogenic bacteria, including both *Y. pseudotuberculosis* and *Y. enterocolitica*. The Hfq protein plays an important role in the regulation of motility, intracellular survival, and production of T3SS effectors in *Y. pseudotuberculosis* (Schiano, Bellows and Lathem, 2010).

1.2.4.4. Pathophysiology of *Y. pseudotuberculosis* infection

After ingestion by humans, *Y. pseudotuberculosis* pass into the small intestine and adhere to the mucosal lining of the ileum, where intracellular infections in Peyer's patches, mucosal cells, and macrophages can occur (Fig. 1.1; Gage, 2012). Invasion of the ileal mucosa is favoured by the presence of numerous virulence-associated genes coding for fimbriae, flagellar proteins, and adhesins, such as invasin (Inv), YadA, and the attachment invasion locus protein (Ail) (Carniel, 2003). The invasin protein prompts the internalisation of the bacteria, which translocate across the epithelium (Pepe and Miller, 1993). As a result, an inflammatory response occurs causing

the characteristic symptoms of fever, abdominal pain, and diarrhoea, which are typical of acute gastroenteritis and mesenteric lymphadenitis. Ulcerative ileitis, and extraintestinal conditions such as mesenteric adenitis, erythema nodosum, and necrosis within Peyer's patches can arise in more advanced cases (Jalava *et al*., 2006). If the regional defences of the ileum are broken, the bacteria can disseminate and cause sepsis, or even abscesses of the liver or spleen (Kaasch *et al*., 2012). Polyarthritis, a type of arthritis that involves five or more joints and usually associated with autoimmune conditions, may also occur at a later stage of illness, particularly in human leukocyte antigen (HLA)-B27-positive individuals (Gage, 2012). Bacterial cells that are replicating in the intestinal tract can also attack the host lymphoid tissues, much like *Y. pestis*. Invasion of these tissues and resistance against the host defences rely on the possession of the pYV plasmid which encodes genes for various Yops and the V antigen (LcrV) (Gage, 2012). The products of this plasmid work together to silence phagocytic immune cells as well as reduce inflammation, resulting in suppression of the host immune response. This favours persistence of these bacteria in the body, allowing them to replicate extracellularly and form aggregates in the mesenteric lymph nodes, which can lead to septicaemia (Gage, 2012).

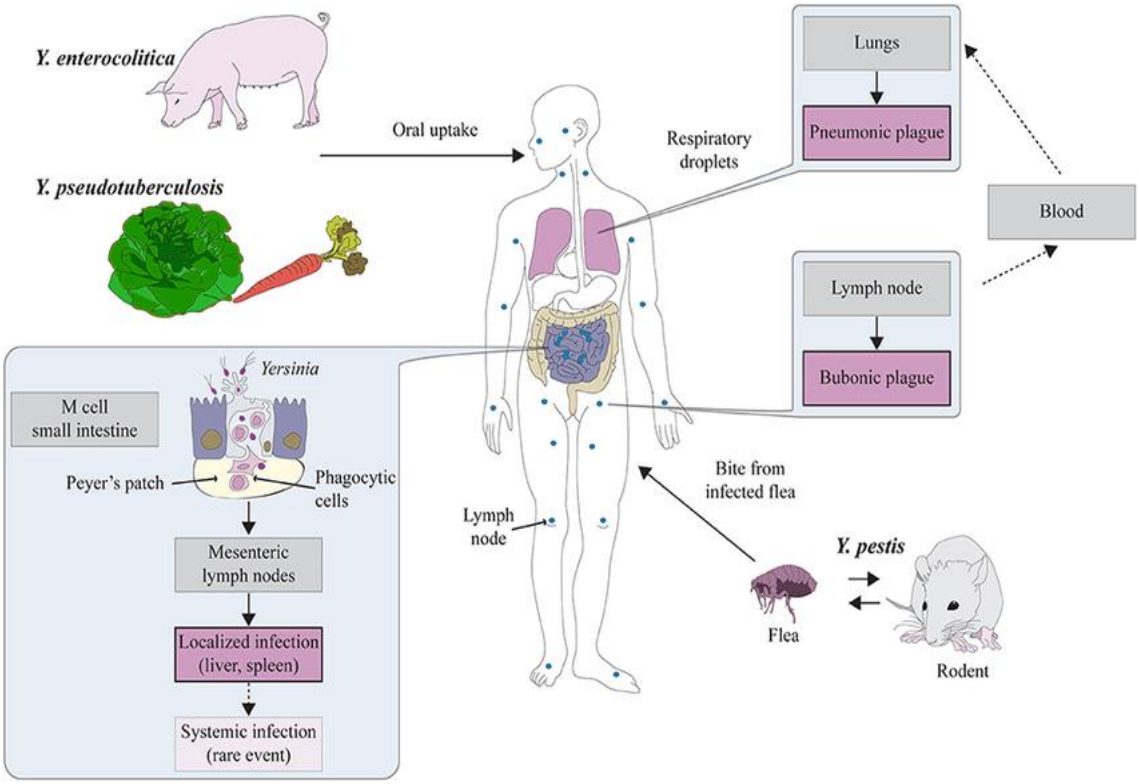


Figure 1.1. Routes of transmission and mechanisms of pathogenesis of the human pathogenic *Yersinia* species.

The routes of transmission for the enteropathogenic *Yersinia* species, *Y. pseudotuberculosis* and *Y. enterocolitica*, are usually associated with the consumption of contaminated raw vegetables (lettuce/carrots) and undercooked meat (mainly pork). They are ingested via the contaminated food and enter the lymphatic system through the M cells in the small intestine. They can then progress to cause a localised infection of the liver or spleen, or in rare cases, lead to systemic infection such as septicaemia. The main reservoirs of *Y. pestis*, on the other hand, are rodents and fleas. Transmission of the bacteria to humans occurs through the bite of an infected flea resulting in bubonic plague. Pneumonic plague can arise when *Y. pestis* reaches the lungs, via the bloodstream, and is transmitted to other individuals through the release and inhalation of respiratory droplets. Figure adapted from Heroven and Dersch (2014).

M cell: microfold cell.

1.2.5. Clinical manifestations of *Y. pseudotuberculosis* infection

After an incubation period of around 3 to 7 days, gastroenteritis usually develops, which can be difficult to distinguish from *Salmonella*- or *Campylobacter*-associated gastroenteritis. Patients with *Y. pseudotuberculosis*-associated infections typically present with clinical syndromes characterised by acute mesenteric adenitis with fever, diarrhoea, vomiting, and bloody stools (Gage, 2012). Young children are more likely to develop enterocolitis, whilst older children more commonly experience acute terminal ileitis, mesenteric adenitis, and systemic disease. Sepsis is generally uncommon and is most likely to occur in individuals with pre-existing conditions and risk factors such as diabetes mellitus, cirrhosis, immunosuppression, older age, and haemochromatosis. Splenic abscesses, meningitis, or endocarditis can develop in septic patients, and the mortality rate in these cases is close to 50%. Erythema nodosum is identified in about one-third of all patients and in approximately 10% of adults (Gage, 2012). *Y pseudotuberculosis* infections are more common in men than in women. However, post-infection complications, such as erythema nodosum and reactive arthritis, are more common in women. Children aged 5 to 15 years comprise greater than 75% of patients with *Y. pseudotuberculosis* infection. Illness typically lasts 1 to 3 weeks, but symptoms may persist in some patients for several months (Gage, 2012). Although fever and self-limiting gastroenteric symptoms are the primary clinical manifestations of *Y. pseudotuberculosis* infection in Europe, those occurring in Japan, Korea, and the Far East of Russia include not only symptoms of gastroenteritis but also a variety of systemic manifestations such as scarletiniform rash, desquamation, and arthritis (Gage, 2012). This disease variant is called Far East scarlet-like fever, which is associated with specific superantigen-containing strains of *Y. pseudotuberculosis*. Two research groups have previously reported considerable geographical heterogeneity between the Far East of Asia and Europe regarding the prevalence of YPMa-producing strains (Ueshiba *et al*., 1998; Yoshino *et al*., 1995). The investigators reported an absence of the *ypmA* gene in strains belonging to serotypes O:1 and O:2 from Europe, however this gene was present in almost all strains belonging to serotypes O:1, O:2, O:4, and O:5 from Asia, indicating the importance of the *ypmA* gene in strains causing Far East scarlet-like fever (Ueshiba *et al*., 1998; Yoshino *et al*., 1995). In Russia and Japan, *Y. pseudotuberculosis* infection is regarded as a national health problem and was added to the national notification system in 1988 (Tseneva *et al*., 2012).

1.2.6. Treatment and prevention

Antibiotics generally do not improve the course of uncomplicated enterocolitis or mesenteric adenitis, and antibiotic treatment regimens are not recommended for intestinal forms of the

disease (Bottone, 1997). However, such therapy has been recommended for immunocompromised patients, individuals with septicaemia, and those with systemic disease or extraintestinal progression of disease. Broad-spectrum cephalosporins, sometimes accompanied by aminoglycosides, have resulted in successful outcomes in patients with extraintestinal forms of yersiniosis, including septicaemia (Bottone, 1997). Antimicrobial resistance is generally less common in *Y. pseudotuberculosis* than in *Y. enterocolitica* (Bonardi *et al*., 2016). Prevention of contracting *Y. pseudotuberculosis* infection relies on measures intended to protect individuals from direct contact with contaminated environments, foods, wastes, as well as other infected humans and animals. Some of these approaches include using proper sewage disposal methods, protecting water supplies from contamination, as well as using appropriate food hygiene, preparation, and storage measures (Bottone, 1997).

1.2.7. Identification and typing of *Y. pseudotuberculosis*

1.2.7.1. Serological characterisation of *Y. pseudotuberculosis*

Diagnosis of yersiniosis begins with isolation of the causative organism, using selective media, from the human host's stool, blood, or vomit, and sometimes at the time of appendectomy (surgical removal of the appendix). Biotyping and serotyping of *Y. pseudotuberculosis* can provide useful epidemiologic information when tracking the source of community outbreaks. Classically, identification and typing of *Y. pseudotuberculosis* is commonly based on the lipopolysaccharide (LPS) O-antigen, which is used for the serological characterisation of strains. The LPS is a lipoglycan present on the cell surface of most Gram-negative bacteria, and for many pathogens including *Y. pseudotuberculosis*, the LPS represents an important virulence factor (Ho *et al*., 2008). The structure of the LPS is comprised of three key components: lipid A, the core oligosaccharide, and O-specific polysaccharide (OPS; also termed the O-antigen). The O-antigen serotyping scheme used for *Y. pseudotuberculosis* involves differentiation of the variable OPS subunits (O units) using serology. It has been suggested that the same serotyping scheme can be implemented for the identification and typing of all members of the *Y. pseudotuberculosis* complex, which includes *Y. pestis*, *Y. similis*, and *Y. wautersii* (Savin *et al*., 2014; Laukkanen-Ninios *et al*., 2011). Some of the 15 major serotypes (O:1 – O:15) of *Y. pseudotuberculosis* are divided into ten subtypes (O:1a, O:1b, O:1c, O:2a, O:2b, O:2c, O:4a, O:4b, O:5a, O:5b), thus resulting in a total of 21 known serotypes (Skurnik, Peippo and Ervela, 2000). The efficacy of serotyping in investigating potential outbreaks of *Y. pseudotuberculosis* in Europe is very limited, due to the vast majority of strains isolated from human cases belonging to serotypes O:1 and O:3, whereas serotypes O:2 and O:4–O:15 are primarily found in Asia (Fukushima *et al*., 2001). Previous studies have indicated that a large proportion of strains isolated from human cases

belong to serotypes O:1a, O:1b, and O:3 (Williamson *et al*., 2016; Laukkanen-Ninios *et al*., 2011). The application of serotyping methods to *Y. pseudotuberculosis* therefore provides only low-level resolution when studying the ecology and diversity of this important human pathogen, thus suggesting the need for higher resolution typing methods.

1.2.7.2. *Y. pseudotuberculosis* multilocus sequence typing (MLST)

Population genomic analyses of *Y. pseudotuberculosis* have revealed greater details about the population structure of this species and its relationship to the other closely related species of the *Y. pseudotuberculosis* complex. Sequence-based analysis provides a more detailed classification system than biochemical characterisation to differentiate between phenotypically indistinguishable strains, as has been demonstrated with the recently described species *Y. similis* (Sprague *et al*., 2008), which is biochemically similar to *Y. pseudotuberculosis*. Multilocus sequence typing (MLST) is a molecular typing method, developed in 1998, with the aim of improving the portability and accuracy of epidemiological and molecular typing information. It was proposed to provide a highly discriminating typing system for the unambiguous characterisation of isolates of most bacteria and other organisms (Maiden *et al*., 1998). Soon after MLST analysis was first described, a research group led by Mark Achtman developed a new MLST scheme for *Y. pseudotuberculosis* (Laukkanen-Ninios *et al*., 2011). The MLST scheme was based on allele profiling of fragments of seven *Y. pseudotuberculosis* housekeeping genes: *glnA*, *thrA*, *tmk*, *trpE*, *adk*, *argA*, and *aroA* (Laukkanen-Ninios *et al*., 2011). The MLST scheme was applied to a diverse collection of 417 isolates from 29 countries representing all continents, in order to characterise the molecular epidemiology, population structure, and diversity of *Y. pseudotuberculosis*, adding further granularity to the serotype classification method. This analysis grouped serotype O:3 strains into a distinct clone designated ST19, which consists of single-locus variants ST50 and ST57. It has previously been suggested that these strains are associated with lowered pathogenicity (Fukushima *et al*., 2001); however, these strains harbour the pYV plasmid and the chromosomal *inv* gene and are sometimes responsible for fatal diarrhoea in cattle (Martins, Bauab and Falcao, 1998). It was revealed that serotype O:1 strains formed a distinct clade of strains which represented a large number of sequence type complexes. This indicated a genotypically diverse population of bacteria within the serotype O:1 group and thus an overall high diversity among disease-causing *Y. pseudotuberculosis* strains (Laukkanen-Ninios *et al*., 2011).

1.3. *Escherichia coli*

1.3.1. An introduction to the *E. coli* species

The second model organism used in this study to investigate microbial ecology is *Escherichia coli*. The species was first discovered in the late nineteenth century by Theodor Escherich, a German-Austrian paediatrician who believed that intestinal diseases in infants were caused by the infant gut microbes. Escherich isolated *E. coli* from the faeces of infants and subsequently named the organism '*Bacterium coli commune*', which was later renamed *Escherichia coli* after Escherich's death (Hacker and Blum-Oehler, 2007). *E. coli* is a member of the Enterobacteriaceae, which are widespread in nature and are most commonly found in the intestinal tracts of mammals, but can also be isolated in high concentrations from soil, water, and agricultural land, due to faecal contamination (Winn *et al*., 2006). The genus *Escherichia*, which contains six distinct species (*E. coli*, *E. albertii*, *E. fergusonii*, *E. vulneris*, *E. hermannii*, and *E. marmotae*), is the most commonly encountered genus of the Enterobacteriaceae in the clinical setting. Within the genus *Escherichia*, *E. coli* is the only member to exhibit pathogenic traits, and due to its environmental ubiquity and ability to grow easily under laboratory conditions, it is also one of the most extensively studied model organisms in microbial genetics.

The bacterium is a Gram-negative, rod-shaped, non-sporulating, facultatively anaerobic coliform, typically 2 µm in length and 0.5 µm in diameter (Winn *et al*., 2006). The bacterial cell characteristically has peritrichous flagella projecting in all directions, which enable a level of motility for the bacterium, a trait which is thought to contribute to bacterial fitness and virulence (Lane *et al*., 2005). Despite the metabolic complexity of *E. coli*, the species can be distinguished from other members of the Enterobacteriaceae based on distinct metabolic characteristics, which includes its ability to produce indole from the metabolism of the amino acid tryptophan. *E. coli* can also reduce nitrate to nitrite, and produce pyruvic acid from glucose, which can be demonstrated using the methyl-red indicator (Winn *et al*., 2006). Most strains of *E. coli* are also able to ferment lactose, which is often used as a feature to distinguish *E. coli* from the closely related species *Shigella* (Winn *et al*., 2006), and this can easily be demonstrated by a change in colour of the indicator on differential culture media, such as CLED (cysteine-, lactose-, and electrolyte-deficient) agar and MacConkey agar. The metabolic adaptability of *E. coli* may provide it with a fitness advantage over fastidious organisms, allowing it to survive and multiply in nutritionally poor microenvironments, such as the human bladder.

1.3.2. Commensal *E. coli*

Commensal *E. coli* are strains which colonise humans and animals, but do not trigger an immune response or cause disease in the host. *E. coli* are one of the first bacteria to colonise the human intestine, with initial colonisation occurring during the early stages of infancy. This may be due to the transmission of *E. coli* from the mother to the neonate, or even through nursing staff in the hospital (Watt *et al*., 2003). *E. coli* are lifelong commensal colonisers of adults, where the species has adapted its metabolism very successfully to the nutritional ecological niche of the gut, withstanding competition from more than 500 other bacterial species (Tenaillon *et al*., 2010). As a consequence of being a gut microbe, *E. coli* are regularly excreted into the wider environment, and despite being intricately adapted to life inside a host, *E. coli* must also be adapted to successfully acclimatise to harsher conditions outside of the host (Savageau, 1983). *E. coli* are able to persist in the environment until the next host consumes viable bacteria in contaminated food or water. One of the stress conditions faced by *E. coli* following ingestion is the acidic pH of the stomach, which it survives by evoking protective acid resistance systems (Foster, 2004). *E. coli* must then acquire the nutrients required to proliferate once reaching the colon. Successful colonisation of the colon by *E. coli* depends on competition for nutrients with an extremely large and diverse microbiota, penetration of the intestinal mucus layer, ability to avoid the host defences, and grow rapidly, beyond the turnover rate of the mucus layer (Conway and Cohen, 2015). *E. coli* persists in the mucus whilst some cells that are sloughed off into the lumen of the intestine are eliminated in the host faeces and the cycle repeats with a new host. This circle of colonisation and extraintestinal survival represents the life cycle for both commensal and pathogenic strains of *E. coli*.

In addition to colonising the intestinal systems of humans, commensal *E. coli* also typically colonise the urinary tract and cause a condition known as asymptomatic bacteriuria (ABU), which is characterised by the presence of *E. coli* in urine but the host does not exhibit any of the classical symptoms associated with a urinary tract infection (UTI). It was traditionally thought that commensal *E. coli* strains simply did not possess virulence-associated genes in their genome and were therefore unable to cause disease (Mabbett *et al*., 2009). In fact, more recent studies of the ABU isolate *E. coli* 83972 have shown that this strain does possess virulence genes. Rather they have become attenuated due to deletion events and are no longer functional. It was postulated that this occurred through adaptation to the new environment of the bladder, giving avirulent strains the advantage of residing within the host without eliciting the lethal action of the immune response (Mabbett *et al*., 2009).

1.3.3. Intestinal pathogenic *E. coli* (IPEC)

Despite the fact that *E. coli* exists as a commensal species in the intestinal microbiota of a variety of animals including humans, not all strains are harmless, and some can cause debilitating and sometimes life-threatening diseases in humans as well as mammals and birds (Belanger *et al*., 2011). Pathogenic *E. coli* strains are classified into two groups: those that cause intestinal infection and those that cause extraintestinal infection. Unlike the intestinal commensal *E. coli*, intestinal pathogenic *E. coli* strains have acquired virulence-associated genes (VAGs), giving them the ability to cause many serious intestinal diseases (Kaper, Nataro and Mobley, 2004). Among the intestinal *E. coli* there are eight recognised pathotypes (Fig. 1.2): enteropathogenic *E. coli* (EPEC), enterotoxigenic *E. coli* (ETEC), enteroinvasive *E. coli* (EIEC), enteroaggregative *E. coli* (EHAC), enterohaemorrhagic (Shiga toxin-producing) *E. coli* (EHEC/STEC), diffusely adherent *E. coli* (DAEC), entero-aggregative-haemorrhagic *E. coli* (EAHEC), and adherent invasive *E. coli* (AIEC) (Clements *et al*., 2012). These pathotypes are capable of causing varying severities of disease ranging from mild, self-limiting diarrhoea to diseases such as haemolytic uraemic syndrome (HUS), which is characterised by haemolytic anaemia, acute kidney failure, and low platelet count. The type of disease caused by an intestinal pathogenic strain of *E. coli* is influenced by the types of virulence-associated genes that the strain possesses. The profile of VAGs and the type of disease caused by strains is used to broadly classify intestinal *E. coli* into one of the eight pathotypes, which are used to inform diagnosis and treatment of diseases. There is increasing crossover between *E. coli* strains of different pathotypes as a result of horizontal recombination, and thus new pathotypes are often proposed, increasing our understanding of the evolution of this important pathogen to humans.

Enteropathogenic *E. coli* (EPEC) are pathogens that colonise the small intestines and are a common cause of severe, watery diarrhoea in infants of developing countries (Trabulsi, Keller and Tardelli Gomes, 2002). In industrialised countries, the prevalence of these organisms has decreased, but they continue to be an important cause of diarrhoea (Nataro and Kaper, 1998). The primary mechanism of EPEC pathogenesis involves attaching and effacing (A/E) lesions, which are characterised by microvilli destruction, adherence of the bacteria to the intestinal epithelium, pedestal formation, and aggregation of polarised actin and other elements of the cytoskeleton at sites of bacterial attachment (Nataro and Kaper, 1998).

Enterotoxigenic *E. coli* (ETEC) are strains that colonise the mucosa of the small intestines of humans and cause a mild, self-limiting diarrhoeal disease. In immunocompromised hosts, the disease may progress to a more severe, longer lasting infection akin to that of cholera. ETEC is one of the leading causes of diarrhoea in the developing world, and it is the most common cause

of travellers' diarrhoea, a significant disease in children estimated to be responsible for approximately 210 million cases and 380,000 deaths per year (Jelinek and Kollaritsch, 2008).

Enteroinvasive *E. coli* (EIEC) is an intracellular *E. coli* pathotype which is genetically and biochemically very similar to *Shigella*, another genus of enteric pathogen within the Enterobacteriaceae. Both EIEC and *Shigella* possess the *ipaH* invasive gene (Kaper, Nataro and Mobley, 2004) and cause invasive disease which may lead to severe illness in otherwise healthy individuals. Since EIEC and *Shigella* are so closely related it has previously been suggested that they should be classified as a single pathotype of *E. coli*, however *Shigella* keeps its species designation due to the association with the disease shigellosis (Croxen and Finlay, 2010).

Enteroaggregative *E. coli* (EAEC) was first described in 1987 in a child suffering from acute diarrhoea in Lima, Peru (Nataro *et al*., 1987). Since its discovery, EAEC have been associated with persistent diarrhoea in children living in EAEC-endemic areas (Nataro *et al*., 1987), individuals infected with the human immunodeficiency virus (HIV) (Mathewson *et al*., 1995), and as a cause of diarrhoea in travellers from industrialised countries visiting the developing world. The pathogenesis of EAEC is determined by its ability to adhere to intestinal cells, produce enterotoxins and cytotoxins, and induce inflammation of the intestinal wall. EAEC are characterised by the ability to colonise either the small or large intestinal mucosa, but primarily the colon, by aggregative adhesion.

The diffusely adherent *E. coli* (DAEC) are considered a diarrhoeagenic class of organisms that colonise the small intestines, causing diarrhoea in children between the age of 18 months and 5 years in developing countries (Mansan-Almeida, Pereira and Giugliano, 2013). These strains are characterised by the diffuse adherence pattern on cultured epithelial cells HeLa or Hep-2 (Croxen and Finlay, 2010). DAEC strains are able to produce finger-like projections that extend from the surface of infected Caco-2 or HEp-2 cells (Cookson and Nataro, 1996). These projections supposedly "embed" the bacteria, providing some protection against gentamicin but without complete internalisation of the cell, however the role for this phenotype in pathogenesis has not yet been determined.

The enterohaemorrhagic *E. coli* (EHEC) pathotype was first defined in 1983 after two outbreaks of gastrointestinal illness characterised by severe abdominal cramps, watery diarrhoea which progressed to extremely bloody diarrhoea, and was not accompanied by fever (Riley *et al*., 1983; Karmali *et al*., 1983). EHEC was defined based on the serological evidence and presence of a specific cytotoxin derived from these two outbreaks. EHEC strains comprise a subgroup of the Shiga toxin-producing *E. coli* (STEC), which encompasses EHEC and the lesser virulent/avirulent STEC. Owing to their human pathogenicity, some STEC strains are also designated as EHEC

(Nataro and Kaper, 1998). The major defining feature of EHEC is the production of phage-encoded Shiga toxin, Stx1 and/or Stx2, which are responsible for serious disease in humans, such as HUS and HC.

In 2011, an outbreak strain associated with haemolytic-uraemic syndrome and bloody diarrhoea in Europe, was identified as an EAEC strain that acquired the prophage-encoded Shiga toxin of EHEC, thus combining the virulence potentials of two different pathogens (Denamur, 2011). This combination of genomic features, associating characteristics from both EAEC and EHEC, gave rise to a new pathotype: the enteroaggregative-haemorrhagic *E. coli* (EAHEC). This outbreak highlighted the ability of *E. coli* to recombine and produce new combinations of genes, resulting in new lineages. The EAHEC outbreak also provided a good example of the application of modern sequencing technologies to rapidly and accurately identify causative strains, which in the past, relied heavily on low resolution methods such as serotyping of infectious organisms (Denamur, 2011).

The adherent-invasive *E. coli* (AIEC) are a pathotype of *E. coli* that have been implicated in the pathogenesis of Crohn's disease (Darfeuille-Michaud, 2002). AIEC are unusual among intestinal *E. coli* pathotypes in that they are not associated with diarrhoea. The high prevalence of adherent *E. coli* isolated from the ileal mucosa of patients with Chron's disease led to the characterisation of several strains, which failed to detect any virulence-associated genes that are traditionally present in typical pathogenic species. One characteristic of these strains is the ability to adhere to and invade intestinal epithelial cells, as well as the ability to replicate within macrophages, which discerns them from other varieties of *E. coli*, including commensals. These strains were therefore categorised as the specific pathogenic group known as AIEC (Darfeuille-Michaud, 2002).

1.3.4. Extraintestinal pathogenic *E. coli* (ExPEC)

Extraintestinal pathogenic *E. coli* (ExPEC) are facultative pathogens which colonise the gastrointestinal tract of many healthy individuals, where they exist as commensals and do not cause enteric disease, contrary to the intestinal pathogenic *E. coli* (IPEC). ExPEC strains colonise sites outside of the intestinal tract, such as the urinary tract, bloodstream, and brain (Fig. 1.2). ExPEC are considered the primary aetiological agent of urinary tract infections (UTIs), as well as a common cause of bacteraemia and sepsis in the community. Other ExPEC strains are responsible for surgical wound infections, neonatal meningitis, and neonatal sepsis (Ron, 2010; Russo and Johnson, 2003). ExPEC strains which reside in the gastrointestinal tract differ from normal commensal strains, in that they possess virulence traits that allow them to colonise more

inhospitable environments, such as the urogenital tract (Smith, Fratamico and Gunther, 2007). In addition to these bacteria-specific traits, host-specific factors are also required in order to cause disease. ExPEC is therefore considered a necessary but not sufficient cause for extraintestinal *E. coli* infection, and as a result, additional factors are required for an infection to occur (Singer, 2015). Consequently, ExPEC are considered opportunistic pathogens for causing ExPEC-associated disease. When provided with an opportunity in individuals who might be susceptible in some way (e.g., compromised immune system), or through the influence of specific risk factors, the bacterium can be transferred to the urogenital tract where it can cause a UTI (Foxman, 2014). Similar to the intestinal pathogenic *E. coli* described earlier, ExPEC strains are also categorised into pathotypes as defined by the anatomical location of the disease they cause and the molecular virulence-associated genes that they carry, although few traits appear to be exclusive to one specific ExPEC subgroup.

1.3.4.1. Avian pathogenic *E. coli* (APEC)

Avian pathogenic *E. coli* (APEC) are responsible for causing systemic extraintestinal infections such as aerosacculitis, polyserositis, and septicaemia in avian hosts such as chickens, turkeys, and other wild and domesticated birds (Manges, 2016). APEC are typically part of the intestinal microbiota of healthy birds and infections typically result from environmental exposures and increased host susceptibility. More recently, APEC are also thought to be responsible for infections in humans, due to similarities in virulence determinants found in APEC and human ExPEC strains (Johnson *et al*., 2008). For instance, it has been demonstrated that genome content, virulence gene profiles, phylogeny, biofilm formation, and *in vivo* transcriptional activation are shared by APEC strains and the human ExPEC serotypes O18:K1:H7, O78, and O2:K1:H7 (Bauchart *et al*., 2010). Other studies of pathogenesis *in vivo* and *in vitro* have shown that APEC can cause disease in mammalian hosts and conversely, ExPEC isolated from human infections can cause disease in avian models (Jakobsen *et al*., 2012; Tivendale *et al*., 2010). These findings have led to the supposition that APEC is a zoonotic pathogen contributing to the weight of ExPEC infections in humans, particularly UTIs, and that consumption of retail poultry may be a source of infection (Platell *et al*., 2011b; Tivendale *et al*., 2010).

1.3.4.2. Neonatal meningitis *E. coli* (NMEC)

Neonatal meningitis *E. coli* (NMEC) have the ability to cross the blood-brain barrier (BBB) and are the second-leading cause of neonatal meningitis (Heath, Nik Yusoff and Baker, 2003), causing high mortality and neurologic sequelae in affected neonates. NMEC commonly inhabit the lower gastrointestinal tract, but become niche pathogens upon entry to the bloodstream and central

nervous system (CNS). NMEC are able to penetrate the BBB due to their ability to persist in the bloodstream by surviving engulfment by macrophages, multiplying in high numbers, and thus allowing successful invasion of the meninges of infants and causing meningitis (Kaper, Nataro and Mobley, 2004). NMEC are resistant to the host immune response, due to the possession of a KI capsule. This capsule is a thick polysialic acid layer that safeguards the bacterium from ingestion by phagocytic cells, and also from fusion with intracellular lytic vacuoles. As a consequence of the protection that the KI capsule provides, significant numbers of viable bacterial cells are transported across the BBB and into the CNS, where they can cause oedema and neural damage. It is most likely that these factors contribute to *E. coli*-associated neonatal meningitis carrying the significantly high mortality and morbidity rate (10–30%) associated with neonatal meningitis, caused by NMEC during the neonatal period (Kaper, Nataro and Mobley, 2004).

1.3.4.3. Uropathogenic *E. coli* (UPEC)

The pathotype of *E. coli* responsible for causing urinary tract infections (UTIs) is referred to as uropathogenic *E. coli* (UPEC). Among the common urinary pathogens associated with the development of UTIs, UPEC are the primary cause (Terlizzi, Gribaudo and Maffei, 2017). UPEC are defined by their ability to cause extraintestinal infections of the urinary tract (bladder, kidneys, ureter, and urethra), and are characterised by a plethora of both structural (fimbriae, pili, flagella, capsules) and secreted (toxins, iron-acquisition systems, proteins) virulence factors that contribute to their capacity to cause disease. UTIs are globally widespread and affect a large proportion of the human population. Approximately 150 million people worldwide develop a UTI each year (Flores-Mireles *et al*., 2015) and roughly 11% of women suffer an episode of UTI per year (Foxman, 2014). UPEC are thought to be responsible for up to 80% of uncomplicated UTIs in females (Flores-Mireles *et al*., 2015). UPEC are particularly well-adapted to surviving in the urinary tract and possess VAGs enabling them to scavenge iron from the environment and catabolise the amino acid D-serine, which is present in urine (Flores-Mireles *et al*., 2015). Additionally, UPEC are also notable for their ability to adhere to host epithelial cells in the urinary tract, and this represents the most important determinant of pathogenicity for UPEC. In severe cases where a UTI is left untreated, UPEC can ascend the urinary tract to cause infection of the kidneys and bloodstream.

UPEC strains that have invaded the bladder cells may be released and ascend to the kidneys via the ureters. Adherence to the kidney epithelial cells is mediated by binding of P-fimbriae to digalactoside receptors (Kaper, Nataro and Mobley, 2004). Upon colonisation of the kidney, UPEC virulence factors such as haemolysin and secreted autotransporter toxins (SAT) result in

damage to the renal epithelium, leading to pyelonephritis. Bacteria that have made it to this stage are then able to penetrate the endothelial cells of the proximal tubes and gain access to the bloodstream, resulting in bacteraemia (Kaper, Nataro and Mobley, 2004). The process of ascending infection from UTI to bacteraemia, involving the bladder, kidneys, and bloodstreams is known as urosepsis. The incidence of bloodstream infection is becoming more prevalent and is associated with higher rates of mortality. Bacteraemia caused by *E. coli* infection, which can be community or hospital-acquired, is reported with increasing frequency worldwide (Ron, 2010), accounting for 17–37% of bacteraemia cases globally (Russo and Johnson, 2003). Studies have reported that *E. coli* is the most frequent organism isolated from septicaemia resulting from an initial UTI. A UK-based study revealed that *E. coli* accounts for ~75% of Gram-negative bacteraemia cases of urinary origin (Al-Hasan, Eckel-Passow and Baddour, 2010). The association of bacteraemia caused by *E. coli* and UTI origin is therefore significant in the UK.

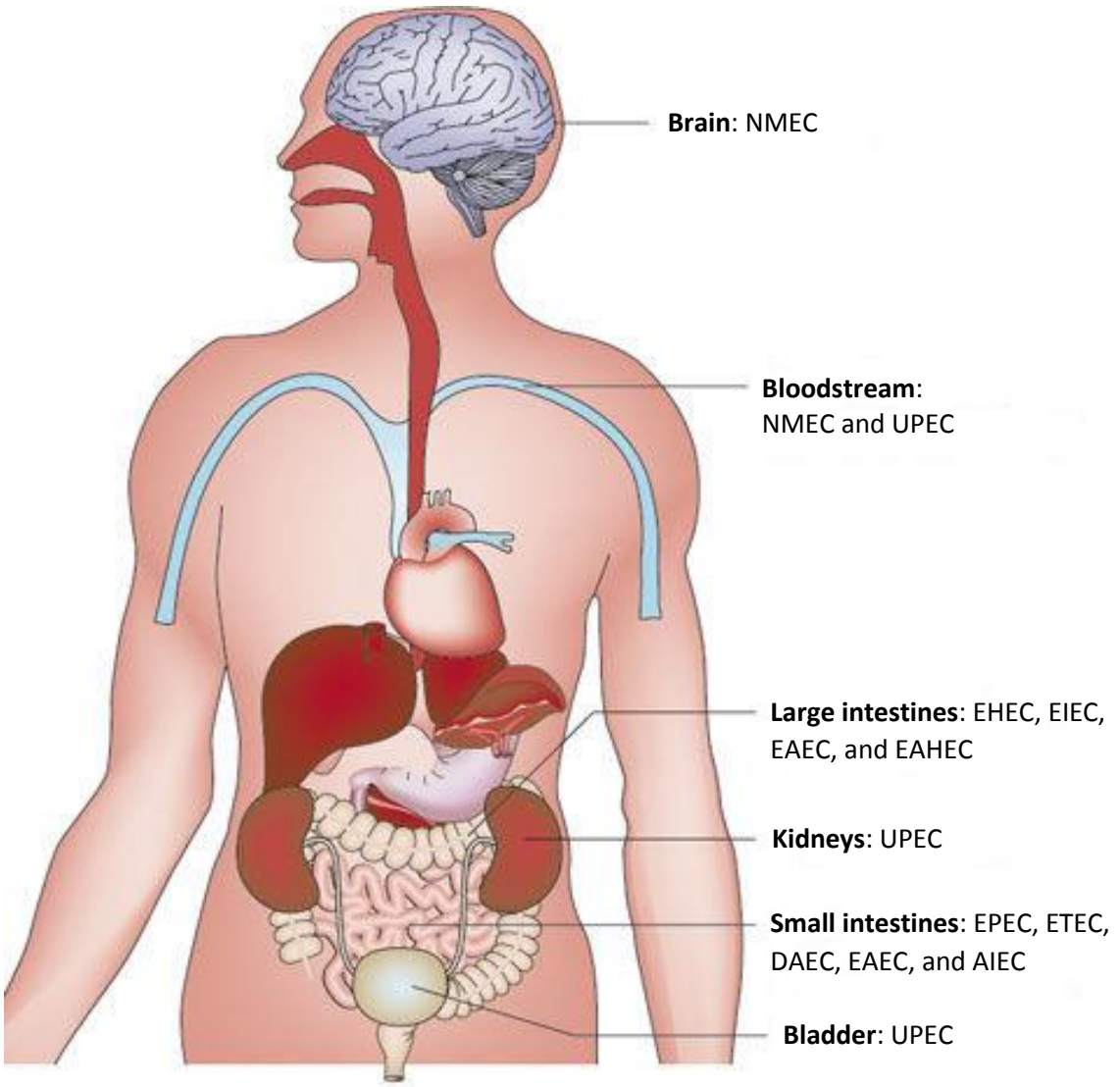


Figure 1.2. Sites of colonisation of the human body by pathogenic strains of *Escherichia coli*.

Pathogenic strains of *E. coli* are able to colonise various parts of the human body and cause subsequent disease, an ability that is attributed to the genome plasticity of the species and the carriage of specific virulence-associated genes (VAGs). The human extraintestinal pathogenic *E. coli* (ExPEC) develop infections at anatomical sites exterior to the gastrointestinal tract. Neonatal meningitis *E. coli* (NMEC) have the ability of crossing the blood-brain barrier (BBB) into the central nervous system and cause meningitis, whilst uropathogenic *E. coli* (UPEC) colonise and cause infection along various parts of the urinary tract, including the bladder and the kidneys, from which they can disseminate to the bloodstream and cause septicaemia. The intestinal pathogenic *E. coli* colonise various parts of the gastrointestinal tract. Enteroinvasive *E. coli* (EIEC), enteroaggregative *E. coli* (EAEC), and the particularly virulent entero-aggregative-haemorrhagic *E. coli* (EAHEC) and enterohaemorrhagic *E. coli* (EHEC) colonise the large intestines, whereas enteropathogenic *E. coli* (EPEC), enterotoxigenic *E. coli* (ETEC), diffusely adherent *E. coli* (DAEC), as well as adherent invasive *E. coli* (AIEC) colonise and cause infection in the small intestines. Enteroaggregative *E. coli* (EAEC) have the ability to cause disease in both the large and small intestines. Figure adapted from Croxen and Finlay (2010) to include the recently identified pathotypes, EAHEC and AIEC.

1.3.5. Antibiotic resistance in extraintestinal pathogenic *E. coli*

Antibiotic resistance in *E. coli* is a major problem facing today's healthcare systems, as it is a common cause of nosocomial infection and the leading causative agent of urinary tract infection and bacteraemia. The prevalence of antimicrobial resistance has been increasing in uropathogenic bacteria in recent times and the management of infections caused by these strains has been hampered as a result. The cephalosporins, fluoroquinolones, and trimethoprim–sulphamethoxazole drug classes have often been used as first-line antibiotics to treat community and hospital-acquired infections caused by *E. coli*. Increasing resistance to these agents is responsible for impediments to the appropriate therapy, resulting in more frequent cases of morbidity and mortality (Tumbarello *et al*., 2007). Up until the late 1990s, ExPEC were known to be relatively susceptible to first-line antibiotics, but surveillance studies during the 2000s in Europe and North and South America have shown that 20–45% of ExPEC are resistant to multiple drug classes, including the cephalosporins, fluoroquinolones, and trimethoprim-sulfamethoxazole (Foxman, 2014). β-Lactam antibiotics, in particular the penicillins and third-generation cephalosporins, are a major drug class used to treat serious *E. coli* infections acquired in the community or the hospital setting (Livermore and Woodford, 2006). Among *E. coli*, β-lactamase production remains the most important cause of β-lactam resistance. β-lactamases are usually plasmid-encoded enzymes produced by many Gram-negative bacteria, that inactivate β-lactam antibiotics by hydrolysing the β-lactam ring of the basic penicillin structure, thus deactivating the molecule's antibacterial properties (Livermore and Woodford, 2006). The first β-lactamases to be described were TEM (named after Temoneira, a patient in Greece from which TEM-1 was first isolated) in the 1960s, and later the SHV (SulfHydryl-Variable) β-lactamase. TEM and SHV β-lactamases are encoded by the bla_{TEM} and bla_{SHV} genes, respectively. These enzymes quickly became common in the hospital environment and allowed many genera of Gram-negative bacteria to become resistant to the commonly used antibiotics at the time.

The TEM-1 enzyme is the most commonly encountered β-lactamase in Gram-negative bacteria. It is thought that up to 90% of ampicillin resistance in *E. coli* is attributed to the production of TEM-1 (Livermore, 1995). Although TEM-type β-lactamases are most often found in *E. coli* and *K. pneumoniae*, they are also found in other species of Gram-negative bacteria with increasing frequency. TEM-1 has the ability to hydrolyse penicillins and early cephalosporins, such as cephalothin and cephaloridine (Paterson and Bonomo, 2005). The TEM-3 derivative, originally reported in 1989, was the first TEM-type β-lactamase that displayed the ESBL phenotype, but in the years since the first report, over 100 derivatives of TEM have been described (Paterson and Bonomo, 2005). SHV-1 shares 68% amino acid homology with TEM-1 and has an overall similar

structure. The SHV-1 β-lactamase is most commonly found in *K. pneumoniae* and is responsible for approximately 20% of the plasmid-mediated ampicillin resistance in this species. A *Klebsiella ozaenae* isolate was identified in Germany in 1983, which possessed a β-lactamase which efficiently hydrolysed cefotaxime, and to a lesser extent ceftazidime (Knothe *et al*., 1983). Sequence analysis of this isolate revealed that the β-lactamase differed from SHV-1, by replacement of glycine by serine at position 238. This point mutation accounts for the extended-spectrum properties of this β-lactamase, which was designated SHV-2 (Paterson and Bonomo, 2005). Within just 15 years of the discovery of this enzyme, organisms possessing the SHV-2 β-lactamase were found in every inhabited continent on Earth (Paterson *et al*., 2003). *E. coli* is also a typical clinical host of SHV-type enzymes, with previous studies reporting a high prevalence of *E. coli* producing SHV-type β-lactamases thought to be a significant cause of community-onset infections (Memariani *et al*., 2015).

Of importance with regards to ExPEC infections in the community and hospital environment is the increasing numbers of isolates developing resistance against newly developed antibiotics. These include the plasmid-mediated AmpC β-lactamases (e.g., CMY types), carbapenemases, and extended-spectrum β-lactamases (ESBLs). ESBLs are able to hydrolyse third and fourth generation cephalosporins, but are catalytically less efficient than the parent enzymes and are therefore susceptible to β-lactamase inhibitors, such as clavulanic acid (Paterson and Bonomo, 2005). To date, over 150 different types of ESBLs have been characterised, with a large proportion of ESBLs derived from point mutations in the parent β-lactamases, TEM-1, TEM-2, and SHV-1, which alter the amino acid configuration around the active site of these β-lactamases. This extends the spectrum of β-lactam antibiotics that can be susceptible to hydrolysis by these enzymes (Paterson and Bonomo, 2005). An increasing number of ESBLs that are not of the TEM or SHV lineage have recently been described, which carry tremendous clinical significance. The CTX-M type ESBLs are so named as they preferentially hydrolyse CefoTaXime and were first discovered in Munich in 1986 (Pitout *et al*., 2005). These enzymes, encoded by $bla_{\text{CTX-M}}$, are plasmid-mediated and are thought to have evolved separately from the TEM and SHV family, as they only have 40% sequence homology to these β-lactamases (Tzouvelekis *et al*., 2000). When these enzymes were first discovered they were named TOHO-l, but this was later changed to CTX-M (Peirano and Pitout, 2010). CTX-M β-lactamases have mainly been found in strains of *Salmonella enterica* serovar Typhimurium and *E. coli*, but have also been described in other species of Enterobacteriaceae (Bradford, 2001). It is thought that the extended spectrum activity of the CTX-M-type β-lactamases is attributed to the serine residue at position 237, which is present in all of the CTX-M enzymes (Tzouvelekis *et al*., 2000). Strains expressing CTX-M-type β-lactamases have been isolated from multiple locations around the world, but have most often

been associated with focal outbreaks (Bradford, 2001). Since the emergence of the CTX-M family of ESBLs in the UK in 2001-02, they have become a major concern in the healthcare setting, particularly CTX-M-15 and -14, due to their highly transmissible nature. The production of CTX-M-15 β-lactamases was first reported in an *E. coli* isolate in India in 2001 (Karim *et al*., 2001) and they have spread rapidly around the world, becoming an important cause of multidrug-resistant hospital- and community-acquired urinary tract infections (Peirano and Pitout, 2010). Consequently, CTX-M-15 is now the most globally widespread CTX-M enzyme, due to plasmid-associated dissemination coinciding with the emergence of a particularly successful *E. coli* clone, O25b:H4 ST131 (Peirano and Pitout, 2010). The impact of ESBL-producing *E. coli* has spread beyond the human-clinical setting. A wide range of ESBL-producing ExPEC isolated from domesticated animals has been reported, primarily of CTX-M-1 in chicken and CTX-M-14 in cattle, while CTX-M-15 prevails among companion animals (Ewers *et al*., 2014). Furthermore, CTX-M-producing strains of clinically-associated clonal groups of *E. coli* have been identified in surface waters (Gomi *et al*., 2017b) and retail poultry (Johnson *et al*., 2017; Leverstein-van Hall *et al*., 2011), suggesting that potential non-human reservoirs of MDR ExPEC may exist.

The OXA-type enzymes (oxacillinase, encoded by bla_{OXA}) are another evolving family of ESBLs. These β-lactamases differ from the TEM, SHV, and CTX-M families (class A) in that they belong to molecular class D and functional group 2d (Bradford, 2001). OXA-type ESBLs confer resistance to multiple antibiotics, including ampicillin and cephalothin, and are characterised by increased hydrolytic activity against oxacillin and cloxacillin, whilst being poorly inhibited by clavulanic acid. While most ESBLs have been found in species such as *E. coli*, *Klebsiella pneumoniae*, and other members of the Enterobacteriaceae, the OXA-type ESBLs are more commonly found in *Pseudomonas aeruginosa* (Bradford, 2001). The most common OXA-type ESBL is OXA-1, which has been identified in roughly 1–10% of *E. coli* isolates. The *bla*OXA-48 gene encodes a carbapenemase class D β-lactamase that was first identified in *K. pneumoniae* from Turkey in 2003 (Poirel *et al*., 2004), followed by subsequent reports of OXA-48 producers in *E. coli* from Israel, Senegal, and North Africa (Poirel *et al*., 2011a; Moquet *et al*., 2011; Cuzon *et al*., 2010). Further reports have indicated that OXA-48-producing *E. coli* have started to spread into the community in Europe (Poirel *et al*., 2011b), and can be an important cause of carbapenem resistance in extraintestinal infections. Aside from class D carbapenemases, there are also class A carbapenemases (mostly of the KPC type, produced by *K. pneumoniae*) and the metallo-β-lactamases of class B (such as NDM).

1.3.6. Source attribution

In nature, *E. coli* is principally a constituent of the gut microbiome of warm-blooded mammals, but it can also be found, albeit less frequently, in the gut microbiome of birds, reptiles and fish, as well as ubiquitously in the environment, in soil, water, plants and in food (Dublan Mde *et al*., 2014; Berthe *et al*., 2013; Platell *et al*., 2011b; Brennan *et al*., 2010; Tenaillon *et al*., 2010). Due to its ubiquity in nature and agriculture, there is a significant risk associated with transmission to humans if the appropriate control measures are not considered. *E. coli* is a cause of zoonoses – infections that can be transferred from animals to humans through a variety of mechanisms. Most commonly, human infections with *E. coli* occur due to poor hygiene, close contact with infected animals, or the consumption of contaminated food products (Frank *et al*., 2011). Although not all *E. coli* strains isolated from the environment are capable of causing disease in humans, it is likely that some pathogenic strains can find their way into the food chain (Fig. 1.3). *E. coli* is shed in the faeces of natural hosts into the environment, where it can survive in soil, water or on food for several days. The successful habitation of *E. coli* in these secondary environments is reliant on several key factors, such as the availability of nutrients and water, temperature, and acidity (van Elsas *et al*., 2011). Considering that *E. coli* is a predominantly intestinal inhabitant, presence of *E. coli* in food or water is an indicator of faecal contamination or poor hygienic practices.

Animals are a significant risk factor for the transmission of pathogenic *E. coli* to humans, due to the abundance of these microbes in the intestines of domesticated animals, such as sheep and ruminants. A previous study has shown that *E. coli* is well adapted to survive in the faeces of these animals, with an extinction range of 2–9 days, and has highlighted the potential risks associated with the contamination of food products by animal manure (Moriarty *et al*., 2010). Several outbreaks have been associated with contaminated or undercooked processed meat products such as beef burgers, sausages, and poultry (Vincent *et al*., 2010). Other studies have isolated *E. coli* from plants and seeds, suggesting that leaching from the soil leads to uptake of the organism into the roots where it can be disseminated to the leaves of plants such as lettuce and cabbage (Oliveira *et al*., 2012). Studies have suggested that the survival of *E. coli* in plants, which may be consumed raw or without the necessary washing or preparation, is of particular concern in the event of contamination with pathogenic *E. coli* capable of causing significant morbidity and mortality (Frank *et al*., 2011). Other studies have indicated *E. coli* outbreaks associated with contamination of animal-derived products such as unpasteurised milk and cheese (Gaulin *et al*., 2012). Studies have reported *E. coli* resembling the ExPEC strains responsible for human extraintestinal infection being recovered from waterways and retail poultry, leading to the suggestion that multiple non-human reservoirs for human ExPEC exist.

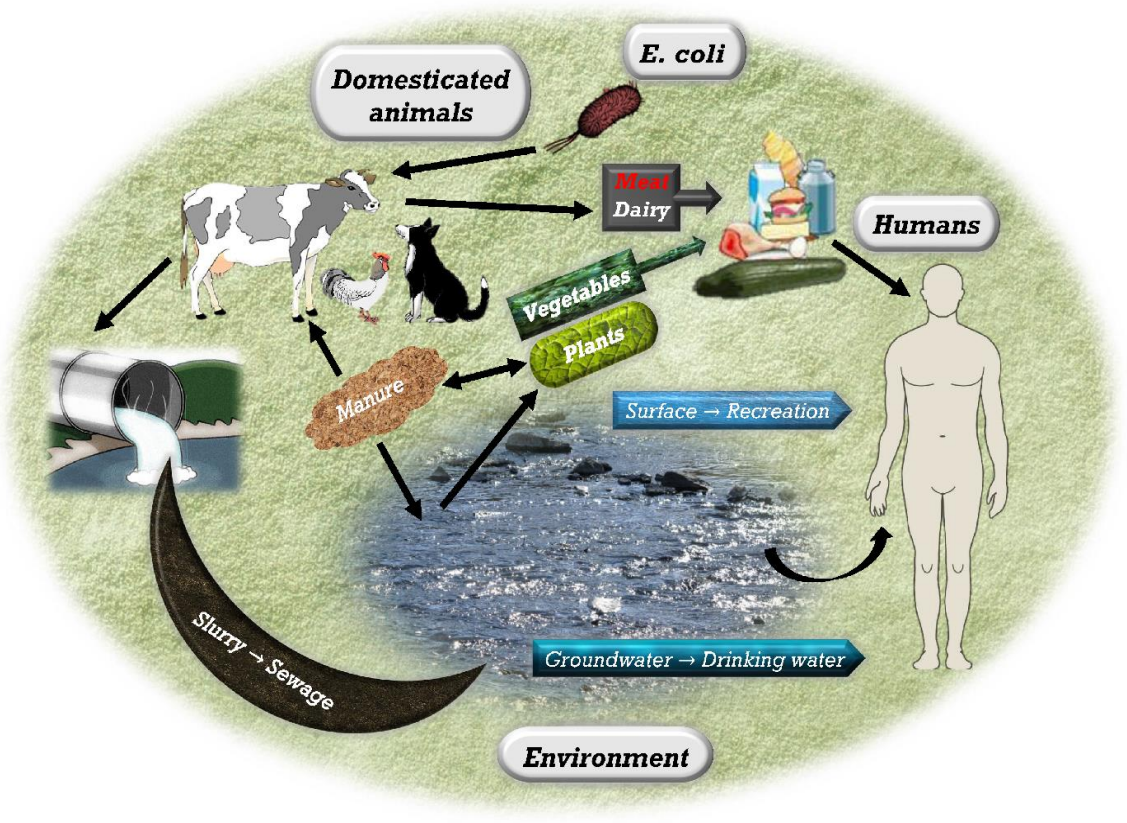


Figure 1.3. Illustration depicting the potential ecological habitat and routes of transmission of *Escherichia coli* in a global ecosystem.

E. coli is a natural constituent of the gut microbiome of warm-blooded mammals, including humans, but it can also be found in the gastrointestinal tracts of birds, reptiles, and fish. Non-pathogenic and pathogenic *E. coli* strains alike, when shed in the faeces of natural hosts into the environment, can survive in manure, soil, water, or on vegetation for several days. Furthermore, some *E. coli* in human faeces can survive the sewage treatment process and be discharged as effluents into natural waterways such as rivers and streams, and bodies of water such as lakes and the sea. If the appropriate control measures are not taken, there is a significant risk for transmission of *E. coli* to humans via the food chain, specifically through the consumption of contaminated water or food. The most common food sources for acquisition of *E. coli* would include fresh produce such as raw vegetables, dairy products such as unpasteurised milk, as well as undercooked ground beef and poultry that have become contaminated due to the slaughtering and food preparation processes. Furthermore, *E. coli* can be transmitted through direct contact with infected animals, as well as between humans, due to poor hygiene and sexual activity.

1.3.6.1. Surface waters as an environmental reservoir for ExPEC

An inevitable consequence of being an intestinal or extraintestinal microbe is to be regularly excreted into external environments. Environmental and urban waterways act as a passive carrier for coliforms such as *E. coli*, which is often used as an indicator of faecal contamination by water quality testing procedures. The major sources of faecal contamination in various watersheds include humans, agricultural and domesticated animals, and wild animals. Surface waters across the globe are contaminated with bacteria carrying antimicrobial resistance genes (Gomi *et al*., 2017b; Muller, Stephan and Nuesch-Inderbinen, 2016; Chen *et al*., 2016; Su *et al*., 2012; Dolejska *et al*., 2011b). Among these bacteria, *E. coli* are recognised as a contributor to the dissemination of antibiotic resistance genes in natural environments. Furthermore, freshwater environments are recognised as reactors for the evolution and dissemination of antibiotic resistance (Marti, Variatza and Balcazar, 2014). The major source of resistant bacteria and resistance genes is human sewage. Most developed countries around the world treat human sewage to reduce the bacterial load before releasing effluents into surrounding lakes, rivers, and oceans, or spreading it on land. However, previous studies indicate that treated sewage effluent, and the water into which it is released, remain heavily contaminated with antimicrobial-resistant bacteria (Gomi *et al*., 2017b; Dolejska *et al*., 2011b).

Recent environmental studies, reporting the presence of MDR and pathogenic ExPEC strains of *E. coli*, indicate that surface water is considered to be one of the important non-human reservoirs of MDR ExPEC (Gomi *et al*., 2017b; Muller, Stephan and Nuesch-Inderbinen, 2016). It has been suggested that contamination of surface waters by clinically important pathogenic clones of *E. coli* may increase the risk of contracting waterborne diseases (Gomi *et al*., 2017b). One of the factors which link the occurrence of ExPEC in surface waters with an increased risk of waterborne disease is the capability of *E. coli* strains to survive in open environments. Every aspect of the external environment, whether it concerns nutrition, temperature, oxygen, moisture, pH, and/or the surrounding microbial community, can vary drastically (Savageau, 1983). The ability to use nutrients and develop methods of overcoming these various stressors plays a crucial role in their survival in such environments. *E. coli* from livestock faeces is known to survive on grass pastures for 5 months or more, allowing the opportunity for pathogenic *E. coli* to be recycled by wild and domesticated animals (Avery, Moore and Hutchison, 2004), or to be introduced into aquatic environments following rainfall or irrigation. The presence of ExPEC in aquatic environments may appear to be a common denominator linking a diverse range of transitory habitats and transmission to animals and humans.

1.3.6.2. Retail poultry meat as a reservoir for ExPEC

With the widespread prevalence of *E. coli* in the environment it is possible that pathogenic strains may be introduced into the food chain. Environmental *E. coli* that resemble the ExPEC strains responsible for human extraintestinal infections have been identified in environmental and urban waterways, sewage, domesticated and wild animals, soil and other environmental samples, suggesting that various non-human reservoirs for human ExPEC may exist (Platell *et al*., 2011a; Ewers *et al*., 2010). It has also been demonstrated that human-to-human transmission of genetically nearly indistinguishable ExPEC occurs between household members and sexual contacts (Johnson and Clabots, 2006), indicating that humans are definitely a reservoir for ExPEC. The magnitude of the contribution of these various routes to ExPEC infection, however, is not known. Given that the foodborne route is arguably the major contributor to the transmission of enteric *E. coli* pathotypes, multiple studies have conducted investigations into food-borne transmission routes for human ExPEC (Muller, Stephan and Nuesch-Inderbinen, 2016; Bergeron *et al*., 2012; Platell *et al*., 2011b; Vincent *et al*., 2010).

A number of reports have suggested a high prevalence of ExPEC on retail chicken, beef, and pork meat, although recovery of ExPEC has evidently been highest from chicken meat (Jakobsen *et al*., 2010; Johnson *et al*., 2005a; 2005b). These studies have also suggested that consumption of contaminated poultry meat may play a role in human extraintestinal infections. The hypothesis that retail poultry meat products may provide a reservoir for human extraintestinal infection is based on several lines of evidence. Studies have identified genetic relationships between avian pathogenic *E. coli* (APEC) and human ExPEC (Zhao *et al*., 2009), and additionally, there are experimental studies showing the pathogenic potential of APEC in mammalian models and the pathogenic potential of human ExPEC in avian models (Jakobsen *et al*., 2012; Tivendale *et al*., 2010). Furthermore, close genetic relationships between *E. coli* isolates recovered from human extraintestinal infections, poultry, and retail chicken meat have been shown through molecular epidemiological data (Johnson *et al*., 2008; Moulin-Schouleur *et al*., 2006). *E. coli* ST131 and other pandemic ExPEC lineages (ST69, ST394, ST95, ST10 and ST117) have previously been identified in both human extraintestinal infections and in poultry reared for consumption or retail meat sources (Bergeron *et al*., 2012; Vincent *et al*., 2010). A study in Sweden revealed that *E. coli* sequence types ST69, ST117, and ST10 comprised 50% of the extended-spectrum β-lactamase (ESBL)-producing *E. coli* population recovered from domestic chicken meat. It was found that a substantial amount of chicken meat and chickens imported into Sweden, that were contaminated with ESBL-positive *E. coli*, had actually spread from imported parent broilers to broiler meat (Egervarn *et al*., 2014). This indicated that the occurrence of these antimicrobial-resistant ExPEC lineages on chicken meat was due to faecal contamination at slaughter. A study

from the Netherlands described four sets of *E. coli* isolates originating from human and poultry or retail chicken meat with indistinguishable ESBL gene types ($bla_{CTX\text{-}M\text{-}1}$ and $bla_{TEM\text{-}52}$), plasmids, and MLST genotypes (ST10, ST58, ST117 and ST10) (Leverstein-van Hall *et al*., 2011). The Netherlands study however, much like many previous studies, focussed primarily on antimicrobial-resistant ExPEC, specifically ESBL-producing *E. coli*, and therefore studies of ESBL-positive ExPEC lineages tend to be over-represented in the literature, and thus the true population structure of *E. coli* from non-human sources has not been accurately represented.

1.3.7. Population structure of extraintestinal pathogenic *E. coli*

1.3.7.1. *E. coli* phylo-typing

The first bacterial species for which population genomic techniques were described was *E. coli*. Through a technique known as multilocus enzyme electrophoresis (MLEE), it was revealed that certain combinations of alleles had appeared on multiple occasions, leading to the interpretation that *E. coli* is characterised by a clonal population structure with infrequent recombination (Milkman, 1973). Further support for this conclusion was achieved through subsequent MLEE analyses of thousands of natural and clinical isolates from humans and other hosts; 72 of these isolates, known as the *E. coli* Reference (ECOR) collection, were chosen to represent the genetic diversity of the species known at that time (Ochman and Selander, 1984). Based on the analysis of the ECOR collection, the *E. coli* species was traditionally split into four main phylogenetic groups: A, B1, B2, and D. Phylogroups B2 and D are largely associated with *E. coli* strains causing extraintestinal infections in humans, including those responsible for UTIs, sepsis, and neonatal meningitis. In contrast, phylogroups A and B1 are mainly associated with commensal and non-pathogenic strains of *E. coli* (Clermont, Bonacorsi and Bingen, 2000). Phylogenetic analysis of housekeeping gene sequences from the ECOR collection indicated that phylogroup D had diverged first, with groups A and B1 being sister groups that separated later (Nelson *et al*., 1997). Subsequent analyses suggest that perhaps phylogroup B2, rather than D, is the ancestral group (Escobar-Paramo *et al*., 2004). In 2000, Clermont and colleagues described a phylogrouping technique based on triplex PCR (Clermont, Bonacorsi and Bingen, 2000), designed to be a simple and rapid alternative to the traditional phylogrouping methods of MLEE and ribotyping, which are both complex and time-consuming techniques. The triplex PCR method is of importance in bio-clinical practice and as a biotechnological screening tool for elimination of potentially pathogenic strains, given the established link between phylogeny and virulence. With the growing body of MLST and phylogenetic analyses based on whole-genome sequence data for *E. coli*, our understanding of the phylogroup structure for this species has been refined in recent years. An improved phylo-typing method based on quadruplex PCR was

developed by Clermont and colleagues (2013), who described a total of eight recognised *E. coli* phylogroups. Seven of these phylogroups (A, B1, B2, C, D, E, and F) belong to *E. coli sensu stricto*, whereas the *Escherichia* cryptic clade C-I is considered to be the eighth *E. coli* phylogroup. Four other cryptic lineages of the *Escherichia* genus (C-II, C-III, C-IV, and C-V) have also been described, which are phylogenetically distinct but phenotypically indistinguishable from typical *E. coli* (Clermont *et al*., 2013). The current generation of enhanced phylo-typing and sequence typing methods has revealed that *E. coli* strains, even within a single pathotype, can vary immensely in terms of their evolutionary trajectory, which in turn can affect pathogenic potential and fitness of *E. coli* strains associated with human infection (Wirth *et al*., 2006).

1.3.7.2. ExPEC genotyping

Due to its reproducibility and comparability between different laboratories, MLST is considered to be the gold standard for ExPEC genotyping (Tartof *et al*., 2005). The seven housekeeping genes for *E. coli* are *adk*, *idh*, *fumC*, *mdh*, *purA*, *gyrB*, and *recA*. The Achtman scheme is the established MLST typing scheme for *E. coli* and the database is available via a publicly accessible website (Wirth *et al*., 2006). Analysis of MLST data can be used to track dissemination of pathogenic variants in epidemiological studies. A recent retrospective study by Kallonen and co-authors (Kallonen *et al*., 2017) analysed whole-genome sequence (WGS) data for 1509 *E. coli* isolates derived from the national British Society for Antimicrobial Chemotherapy (BSAC) collection (n = 1094) and a local collection from Cambridge University Hospital (n = 415). The combined collection comprised *E. coli* isolates associated with cases of bacteraemia between 2001 and 2012. The 1509 *E. coli* isolates were resolved into 228 unique STs. The most prevalent STs detected in the population were ST73 (17.3%), ST131 (14.4%), ST95 (10.6%), ST69 (5.5%), ST12 (4.6%), and ST10 (2.7%) (Kallonen *et al*., 2017). This study confirmed the findings of several genetic studies of *E. coli* lineages associated with UTIs and/or bacteraemia in England and the US, which reported that the most prevalent sequences types are ST131, ST73, ST95, and ST69 (Alhashash *et al*., 2013; Adams-Sapper *et al*., 2013; Gibreel *et al*., 2012). ST131, in particular, has received much scrutiny by investigators following its apparent emergence within the past two decades, due to its rapid dissemination across the globe and frequent multidrug-resistant phenotype.

In the mid-2000s, pulsed-field gel electrophoresis (PFGE) analyses were performed on two CTX-M-15 ESBL-producing *E. coli* strains that were associated with community- and hospital-onset UTI epidemics in the UK and Canada (Pitout *et al*., 2007; Woodford *et al*., 2004). Through the introduction of MLST typing methods, these strains were revealed to belong to a single sequence type, ST131, which were characterised as serotype O25b:H4 and phylogenetic group B2, as

determined by multilocus enzyme electrophoresis (MLEE) (Nicolas-Chanoine *et al*., 2008). Infections caused by ST131 are increasingly more frequent and are reported to be associated with increased morbidity and mortality (Johnson *et al*., 2010). Several studies have reported a high prevalence of ST131 isolates producing CTX-M-15 among ESBL producers and ST131 was significantly associated with fluoroquinolone resistance in a population of ESBL-negative strains (Johnson *et al*., 2010). ST131 is the leading *E. coli* sequence type causing antibiotic-resistant and MDR urinary tract infections in several countries in Europe and Asia, as well as Canada and Australia (Kallonen *et al*., 2017; Peirano and Pitout, 2010). *E. coli* ST131 are prevalent in both community- and hospital-acquired infections, but the source of infection is not well characterised. The potential of companion and domesticated animals to transmit this pathogen has previously been highlighted (McNally *et al*., 2016a), as has an association with the food chain (Platell *et al*., 2011b) and the environment (Gomi *et al*., 2017b).

ST69 strains have been isolated worldwide from cases of UTI and bacteraemia from both community-onset and hospital-acquired infections (Kallonen *et al*., 2017; Alhashash *et al*., 2013; Croxall *et al*., 2011b). Most ST69 strains express an MDR phenotype, although they do not commonly produce ESBLs (Ajiboye *et al*., 2009). In a hospital-based study of *E. coli* bacteraemia isolates collected in San Francisco between 2007 and 2010, the ST69 complex was found to be the fourth most prevalent ExPEC ST after ST131, ST73, ST95 (Adams-Sapper *et al*., 2013), parallel to a recent UK-based study on *E. coli* bacteraemia isolates (Kallonen *et al*., 2017). ST69 strains were obtained less than 48 hours after admission from 83% of all cases, which may suggest that ST69 is a clonal ExPEC group circulating predominantly in the community as opposed to hospital settings. It is also thought that there may be a non-human reservoir for *E. coli* ST69, with isolates having been recovered from retail meats, domesticated animals, and the environment.

E. coli ST95 strains belong to phylogenetic group B2 and comprise K1 capsular serotypes (O1:K1:H7, O2:K1:H7, and O18:K1:H7) that are traditionally linked to neonatal meningitis (Tivendale *et al*., 2010). The ST95 lineage also includes avian pathogenic *E. coli* strains which are responsible for colibacillosis in wild and domesticated birds (Mora *et al*., 2009). ST95 isolates were identified as the second most prevalent clonal group isolated from bacteraemia ExPEC infections in the United States (Adams-Sapper *et al*., 2013), and the third most common ST among bacteraemia isolates from UK-based studies (Kallonen *et al*., 2017; Alhashash *et al*., 2013). In a French hospital-based study, the ST95 complex was the most common *E. coli* lineage isolated from blood and ascitic fluid cultures between 1997 and 2006 (Bert *et al*., 2010). One noticeable distinction of ST95 strains is that they are typically characterised by a low frequency of multidrug resistance. More than half of all ST95 isolates obtained from a San Francisco hospital-based study were susceptible to all antibiotics tested, demonstrating significantly less

resistance compared to ST131 isolates obtained from the same study. (Adams-Sapper *et al*., 2013). In addition to animal hosts, *E. coli* ST95 is also commonly reported by environmental studies, and the ST95 complex was identified as the most prevalent clinically important clonal group (31%) among ExPEC isolates (*n* = 58) obtained from river water in Japan (Gomi *et al*., 2017).

E. coli ST73 strains belong to phylogenetic group B2 and are only associated with serotype O6:H1 (Johnson *et al*., 2008). In the UK, ST73 was identified as the most commonly encountered ST among major ExPEC clonal groups isolated from cases of bacteraemia (Kallonen *et al*., 2017; Alhashash *et al*., 2013) and urinary tract infection (Gibreel *et al*., 2012), suggesting that the *E. coli* strain O6:H1-B2-ST73 is a leading cause of human extraintestinal infection in this country. Other UK-based studies have reported a high prevalence of ST73 strains associated with both community-onset and hospital-acquired infections (Croxall *et al*., 2011b). ST73 has also been reported to be one of the predominant ExPEC STs associated with expression of the CTX-M-15 ESBL (Gibreel *et al*., 2012). ST73 strains with closely related PFGE types have been isolated from humans, dogs, and cats, suggesting cross-species transmission of this clone (Johnson *et al*., 2008). It may be the case that ST73 represents a long-standing, human-adapted ExPEC clonal group as it has been responsible for causing UTIs in women from widely separated geographic locales over a considerable period of time (Manges *et al*., 2008).

E. coli ST10 and closely related STs of the ST10 clonal complex belong to phylogenetic group A, which is typically associated with commensal colonisation of the human gastrointestinal tract. There are numerous serotypes associated with *E. coli* ST10. Although the ST10 clonal complex is commonly encountered as a human intestinal coloniser of low virulence and low antimicrobial resistance, it has also been associated with human infections and ESBL production (hospital- and community-acquired infections), meat products, and food animals (Peirano *et al*., 2012; Cohen Stuart *et al*., 2012). *E. coli* ST10 are also widespread in the environment and are commonly identified in surface waters (Gomi *et al*., 2017b; Jorgensen *et al*., 2017). In a study carried out in the Netherlands, CTX-M-producing ST10 strains were isolated from human blood cultures and poultry, and TEM-producing ST10 isolates were recovered from human urine samples as well as poultry (Leverstein-van Hall *et al*., 2011). Furthermore, a study from Canada identified multidrug-resistant *E. coli* ST10 strains from human-clinical samples, chicken faeces, retail chicken meat, pig faeces, and pork meat, indicating a strong association of this genotype with the food chain (Bergeron *et al*., 2012).

1.4. Aims and objectives

The primary aim of this study was to carry out population genomic analyses to uncover novel information on the ecology of *Y. pseudotuberculosis* and *E. coli*. Comprehensive genome-scale analyses focussed on the enteric pathogen *Y. enterocolitica* suggested that distinct phylogroups of the species may be ecologically separated, through an exhibition of restricted genetic exchange between phylogroups. The dearth of such large-scale population genomic studies for *Y. pseudotuberculosis* means the ecology of this organism is not fully understood. To investigate whether similar hidden ecological patterns can be uncovered for this model organism, this study offers a high-resolution contribution to the understanding of the ecology, evolution, and dissemination of this important human pathogen. A large data set of globally and temporally distributed *Y. pseudotuberculosis* genomes from multiple ecosystems were analysed in this study. To identify any genomic signatures associated with the ecology of the *Y. pseudotuberculosis* population, core genome phylogenetic analysis was complemented with a pan-genome approach to effectively compare the entire gene contents of multiple genomes across the population.

In addition to *Y. pseudotuberculosis, E. coli* was also investigated in this study. There are many reports in the literature of non-human *E. coli* resembling the ExPEC strains responsible for human extraintestinal infection that have been recovered from environmental and food sources, particularly river water and retail chicken meat. This has led to the suggestion that there may be several non-human reservoirs for human multidrug-resistant ExPEC. The majority of these studies, however, selectively enrich for antimicrobial-resistant isolates, and thus, reports of ESBL-producing *E. coli* tend to be overrepresented in the literature. The relative abundance of MDR *E. coli* and ExPEC strains in the wider non-human population of *E. coli* is therefore largely unknown. In this study, we sought to determine the true population structure of non-human *E. coli* from river water and retail chicken by taking an unbiased culture-based approach to sampling and not selectively cultivating resistant isolates. This strategy was combined with whole-genome sequencing of single isolates to allow comparative genomic analyses to be performed between the non-human population and human-clinical isolates of *E. coli*, previously obtained from the same region. A pan-genome approach was applied to the two populations, providing high-resolution genomic comparison to determine the extent of genetic overlap between the non-human and human-clinical populations and whether any genetic overlap exists between the non-human and human-clinical populations of *E. coli* in Nottingham, and whether non-human sources of *E. coli* are likely to contribute to the weight of extraintestinal infections in this region.

The primary objectives of the study were to:

- Reconstruct the phylogeny of a globally dispersed population of *Y. pseudotuberculosis* from multiple ecological niches and identify any genomic signatures associated with the ecology of the *Y. pseudotuberculosis* population.
- Investigate the evolutionary history of *Y. pseudotuberculosis* by performing dating analysis to date the phylogeny of the species.
- Determine whether any patterns of gene flow exist within the *Y. pseudotuberculosis* population by analysing the core and accessory genomes of these strains.
- Isolate a population of *E. coli* from non-human (river water and retail chicken meat) samples and sequence the genomes of a non-biased representative proportion of the population, to generate a snapshot of the population structure of non-human *E. coli*, as determined by *in silico* multilocus sequence typing and phylogenetic analyses.
- Detect antimicrobial resistance genes and virulence-associated genes in the non-human population of *E. coli* to create a snapshot of the prevalence of potentially multidrug-resistant strains as well as human ExPEC strains.
- Compare the population structures of non-human and human-clinical *E. coli* isolated from the same region, with regards to phylogeny and the prevalence of clinically important clonal groups, antimicrobial resistance determinants, and human ExPEC strains in both populations.
- Use phylogenetic analysis to situate representative non-human strains of clinically important clonal groups within the wider populations of those clones obtained from multiple hosts.
- Identify genomic signatures of ecological separation by performing comparative genomic analysis of all non-human and human-clinical strains of *E. coli*, using a pan-genome approach.
- Determine the extent of gene movement between closely related strains of the human-clinical and non-human populations of *E. coli*, by comparing the pan-genomes and detected core genome recombination events between strains of clinically important clonal groups present in both populations.

CHAPTER 2

Materials and methods

2.1. Media and reagents

2.1.1. Growth and storage media

Cystine Lactose Electrolyte Deficient (CLED, with Andrade's indicator) agar was prepared with 36.2 g of CLED agar powder (CM0423, Oxoid, Basingstoke, UK) per 1 L of distilled water.

HiCrome™ UTI agar was prepared with 55.4 g of HiCrome™ UTI agar powder (16636, Sigma-Aldrich, Dorset, UK) per 1 L of distilled water.

LB agar was prepared with 40 g of LB agar powder, Miller (tryptone 10 g/L, yeast extract 5 g/L, sodium chloride 10 g/L; BP1425 Fisher BioReagents, Loughborough, UK) per 1 L of distilled water.

Tryptone Soya agar (TSA) was prepared with 40 g of TSA powder (CM0131, Oxoid, Basingstoke, UK) per 1 L of distilled water. All agar solutions were sterilised by autoclaving at 121 °C for 15 minutes and then cooled to 50 °C before pouring into sterile Petri dishes.

Lysogeny broth (LB) was prepared with 40 g of LB broth powder (BP1426, Fisher BioReagents, Loughborough, UK) per 1 L of distilled water.

Buffered peptone water, a broth used for the culture of organisms for detecting indole production with Kovac's reagent, was prepared with 20 g of buffered peptone water powder (CM0509, Oxoid, Basingstoke, UK) per 1 L of distilled water. All broth solutions were sterilised by autoclaving at 121 °C for 15 minutes.

A broth used for the storage of bacterial cultures at –80 °C was prepared in a 1 mL cryotube vial by suspending a single colony of bacterial culture from a CLED agar plate in 800 µL of LB broth with 200 µL (20% v/v) of glycerol (Sigma-Aldrich).

2.1.2. API identification kit

Identification of bacterial species was carried out using the API 20 E (20100, BioMérieux, Marcy-l'Etoile, France) identification system for Enterobacteriaceae and other Gram-negative rods, by following the manufacturer's protocol. The API Reagent Kit (20120) was used for tests that require the addition of reagents when reading and interpreting the test strip. API Suspension Medium (20150) was used for the preparation of the inoculum for the API test strip. The McFarland Standard kit (70900) was used to produce a standard inoculum for API testing (0.5 McFarland standard). Mineral oil (70100) was required to produce anaerobic test conditions for certain biochemical reactions within the API test strip. The api*web* identification database

(40011, BioMérieux, Marcy-l'Etoile, France) was used to interpret the results and identify the species of the test organism.

2.1.3. Molecular microbiology reagents

Genomic DNA (gDNA) was prepared using the GenElute™ Bacterial Genomic DNA Kit (NA211O, Sigma Aldrich, Dorset, UK), as per the manufacturer's instructions. Polymerase chain reaction (PCR) amplification assays were performed using the GoTaq® Flexi DNA Polymerase kit (M8306, Promega, Southampton, UK). Agarose gels were prepared using Agarose Molecular Biology grade powder (10766834, Fisher Scientific, Loughborough, UK), 50x Tris-acetate-EDTA (TAE) buffer (EC-872, National Diagnostics supplied by Fisher Scientific, Loughborough, UK), and SYBR™ Safe DNA gel stain (S33102, Invitrogen, Renfrew, UK). PCR amplicons were electrophoresed on each gel with a 100 bp DNA Ladder (N3231, New England Biolabs, Hitchin, UK).

2.1.4. NGS reagents and sequencing kits

Qubit™ dsDNA BR Assay Kit
Q32850, Invitrogen, Renfrew, UK

Qubit™ dsDNA HS Assay Kit
Q32854, Invitrogen, Renfrew, UK

High Sensitivity D1000 ScreenTape
5067-5584, Agilent Technologies, Stockport, UK

High Sensitivity D1000 Reagents
5067-5585, Agilent Technologies, Stockport, UK

Agencourt AMPure XP
A63881, Beckman Coulter, High Wycombe, UK

Nextera XT DNA Library Prep Kit
FC-131-1024, Illumina, Cambridge, UK

Nextera XT DNA Library Index Kit v2 Set A
FC-131-2001, Illumina, Cambridge, UK

MiSeq Reagent Kit v2 (500 cycles)
MS-102-2003, Illumina, Cambridge, UK

PhiX Control Kit v3

FC-110-3001, Illumina, Cambridge, UK

Sodium Hydroxide 10 M

10488790, Fisher Scientific, Loughborough, UK

2.1.5. Buffers and reagents

0.85% saline solution was prepared by dissolving 1 saline tablet (BR0053, Oxoid, Basingstoke, UK) in 500 mL of distilled water. The solution was then sterilised by autoclaving at 121 °C for 15 minutes to obtain 0.85% ('normal', physiological, or isotonic) saline solution.

1M Tris-HCl (pH 8.0) was prepared with 12.1 g of Tris base (BPE 152-1, Fisher Scientific, Loughborough, UK) per 80 mL of distilled water. The pH was adjusted to 8.0 using 1M HCl. The final volume of the solution was adjusted to 1 L using distilled water. The solution was sterilised by autoclaving at 121 °C for 15 minutes and any final pH adjustments were made once the solution cooled to room temperature.

Oxidase discs (70439, Sigma-Aldrich, Dorset, UK) were used to detect oxidase-producing organisms.

Kovac's reagent for indoles (60983, Sigma-Aldrich, Dorset, UK) was used to detect indole-producing organisms.

2.2. *Y. pseudotuberculosis* phylogenomics

2.2.1. Determining phylogenetic relationships

A core genome alignment of the strains was constructed from localised co-linear blocks using the Parsnp tool (v1.2) from the Harvest suite (Treangen *et al*., 2014). Parsnp was run with default parameters, using the following command-line `parsnp -r ! -d <genome_dir> -c` where the (r)eference genome was set to '!' to pick a random reference from the genome directory, the path was specified to the (d)irectory of genomes, and the (c)urated genome directory flag was used to force inclusion of all genomes in the directory. Parsnp takes the directory of FASTA files to be aligned and generates a maximum-likelihood phylogenetic tree, reconstructed from the alignment, based on core genome SNP analysis. Metadata encompassing information on isolation (continent, country and host), serotype, and CRISPR motif for each strain were

superimposed on the tree as coloured bars, using the Interactive Tree of Life (iTOL) web-based tool (Letunic and Bork, 2016).

2.2.2. Analysis of CRISPR loci

Katja Koskela (University of Helsinki) had searched the genome assemblies for CRISPR loci with BLASTN, using the *Y. pseudotuberculosis*–specific CRISPR direct repeat sequence (5'-TTTCTAAGCTGCCTGTGCGGCAGTGAAC-3'), its complementary sequence, the 5'– and 3'–flanking sequences of the YP1, YP2 and YP3 loci, and their complementary sequences (Koskela *et al*., 2015). Identified sequences were submitted to the CRISPRFinder tool on the CRISPRs Web Server (http://crispr.i2bc.paris-saclay.fr/Server/), together with the spacer dictionary compiled earlier (Koskela *et al*., 2015). This analysis increased the number of identified spacers in the *Y. pseudotuberculosis* spacer dictionary from 1902 to 2969. The complete list of strains and spacer arrays used for CRISPR spacer clustering is available from doi: http://dx.doi.org/10.1099/mgen.0.000133 as supplementary material (Seecharran *et al*., 2017).

2.2.3. Pan-genome and accessory genome analyses

The Large-Scale Blast Score Ratio (LS-BSR) v3.0 pipeline (Sahl *et al*., 2014) was used to create pan-genomes from genome assemblies of all strains. The post-matrix script (filter_BSR_variome.py) was run to isolate the accessory genomes from the pan-genomes. The resulting accessory genome matrix was then transposed according to the order of the strains on the phylogenetic tree. The output was used to visualise the presence or absence of all accessory genes in each individual genome by generating a heat map using the ggplot2 package of the R statistical software v3.2.0 (http://www.r-project.org/; R Core Team, 2015; Wickham, 2016). Genes with > 90% prevalence, and also those found in fewer than 5 strains, were excluded from this analysis. The Python script compare_BSR.py from LS-BSR was used to look for unique coding sequences (CDSs) between two defined populations in the pan-genome matrix. Comparisons were made between the 'European' clade of strains and the 'Asian' clade, as well as between each CRISPR cluster and the rest of the population. Any unique CDSs detected were compared to the non-redundant nucleotide database using nucleotide BLAST with default parameters (http://blast.ncbi.nlm.nih.gov/) to determine the genes they encode.

The following analysis was performed by Jukka Corander (University of Oslo). KPAX2 software was used to cluster the strains based on their CRISPR spacer profiles (Pessia *et al*., 2015), resulting in 33 identified 'CRISPR cluster' labels. Input to the software was a binary matrix with

columns representing an absence/presence variable for each of the 2,969 spacers in each detected CRISPR cassette. KPAX2 was used with default prior hyperparameters and an upper bound for the number of clusters equal to 50. Five independent runs of the inference algorithm were performed and the clustering solution with the highest posterior probability was chosen. All estimation runs converged to a number of clusters well below the chosen upper bound, indicating that it was sufficiently large to accommodate the region of high posterior density. To analyse the association between CRISPR spacer patterns and the accessory genome content, an average accessory genome dissimilarity (Hamming distance normalised by the number of CRISPR spacers) matrix was calculated for all detected CRISPR clusters with > 1 strain (18 clusters).

To assess the significance of the observed dissimilarity pattern, a standard permutation test was performed by Jukka Corander. Under the null hypothesis of no association between CRISPR clusters and accessory genome content, the cluster label of a strain can be permuted randomly. For each of 10,000 random permutations of the label, the average dissimilarity for each cluster was recalculated, and it was recorded how often the observed value is smaller than the observed dissimilarity in the original data matrix. Under the global significance level of 0.05, 12 out of 18 CRISPR clusters had a significantly smaller average distance than expected under the null hypothesis.

2.2.4. Detection of core genome recombination events

Core genome alignments were constructed using Parsnp (Treangen *et al*., 2014), as previously described in section 2.2.1. Core genome recombination events were detected by running the software package BratNextGen (v1.0) (Marttinen *et al*., 2012) on the core genome alignment. BratNextGen was run by Alan McNally (University of Birmingham), using the default prior settings, 20 iterations of the HMM estimation algorithm and 100 runs executed in parallel for the permutation test of significance.

2.2.5. Dating analysis

Bayesian Evolutionary Analysis by Sampling Trees (BEAST 2, v2.4.0) (Bouckaert *et al*., 2014) was used to date the phylogeographic split within the species, and the formation of the distinct CRISPR clusters. Of the 134 genomes sequenced, isolation dates were available for 46 strains which represent the full diversity of the phylogeny. A core genome alignment of the 46 strains, constructed using Parsnp (Treangen *et al*., 2014), was stripped of recombination detected using BratNextGen (Marttinen *et al*., 2012), and the resulting alignment was used as input for BEAST 2 with all known dates of isolation to date individual taxa. BEAST 2 was run by Alan McNally

(University of Birmingham). By assessing ESS (effective sample size) scores for priors, the following parameters were chosen for the best fitting model: HKY model of substitution with estimated base frequencies and a relaxed molecular clock. The analysis was run for a total of 50 million iterations with the initial 5 million used as burn-in. From this, a maximum clade credibility tree was produced and visualised in Figtree (http://en.bio-soft.net/tree/figtree.html). For the Skyline analysis, a stepwise constant variant was selected with the age of youngest tip set to zero.

2.3. Non-human *E. coli* strain collection

2.3.1 River water sampling

Nine water samples were collected by Jody Winter (Nottingham Trent University) in July 2015. These samples were taken from different sites along rivers/streams and wetlands at 4 geographically distinct locations within the Trent River basin, in Nottinghamshire and Derbyshire:

1. **Giltbrook** – Two samples were taken from the *Gilt Brook* near Giltbrook, Nottinghamshire (Fig. 2.1); one sample upstream and one sample downstream of Severn Trent Water Ltd waste water and sewage treatment plant.
2. **Erewash Pinxton** – Two samples were obtained from the *River Erewash* near Pinxton, Derbyshire (Fig. 2.2); one sample upstream and one sample downstream of Amber Valley Water Services wastewater treatment plant.
3. **East Leake** – Two samples were extracted from the *Kingston Brook* near East Leake, Nottinghamshire (Fig. 2.3); one sample upstream and one sample downstream of Brook Furlong Farm.
4. **Keyworth** – Three samples were collected from a tributary of the *Fairham Brook* near Keyworth, Nottinghamshire (Fig. 2.4); one sample upstream and one sample downstream of Hillside Farm, and one sample from the wetland area near a cattle field.

The samples were collected in sterile universal containers at a depth of 0.5 m and were transported to the laboratory on the day of collection. The samples were then stored at 4 °C and microbiological cultivation was carried out within 24 hours of collection.

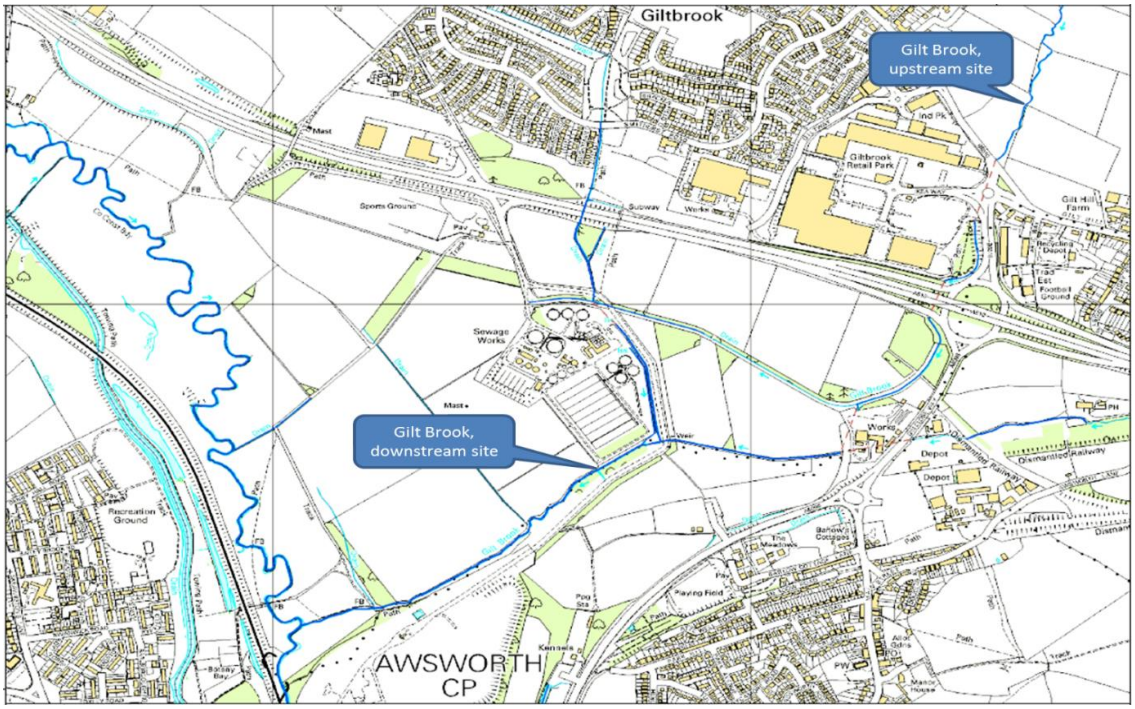


Figure 2.1. Sample sites upstream and downstream of Severn Trent Water Ltd wastewater treatment plant, located on the *Gilt Brook* near Giltbrook, Nottinghamshire.

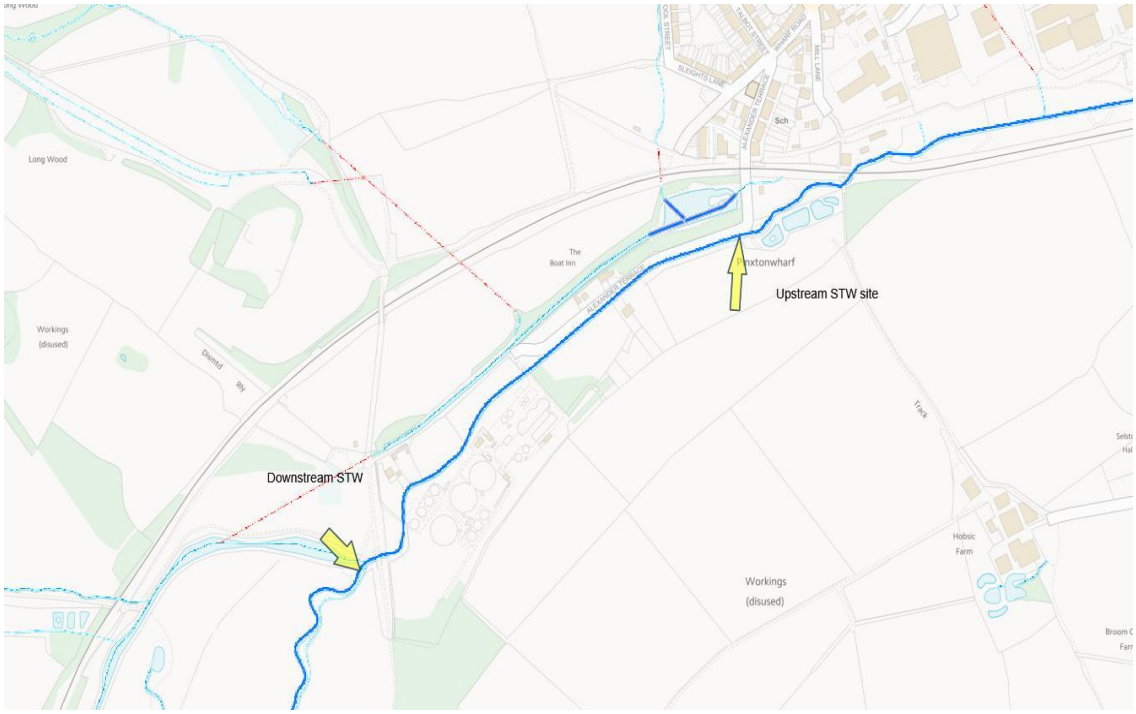


Figure 2.2. Sample sites upstream and downstream of Amber Valley Water Services wastewater treatment plant, located on the *River Erewash* near Pinxton, Derbyshire.

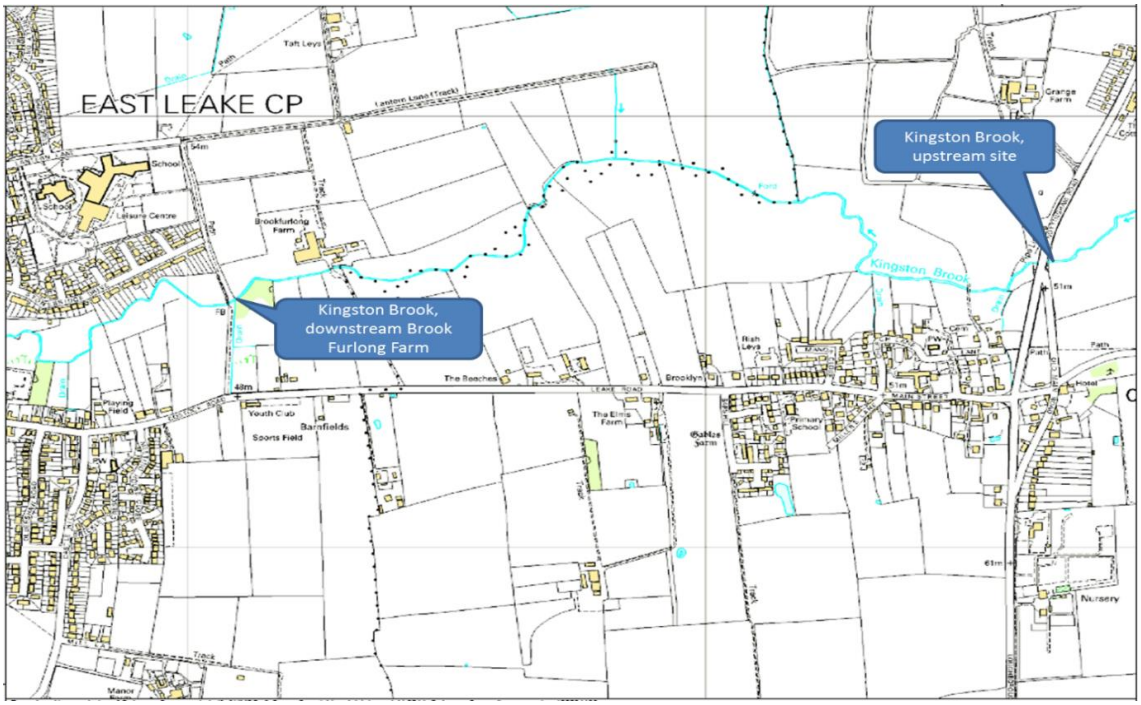


Figure 2.3. Sample sites upstream and downstream of Brook Furlong Farm, located on the *Kingston Brook* near East Leake, Nottinghamshire.

Figure 2.4. Sample sites upstream and downstream of Hillside Farm and sample site of wetland area near cattle field, located on the *Fairham Brook* near Keyworth, Nottinghamshire.

2.3.2. Retail poultry sampling

A total of 20 whole chickens, weighing approximately 1.4–1.6 kg, were obtained from 6 major retail outlets in the Greater Nottingham area, in October 2015 (Table 2.1). All selected chickens were reared in the UK and represent both caged and free-range chickens. The chicken samples were kept in their original packaging and transported to the laboratory where they were stored at 4 °C and processed within 2 hours, following a previously published protocol (Asensi *et al*., 2009). Each whole chicken was hand-rinsed, under aseptic conditions, with 100 mL of 0.85% saline solution (Oxoid) in a sterile plastic bag, for 3 minutes. The chicken rinse solution was then immediately subjected to microbiological cultivation.

Table 2.1. Numbers of whole chickens obtained from 6 major supermarket chains and a snapshot of chicken processing companies in the UK.

		Poultry processing companies				
Supermarket	**Number of chickens**	**Faccenda Group**	**Moy Park Ltd**	**2 Sisters Food Group Ltd**	**Cargill PLC**	**Banham Poultry Ltd**
Tesco PLC	4		✓	✓	✓	
Sainsbury's	3		✓	✓		
Asda Stores Ltd	3	✓				
Iceland Food Ltd	4				✓	✓
Morrisons	4			✓		✓
Aldi	2			✓		✓

The majority of all retailers source their fresh chicken from the same 4 processing companies that predominate the farming and abattoir business, which include the largest supplier in the UK, 2 Sisters Food Group Ltd. Asda Stores Ltd is the only major supermarket chain to source fresh chicken from Faccenda Group.

2.3.3. Isolation and identification of *E. coli* from non-human samples

Bacteria were cultivated from river water and chicken rinse samples using standard microbiological techniques as follows. A total of 200 µL of each sample was transferred to CLED Agar with Andrade's indicator (Oxoid) and inoculated using the spread plate technique. The plates were incubated at 37 °C for 18–24 hours.

After incubation, the morphologies of single colonies present on CLED agar plates were recorded and an initial presumptive identification was made based on the characteristics described in Table 2.2. A control strain (*E. coli* UTI18, obtained from the NTU Pathogen Research Group culture collection) was used to test the performance of CLED agar to identify potential *E. coli*. All colonies that were presumptively identified as *E. coli* were subsequently sub-cultured onto HiCrome™ UTI Agar (Sigma-Aldrich), which is a chromogenic differential medium for identification and differentiation of microorganisms mainly causing urinary tract infections, including Enterobacteriaceae. The plates were then incubated at 37 °C for 18–24 hours and colony morphologies were recorded. This additional culture step allowed for purification to single colonies and, as a result, isolates that presented with two or more different morphologies were regarded as a 'mixed culture' and were discarded from further testing.

Pure colonies that exhibited characteristic *E. coli* morphology on HiCrome™ UTI Agar, as described in Table 2.2, were selected for further biochemical testing, which involved performing the oxidase test and the indole test. The oxidase test was performed to support the presumptive identification of *E. coli*, which are oxidase-negative organisms. A plastic inoculation loop was used to transfer a distinct colony from a fresh culture plate (less than 24 hours old) onto an oxidase test disc containing N',N'-dimethyl-p-phenylenediamine (Sigma-Aldrich). A negative reaction was observed, indicated by an absence of colouration after the test was performed. For the indole test, a single colony was inoculated into 10 mL of Buffered Peptone Water (Oxoid) and incubated at 37 °C for 18–24 hours, with shaking (200 rpm). After incubation, 200 µL of Kovac's reagent (Sigma-Aldrich) was added to the culture broth. If indole was present, the surface reagent layer turned red after 1–2 minutes, indicating an indole-positive organism, such as *E. coli*. A negative result was indicated by the surface reagent layer remaining yellow or yellow–orange in colour. *E. coli* strain UTI18 (NTU Pathogen Research Group strain collection) was used a control strain for both the oxidase and indole tests.

Once oxidase-negative and indole-positive isolates had been ascertained, a single colony of each isolate was sub-cultured onto LB Agar (Fisher BioReagents) for purity and incubated at 37 °C for 18–24 hours. Biochemical identification testing was performed from the purity plate using the Analytical Profile Index (API) 20 E identification system (BioMérieux, France). This system was

developed for the identification of members of the Enterobacteriaceae family of bacteria. Bacterial isolates can be identified to the species level based on a profile of 20 biochemical reactions, which can be compared to a vast number of taxa on the regularly updated API database. For the API 20 E protocol, a bacterial suspension was prepared from the pure culture plate by inoculating 5 mL of API Suspension Medium (BioMérieux) with a single colony. The bacterial suspension was standardised according to the turbidity of a 0.5 McFarland Standard (BioMérieux) and distributed into each cupule on the test strip, rehydrating the biochemical substrates that are freeze-dried at the bottom. Some biochemical tests required an overlay of mineral oil (BioMérieux) in the cupule to create anaerobiosis, before the test strip was incubated in an incubation box at 37 °C for 18–24 hours. The results of the test strip were interpreted by referring to the Reading Table provided with the API 20 E kit, which describes a colour or turbidity change for positive and negative tests. Some tests required the addition of reagents (API Reagent Kit, BioMérieux) to the cupule before a colour change could be revealed. Each biochemical test was determined as a positive or negative reaction, allowing a 7–digit numerical profile to be built for each bacterial isolate, based on the score for each set of 3 cupules. This numerical profile was compared against the api*web* database (BioMérieux, apiweb.biomerieux.com) and identification of the bacterial isolates were assigned to the species level, along with a confidence interval (%) for the identification. The confidence value for each species designation is based upon the percentage of biochemical tests for each strain that gave a positive result when the test strips were validated. In this study, identifications of 80% were considered as the minimum threshold for an acceptable identification. Strains identified as *E. coli* with confidence values below this threshold were discarded from further analysis.

All bacterial isolates identified as *E. coli*, with ≥ 80% confidence, were stored at –80 °C in a storage medium of Lysogeny broth (LB, Oxoid) with 20% (v/v) glycerol (Fisher Scientific). Isolates were streaked to single colonies and incubated aerobically overnight at 37 °C on LB Agar plates, and checked for purity, before performing experiments. Broth cultures of isolates were grown in LB Broth (Sigma-Aldrich) with shaking (200 rpm) at 37 °C.

Table 2.2. Identification characteristics used to presumptively identify *E. coli* from other bacterial species present in non-human samples.

	Colony morphology			
Species	**CLED with Andrade's indicator**	**HiCrome™ UTI**	**Oxidase**	**Indole**
Escherichia coli	Bright pink semi-translucent colonies with a surrounding pink halo in the medium	Purple to magenta colonies	-	+
Proteus mirabilis	Blue-green translucent colonies	Light brown colonies	-	-
Klebsiella spp.	Grey-green mucoid colonies	Blue to purple, mucoid colonies	-	-
Pseudomonas aeruginosa	Small, grey-green, translucent colonies	Colourless colonies; greenish pigment may be observed	+	-
Staphylococcus aureus	Smooth, entire, opaque; bright golden yellow colonies. Lactose fermenting	Golden yellow colonies	-	-
Enterococcus faecalis	Similar to *S. aureus* but smaller and a much deeper orange yellow colour	Blue-green, small colonies	-	-
Streptococcus pyogenes	Small opaque grey-green colonies	Pale blue/purple, minute colonies	-	-

Presumptive identifications of *E. coli* were made based on colony morphologies on a combination of agars: CLED with Andrade's indicator and HiCrome™ UTI chromogenic agar. Presumptive identifications of *E. coli* were supported by an oxidase reaction and indole test results. These presumptive identifications helped eliminate any non-Enterobacteriaceae and select for potential *E. coli* isolates in the original samples. The selected isolates were subjected to confirmatory testing and species designation using the API 20 E identification system.

2.4. Molecular characterisation of non-human *E. coli*

2.4.1. Preparation of genomic DNA

Genomic DNA (gDNA) was prepared using the GenElute™ Bacterial Genomic DNA Kit (Sigma-Aldrich), as per the manufacturer's instructions. Briefly, 1.5 mL of overnight culture grown in LB broth was centrifuged at 12,000 x *g* for 2 minutes. The resulting pellet was resuspended in 180 µL of Lysis Solution T, containing 20 µL of RNase A solution. The solution was mixed and incubated for 2 minutes at room temperature before the addition of 20 µL of Proteinase K Solution and incubation at 55 °C for 30 minutes. 200 µL of Lysis Solution C was then added, mixed, and the solution was incubated for a further 10 minutes at 55 °C. The lysis solutions used in this extraction kit contain chaotropic salts which ensure the thorough denaturation of macromolecules. 200 µL of 100% ethanol was then added to the lysate, mixed, and vortexed thoroughly to precipitate the DNA and achieve a homogeneous mixture. The addition of ethanol enables the DNA to bind to a pre-prepared spin column as the contaminants are washed through. The lysate was transferred to a binding column and centrifuged at 6,500 x *g* for 1 minute. The column was then washed twice to remove contaminants; firstly, with centrifugation at 6,500 x *g* for 1 minute, followed by a high-speed spin at 12,000 x *g* to dry the column and remove any excess wash solution. The DNA was then eluted into 50 µL of nuclease-free water by a final centrifugation at 6,500 x *g* for 1 minute. The quantity and quality of eluted gDNA was assessed using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific). Only gDNA eluates with A_{260}/A_{280} ratios of ~1.8 and A_{260}/A_{230} ratios between 2.0 and 2.2 were used as templates in the following multiplex PCR protocol, as these values are indicative of 'pure' DNA without organic or inorganic contaminants.

2.4.2. Detection of β-lactamase genes

All *E. coli* isolated from non-human samples were tested for the presence of the β-lactamase genes bla_{TEM}, bla_{SHV}, $bla_{CTX\text{-}M}$ and bla_{OXA}, using a previously published protocol and primer sequences (Fang *et al*., 2008). The primers used in the multiplex PCR assay were synthesised by Eurofins Genomics (Ebersberg, Germany). The PCR master mix contained 0.2 µM final concentrations of each primer, 2.5 mM $MgCl_2$, 0.2 mM dNTPs (Promega), 1.25 U Taq polymerase (Promega), and 2 µL DNA template. The PCR reaction tubes were cycled with the following parameters: initial denaturation at 94 °C for 5 minutes, 30 cycles of denaturation at 94 °C for 30 seconds, primer annealing at 62 °C for 90 seconds, elongation at 72 °C for 60 seconds, followed by a final elongation at 72 °C for 10 minutes. The reaction tubes were then put on hold at 4 °C. Primer sequences for each set of primers and amplicon size for each target gene are detailed in

Table 2.3. The *Klebsiella pneumoniae* strain UTI448 (NTU Pathogen Research Group culture collection), which contained known β-lactamase genes bla_{SHV}, bla_{TEM}, bla_{CTX-M} and bla_{OXA}, was included as a positive control, alongside a negative control containing no DNA, in order to monitor test performance. Amplification of DNA was performed using a Techne TC-4000 thermal cycler. PCR products were electrophoresed on 2% agarose gels, as described in section 2.4.3 below.

2.4.3. Agarose gel electrophoresis

A 2% agarose gel (which was used in all protocols) was prepared with 1 g of agarose (Fisher Scientific) per 50 mL of 1 X Tris-acetate-EDTA (TAE) buffer. A concentration of 0.1 µL/mL of SYBR® Safe DNA Gel Stain (Invitrogen) was added to the gel solution. The gel solution was then poured into a gel casting tray and once set, it was placed into an electrophoresis tank of 1X TAE buffer. Aliquots of 5 µL of PCR product were loaded into each well along with 5 µL of the appropriate molecular weight marker (100 bp ladder, New England Biolabs) in a separate well. Electrophoresis was performed at 90 V for 90 minutes. The gel was then viewed under an ultraviolet (UV) light to observe the DNA bands, using the InGenius Gel Documentation System (Syngene).

Table 2.3. Primer sequences used in this study.

Primer name	Primer sequence (5' – 3')	Target	Target gene function	Annealing temperature (°C)	Amplicon size (bp)	Reference
shvF	CTTTATCGGCCCTCACTCAA	bla_{SHV}	β-lactamase	62	237	(Fang *et al.*, 2008)
shvR	AGGTGCTCATCATGGGAAAG					
temF	CGCCGCATACACTATTCTCAGAATGA	bla_{TEM}	β-lactamase	62	445	
temR	ACGCTCACCGGCTCCAGATTTAT					
ctxmF	ATGTGCAGYACCAGTAARGTKATGGC	bla_{CTX-M}	ESBL	62	593	
ctxmR	TGGGTRAARTARGTSACCAGAAYCAGCGG					
oxaF	ACACAATACATATCAACTTCGC	bla_{OXA}	ESBL	62	813	
oxaR	AGTGTGTTTAGAATGGTGATC					

Commonly used and referenced primers for β-lactamase genes were selected from the publication indicated in the reference column. The efficacy of the assay was tested using the positive control strain *K. pneumoniae* UTI448, which was run on each gel, where possible.

The nucleotide ambiguity codes (IUPAC) are as follows: A, adenine; C, cytosine; G, guanine; T, thymine; K, keto (T or G); R, purine (A or G); S, strong (C or G); Y, pyrimidine (C or T).

2.5. Whole-genome sequencing

2.5.1. Quality assessment of genomic DNA

Genomic DNA samples with A_{260}/A_{280} ratios of ~1.8 and A_{260}/A_{230} ratios between 2.0 and 2.2 were used in the following library preparation protocol for whole-genome sequencing. The concentration of double-stranded DNA was determined by using the Qubit™ 3.0 Fluorometer. The Qubit™ dsDNA HS Assay Kit (Invitrogen) was used first, as this assay is highly selective for double-stranded DNA within a range of 0.2–100 ng/µL. If DNA samples were out of range (i.e. too high), then the Qubit™ dsDNA BR Assay Kit (Invitrogen) was used, which can measure DNA within the broader range of 2–1000 ng/µL. The Qubit™ working solution was made to a ratio of 1 µL dye:199 µL buffer for each sample. The Qubit™ 3.0 Fluorometer was calibrated using test standards for the specific kit, made to a ratio of 10 µL standard:190 µL Qubit™ solution. After calibration, the DNA samples were prepared using a ratio of 2 µL DNA sample:198 µL Qubit™ solution. The standards and samples were vortex-mixed in 0.5 mL Qubit™ Assay Tubes (Invitrogen), incubated at room temperature for 2 minutes, and then analysed on the Qubit™ 3.0 Fluorometer.

2.5.2. Illumina Nextera XT library preparation

Indexed and paired-end libraries were prepared using the Nextera XT DNA Library Preparation Kit and Nextera XT v2 Index Kit set A (Illumina). In this protocol, a fresh hard-shell skirted PCR plate was used for each set of 24 libraries. A maximum total of 48 libraries were sequenced on a single reagent cartridge, to minimise the likelihood of a reduction in sequencing coverage. Before commencing library preparation, a sample sheet (comma-separated values [CSV] file) was created, which stores the necessary information required to set up, perform, and analyse a sequencing run. The parameters specified in the sample sheet included information on experiment name, analysis workflow, read length, and adapter trimming. The library preparation process consisted of the following 5 steps:

Normalisation of gDNA concentrations across samples

The goal of this step was to normalise the gDNA concentration across samples to achieve uniform reaction efficiency in the tagmentation step. Tagmentation is sensitive to the input gDNA concentration and the optimal concentration varies depending on the organism, DNA type, and the DNA extraction method used. For this protocol, all samples were normalised to an input gDNA concentration of 0.2 ng/µL. Quantification of gDNA was performed using the Qubit™ 3.0 Fluorometer, as described in section 2.5.1.

Tagmentation of gDNA

Input gDNA was tagmented (tagged and fragmented) by the Nextera XT transposome. The adaptor sequences tagged onto the ends of the fragmented DNA allow for indexing and PCR amplification in the next step. In a hard-shell skirted PCR plate, 10 µL of Tagment DNA (TD) Buffer and 5 µL of Amplicon Tagment Mix (ATM) were mixed with 5 µL of normalised gDNA. The plate was then sealed and run on a thermal cycler at 55 °C for 5 minutes, after which, 5 µL of Neutralize Tagment (NT) Buffer was added to each well before incubating at room temperature for 5 minutes.

PCR-mediated adapter ligation and library amplification

PCR was used to incorporate the Illumina adaptor sequences (Table 2.4 and Table 2.5) to the tagmented DNA fragments, which are required for cluster formation on the flow cell. The Index 1 (i7) and Index 2 (i5) adaptors bind fragments to the flow cell, and the barcodes (N7--, N5--) allow for multiplexed sequencing. In the plate containing the tagmented gDNA, 5 µL of each Index 1 (i7) adapter was added down each column (1-6) and 5 µL of each Index 2 (i5) adapter was added across each row (A-D). Nextera PCR Master Mix (NPM) was added in volumes of 15 µL to each well containing index adaptors. The plate was sealed and subsequently run on the thermal cycler with the following parameters: 72 °C for 3 minutes; 95 °C for 30 seconds; 12 cycles of 95 °C for 10 seconds, 55 °C for 30 seconds and 72 °C for 30 seconds; 72 °C for 5 minutes and hold at 10 °C.

Post-PCR clean-up of libraries

The PCR products were purified with Agencourt AMPure XP magnetic beads (Beckman Coulter) to remove short library fragments. In this protocol, 30 µL of homogenised and room temperature beads were added to each well, mixed by pipetting, and incubated at room temperature for 5 minutes, before placing on a magnetic stand to pellet the beads. The supernatant was discarded and the beads were washed 2 times with 200 µL of freshly prepared 80% ethanol, before air-drying the pellet on the magnetic stand for 5–10 minutes, or until the pellet appears “matte” in appearance. The pellet was then resuspended in 52.5 µL of Nextera Resuspension Buffer (RSB) and incubated on the magnetic stand for 2 minutes. 50 µL of the supernatant, which contained the purified libraries, was transferred to a new plate for the subsequent steps.

Library quality control, normalisation, and pooling

Sample concentrations and fragment size distributions were estimated and the quantity of each library was normalised to ensure equal library representation in pooled samples. The DNA concentration of each sample was measured using the Qubit™ 3.0 Fluorometer as described in section 2.5.1 'Quality assessment of genomic DNA'. Fragment size (bp) distribution was measured for each library, using the Agilent 2200 TapeStation, by following the manufacturer's instructions. Briefly, 2 µL of High Sensitivity D1000 sample buffer (Agilent Technologies) was aliquoted into sets of optical tube strips, followed by 2 µL of High Sensitivity D1000 ladder (Agilent Technologies) in the first tube, then 2 µL of gDNA library in the subsequent tubes. After vortex-mixing at 2,000 rpm for 1 minute, the samples were loaded into the 2200 TapeStation, along with the High Sensitivity D1000 ScreenTape and loading tips (Agilent Technologies). The analysis was then initiated by launching the Agilent 2200 TapeStation software, resulting in a rapid determination of the average fragment size for each gDNA sample. Based on individual sample concentrations and the common average fragment length, the DNA molarity of each sample was calculated using the following formula:

$$m = \frac{c}{w \times s} \times 1{,}000{,}000$$

Where m = molarity (nM), c = concentration of gDNA library (ng/µL), w = average molecular weight of DNA (taken to be 660 g/mol), s = average DNA fragment size of library.

Each library was then normalised to 4 nM before pooling an equal 5 µL of each library into a single microcentrifuge tube.

2.5.3. Sequencing on the MiSeq

To prepare libraries for sequencing on the Illumina MiSeq, 5 µL of the 4 nM pooled libraries was denatured with 5 µL 0.2 N NaOH. Incubating for 5 minutes at room temperature enabled denaturation of the gDNA into single strands. The DNA was then diluted with 990 µL of pre-chilled Hybridization Buffer (HT1), resulting in a 20 pM library. This was further diluted to 12 pM – the recommended concentration for the MiSeq Reagent Kit v2 (Illumina), which was used in this protocol. A PhiX control library was prepared by denaturing 2 µL of 10 nM PhiX with 3 µL Tris-HCl (pH 8.0) and 5 µL 0.2 N NaOH. The denatured PhiX library was then diluted to 20 pM using 990 µL HT1 buffer. This was then further diluted to 12.5 pM and combined with the diluted and denatured DNA library at a ratio of 6 µL PhiX:594 µL DNA library. The flow cell was cleaned using 80% ethanol and was loaded into the MiSeq along with Incorporation Buffer. The prepared libraries were loaded onto the reagent cartridge and sequenced on the MiSeq for 500 cycles, to generate 250 bp paired-end reads.

Table 2.4. Illumina adapter sequences for Nextera XT Index Kit v2 – index 1 (i7) adapters

Bases in adapter	i7 index name	i7 bases for entry on sample sheet
TCGCCTTA	N701	TAAGGCGA
CTAGTACG	N702	CGTACTAG
TTCTGCCT	N703	AGGCAGAA
GCTCAGGA	N704	TCCTGAGC
AGGAGTCC	N705	GGACTCCT
CATGCCTA	N706	TAGGCATG
GTAGAGAG	N707	CTCTCTAC
CAGCCTCG	N710	CGAGGCTG
TGCCTCTT	N711	AAGAGGCA
TCCTCTAC	N712	GTAGAGGA
TCATGAGC	N714	GCTCATGA
CCTGAGAT	N715	ACTCGCTA
CCTGAGAT	N716	GGAGCTAC
GTAGCTCC	N718	GGAGCTAC
TACTACGC	N719	GCGTAGTA
AGGCTCCG	N720	CGGAGCCT
GCAGCGTA	N721	TACGCTGC
CTGCGCAT	N722	ATGCGCAG
GAGCGCTA	N723	TAGCGCTC
CGCTCAGT	N724	ACTGAGCG
GTCTTAGG	N726	CCTAAGAC
ACTGATCG	N727	CGATCAGT
TAGCTGCA	N728	TGCAGCTA
GACGTCGA	N729	TCGACGTC

Table 2.5. Illumina adapter sequences for Nextera XT Index Kit v2 – index 2 (i5) adapters

Bases in adapter	i5 index name	i5 bases for entry on sample sheet
CTCTCTAT	S502	ATAGAGAG
TATCCTCT	S503	AGAGGATA
GTAAGGAG	S505	CTCCTTAC
ACTGCATA	S506	TATGCAGT
AAGGAGTA	S507	TACTCCTT
CTAAGCCT	S508	AGGCTTAG
CGTCTAAT	S510	ATTAGACG
TCTCTCCG	S511	CGGAGAGA
TCGACTAG	S513	CTAGTCGA
TTCTAGCT	S515	AGCTAGAA
CCTAGAGT	S516	ACTCTAGG
GCGTAAGA	S517	TCTTACGC
CTATTAAG	S518	CTTAATAG
AAGGCTAT	S520	ATAGCCTT
GAGCCTTA	S521	TAAGGCTC
TTATGCGA	S522	TCGCATAA

2.6. Analysis of *E. coli* sequence data

2.6.1. Genome assembly, annotation, and quality assessment

The Illumina software MiSeq Reporter performed secondary analysis which involved demultiplexing of all indexed reads – dividing all sequenced reads into separate files for each indexed tag/sample. The analysis also included generation of FASTQ files, which contained the non-indexed reads for each isolate, excluding reads identified as inline controls and reads that did not pass filter. FASTQ generation performed quality trimming of the 5' portion of the adapter sequence for all reads, to prevent reporting sequence data beyond the sample DNA. *De novo* assemblies of raw reads into contigs and scaffolds were performed using SPAdes v3.9.0 (Bankevich *et al*., 2012). The genome assemblies were annotated using Prokka v1.12 (Seemann, 2014), by running the python script autoprokka.py (https://github.com/stevenjdunn/autoprokka) on the FASTA genome files. The quality of *de novo* assemblies was evaluated by running QUAST v3.2, a bioinformatics quality assessment tool for genome assemblies (Gurevich *et al*., 2013). Any incomplete assembled genomes, i.e. assemblies with N50 values < 1,900 bp and genome sizes < 4.3 Mbp were excluded from further analyses. The assembled genomes passing these criteria were analysed as described in sections 2.6.2 – 2.6.7.

2.6.2. Sequence-typing and clonal complex assignment

Sequence typing of isolates was carried out by running the multilocus sequence typing (MLST) script (https://github.com/tseemann/mlst), which scans the assembled scaffolds against PubMLST databases and assigns a sequence type (ST) to each isolate. Closely related STs were grouped into ST complexes using the PHYLOViZ v3.0 platform (Francisco *et al*., 2012). PHYLOViZ uses the goeBURST algorithm, a refinement of eBURST (Feil *et al*., 2004), and its expansion to generate a complete minimum spanning tree (MST) of all STs. It achieves this by dividing an MLST data set into groups of related isolates and clonal complexes, predicting the ancestral genotype of each clonal complex, and computing the bootstrap support for the assignment.

2.6.3. Detecting antibiotic resistance genes and ExPEC virulence determinants

Genome sequence files were screened for the presence of acquired antibiotic resistance genes by running the bioinformatics tool ABRicate v2.0 (https://github.com/tseemann/abricate), which uses the ResFinder database to generate *in silico* antibiotic resistance profiles for each isolate. ABRicate was also run using VFDB – Virulence Factors Database (Chen *et al*., 2015) – to

mass screen contigs for the presence of virulence determinants that define ExPEC strains. All redundant isolates (i.e. isolates that were collected from the same sample source, assigned to the same ST, and exhibiting the same antibiotic resistance gene profiles) were excluded from further genomic analyses.

2.6.4. Reconstructing phylogenetic trees

Core genome alignments of strains were constructed from genome assemblies using Parsnp v1.2 (Treangen *et al*., 2014), as described in section 2.2.1. One of the output files from Parsnp is a maximum-likelihood phylogenetic tree, reconstructed from the alignment, based on core genome SNP analysis. Phylogenetic trees were visualised using iTOL (http://itol.embl.de/) (Letunic and Bork, 2016) and trees were annotated with coloured bars to display metadata encompassing information on isolation source and presence of antibiotic resistance genes and virulence-associated genes. Reference genomes belonging to each of the following phylogroups (A, B1, B2, D, E, and F) or cryptic clades (clade I to V) were included in the initial phylogenetic tree. Reference genomes were selected from the genomes analysed in previous studies (Kaas *et al*., 2012; Luo *et al*., 2011). Each strain was then assigned to a phylogroup or a cryptic clade based on the position in the phylogeny.

2.6.5. Pan-genome analysis

Pan-genomes were constructed from the annotated genome assemblies produced by Prokka (GFF3 format) using Roary v3.8.2 (Page *et al*., 2015). Each input file should have a unique locus tag for the gene IDs to make it easier to identify where the genes came from. A gene presence and absence spreadsheet (CSV file) was produced as an output of Roary, which lists each gene and the strains they are present in. The accessory genomes of strains were analysed by excluding genes present in ≥ 85% of strains. The gene product functions of excluded genes were evaluated and were confirmed to be core genes of the *E. coli* genome. The query_pan_genome script, which takes the annotated GFF files and clustered_proteins output file from Roary, was run to identify genes that were either unique or shared between the human-clinical and non-human populations of *E. coli.* Any unique or shared genes were listed with strains in separate CSV files. Gene presence and absence matrices were annotated on core genome phylogenetic trees and visualised with associated metadata, using the interactive tool Phandango v1.1.0 (Hadfield *et al*., 2018; https://github.com/jameshadfield/phandango).

2.6.6. Detection of recombinant genomic regions

Core genome alignments were constructed using Parsnp (Treangen *et al*., 2014), as described in section 2.2.1. Genomic recombination events between strains were detected by running the Gubbins v2.2.0 (Genealogies Unbiased By recomBinations In Nucleotide Sequences) (Croucher *et al*., 2015) algorithm on the core genome FASTA alignment, with default parameters. Recombination predictions were output in a GFF file and the recombination blocks were then visualised against the core phylogenetic Newick-formatted tree and associated metadata, using Phadango (Hadfield *et al*., 2018).

2.6.7. Statistical analyses

In this study, statistical analyses were performed to compare the prevalence of extended spectrum β-lactamase genes and sequence types between populations. *p*-values were calculated with two tails using the GraphPad QuickCalcs Fisher's test calculator (https://www.graphpad.com). A *p*-value of $p < 0.05$ would indicate a statistically significant association. Moreover, *p* values of $p < 0.001$ and $p < 0.0001$ were considered very statistically significant, and highly statistically significant, respectively.

Shannon diversity index values and the Hutcheson t-test were used to compare the diversity of the human-clinical and non-human populations of *E. coli*. These were calculated using a custom Excel spreadsheet.

Comparisons between the observed and expected proportions of genes shared between the human-clinical and non-human populations were made by performing a permutation test with pseudo-random re-sampling of the population without replacement of genomes. This approach considered gene frequencies, and all strain-specific genes were excluded from the analysis. Permutations were carried out 1,000 times iterating for each gene category, and each permutation involved picking the same number of genomes as there are genes in that category. To test the significance of differences between expected and observed frequencies of gene sharing, one-tailed empirical *p*-values were calculated. The permutation script was written by Ben Dickins (NTU) and simulated proportions were plotted as histograms, with the observed proportions mapped on the graphs for comparison. Graphs were produced using the tidyverse package (Wickham, 2017) of the R statistical software v3.5.1 (http://www.r-project.org/; R Core Team, 2015).

CHAPTER 3

Ecology and evolution of a global population of *Yersinia pseudotuberculosis*

3.1. Introduction

The genus *Yersinia* belongs to the Gram-negative Enterobacteriaceae and it constitutes a model genus for studying bacterial pathogen ecology and evolution (McNally *et al*., 2016b). The majority of *Yersinia* species are found widely in the environment, including soil, and do not usually cause disease in mammals. There are three main species of *Yersinia* which are well-recognised human pathogens: the plague bacillus *Yersinia pestis*, and the enteropathogenic *Yersinia enterocolitica* and *Yersinia pseudotuberculosis* (McNally *et al*., 2016b). *Y. pseudotuberculosis*, which causes infection in a wide range of hosts such as domesticated and wild animals, has also been implicated in foodborne infection in humans, which is known as yersiniosis. Transmission of the bacterium is usually via the faecal–oral route, and human infection can result from ingestion of contaminated food products or water, or alternatively by direct contact with an infected animal or human (Savin *et al*., 2014; Chiles *et al*., 2002). Although *Y. pseudotuberculosis* is not as frequently associated with human gastrointestinal yersiniosis as *Y. enterocolitica*, both are important pathogens with similar clinical manifestations associated with infection. These can include fever, abdominal pain, and diarrhoea. In some uncommon cases, extraintestinal symptoms such as reactive arthritis and erythema nodosum can occur (Jalava *et al*., 2006). Additionally, in more severe cases, infection can disseminate to the bloodstream and deep tissues (Kaasch *et al*., 2012).

Classically, identification and typing of *Y. pseudotuberculosis* has been based on serotyping of the lipopolysaccharide O-antigen, with a total of 21 serotypes identified, including 6 originally classified as subtypes of either O:1, O:2, O:4 or O:5 (Skurnik, Peippo and Ervela, 2000). Previous studies have indicated that the majority of strains isolated from human cases belong to serotypes O:1a, O:1b, and O:3 (Williamson *et al*., 2016; Laukkanen-Ninios *et al*., 2011). The application of serotyping methods provides only a low-level resolution when investigating potential outbreaks of *Y. pseudotuberculosis*, suggesting the need for higher resolution typing methods. The population structure of *Y. pseudotuberculosis* has been defined by multilocus sequence typing (MLST) previously (Laukkanen-Ninios *et al*., 2011), adding detail to the serotype differentiation. It was revealed that serotype O:1 strains formed a distinct clade of strains which represented a large number of sequence type complexes, suggesting a highly diverse population of bacteria within the serotype O:1 group, and thus an overall high diversity of the *Y. pseudotuberculosis* species (Laukkanen-Ninios *et al*., 2011). In addition to MLST genotyping, a recent study has also analysed clustered regularly interspaced short palindromic repeat (CRISPR) loci of a large collection (n = 355) of *Y. pseudotuberculosis* isolates (Koskela *et al*., 2015). The CRISPR system is an adaptive RNA-based immune system that constitutes a bacterial defence against foreign nucleic acids derived from invading bacteriophages or plasmids. CRISPRs are

constructed from a chain of 21 to 47 bp repeated sequences called direct repeats (DR), and in between DRs are unique spacer sequences, which represent the acquired foreign DNA. One of the key findings from the study by Koskela and colleagues (2015) was that despite the genetic similarity between *Y. pseudotuberculosis* and *Y. pestis*, strains of these species shared very few CRISPR spacers, suggesting that CRISPR analysis can be used to inform bacterial evolution.

Although most infections with *Y. pseudotuberculosis* are thought to be sporadic (Sunahara, Yamanaka and Yamanishi, 2000), this pathogen has been responsible for several nationwide gastrointestinal outbreaks of foodborne infection in countries of largely temperate climates, such as Finland, Russia, France and New Zealand (Williamson *et al*., 2016; Kangas *et al*., 2008; Jalava *et al*., 2006; Nuorti *et al*., 2004). More recently, in 2014, a sustained outbreak of yersiniosis caused by *Y. pseudotuberculosis* occurred in New Zealand, affecting all the major cities of the country (Williamson *et al*., 2016). With a total of 220 laboratory-confirmed cases of infection reported, this outbreak constitutes one of the largest globally reported outbreaks of human yersiniosis due to *Y. pseudotuberculosis*, to date. Prior to the present study, the most inclusive genome-scale analysis of *Y. pseudotuberculosis* focussed on the outbreak in New Zealand (Williamson *et al*., 2016). Genomic and epidemiological analyses in that study suggested a single point-source contamination of the food chain, with subsequent nationwide distribution of contaminated produce. Additionally, through incorporation of publicly available reference genomes within the context of a globally and taxonomically diverse dataset, the study indicates that *Y. pseudotuberculosis* is a highly diverse species and that the New Zealand strains represented a geographically isolated clade of *Y. pseudotuberculosis.*

Large-scale population genomics studies have been performed on the two other human pathogenic *Yersinia*, *Y. pestis* and *Y. enterocolitica* (Reuter *et al*., 2015; Reuter *et al*., 2014; Morelli *et al*., 2010), allowing a high-resolution understanding of the ecology, evolution and dissemination of these pathogens. Global phylogenomic studies of *Y. pestis* have identified evolution from a clone of *Y. pseudotuberculosis*, as a result of gene loss and subsequent global dissemination (McNally *et al*., 2016b; Morelli *et al*., 2010). In contrast, similar global genomic studies of *Y. enterocolitica* have suggested an evolutionary path from a non-pathogenic ancestor through gene gain and loss, resulting in apparently ecologically separated clades within the species (Reuter *et al*., 2015; Reuter *et al*., 2014). A recent study by our group involved population genomic analysis of *Y. enterocolitica*, through means of examining patterns of recombination in both the core and accessory genome of the species (Reuter *et al*., 2015). The study highlights that genetic flow in the species does not occur frequently between phylogroups of *Y. enterocolitica*, and when it does, it is primarily unidirectional with one phylogroup acting largely as a genetic reservoir for the rest of the species. Additionally, the data revealed hidden

ecological patterns and the analysis suggests that the distinct phylogroups of *Y. enterocolitica* may be ecologically separated, with phylogroup 1 (PG1) being ubiquitous and most commonly isolated from non-human environments, whilst PG2–5 are more commonly isolated from human disease cases (Reuter *et al*., 2015). The *Y. enterocolitica* study draws parallels with a study of *Escherichia coli* where core genome recombination was not observed between environmental and human/animal isolates; recombination was only detected between environmental isolates or between human and animal isolates (Luo *et al*., 2011). Furthermore, ecological separation within a hospital environment, leading to reduced recombination between isolated subpopulations of the nosocomial pathogen *Enterococcus faecium* has been reported previously (Willems *et al*., 2012). The observation of host-restricted lineages of *Campylobacter jejuni* (Sheppard *et al*., 2014) is further supportive of ecological barriers playing a major role in restricting genetic flow and recombination, and thus leading to the formation of distinct ecotypes within a bacterial species.

3.1.1. Aim and objectives

Given the observation of ecologically separated lineages in *Y. enterocolitica*, and considering that *Y. pseudotuberculosis* is a closely related species that is also heterogeneous and ubiquitous in nature, it would be beneficial for our understanding of microbial evolution to investigate whether similar ecological inferences can be made for *Y. pseudotuberculosis.* To contribute a high-resolution genomic view into the ecology, evolution and dissemination of this important human pathogen, a data set of globally and temporally distributed *Y. pseudotuberculosis* genomes, from multiple ecosystems, were analysed in this chapter. This was achieved by employing highly discriminatory comparative genomic techniques, such as pan-genome analysis and core genome recombination analysis.

Specific objectives of this chapter were:

- To reconstruct the phylogeny of a global population of *Y. pseudotuberculosis* and identify any genomic signatures correlated with the ecology of the *Y. pseudotuberculosis* population.
- To investigate the evolutionary history of *Y. pseudotuberculosis* by performing dating analysis to date the phylogeny of the species.
- To determine whether any patterns of gene flow exist within the *Y. pseudotuberculosis* population by analysing the core and accessory genomes of these strains.

3.2. Materials and Methods

The key methods, bioinformatic tools and scripts used in this chapter were described previously in section 2.2. of chapter 2 'Materials and Methods'. A total of 134 *Y. pseudotuberculosis* genomes were analysed in this chapter (Table 3.1), of which 108 were recently sequenced and were provided as genome sequence (FASTA) files by Mikael Skurnik (University of Helsinki). These isolates were collected over a 46-year time frame from 19 different countries across 6 continents, and represent the full spectrum of serotypes possible. Additionally, the strains were isolated from a wide range of hosts including livestock, wild animals and companion animals, human-clinical and environmental sources. Library preparation and sequencing of these isolates were performed using the Illumina Nextera kit and Genome Analyzer IIx instrument to create 150 bp paired-end reads. This was carried out by Mikael Skurnik and Laura Kalin-Mänttäri at the FIMM Sequencing unit (Helsinki, Finland). *De novo* assemblies were achieved using Velvet v1.2.1 (Zerbino and Birney, 2008) and annotated using Prokka v1.12 (Seemann, 2014). The raw sequence reads were deposited to the European Nucleotide Archive (ENA) under project PRJEB14064. Additionally, *de novo* assemblies of all genomes used are available on Enterobase (https://enterobase.warwick.ac.uk/species/index/*Yersinia*), searchable by the strain names or ENA accession numbers listed in Table 3.1. Metadata regarding information on isolation (continent, country, host, and year), sequence type, serotype, and CRISPR spacer are available for the majority of strains. The CRISPR loci were identified by Katja Koskela (University of Helsinki) using BLASTN (Koskela *et al*., 2015). The full list of spacer arrays used for CRISPR spacer clustering for the strains analysed in this study are available from doi: 10.1099/mgen.0.000133 as supplementary material (Seecharran *et al*., 2017).

Table 3.1. ***Y. pseudotuberculosis*** **genomes used in this study.**

Strain	Year	ST*	Serotype	CRISPR cluster	Host	Continent	Country	Accession number
1180/95	1995	-	1a	20	-	-	-	ERR1448065
2384	-	19	3	35	-	-	-	ERR1448071
CIP 55.85	1952	-	1	11	Turkey	-	-	ERR1447956
PST25	-	43	-	38	-	-	-	ERR1447955
2515	-	-	2	32	-	-	-	ERR1448073
P 105	1990	14	-	9	Buffalo	-	-	ERR1448074
PST2660	-	14	3	9	-	-	-	NC_010465
PST1813	-	14	4	26	-	-	-	ERR1448070
YPIII	-	-	3	9	-	-	-	ERR1448072
IP32544	-	19	3	-	Pig	Africa	South Africa	ERR024924
488	-	-	1a	-	Salmon	Asia	Russia	ERR024916
514	-	-	1a	20	Salmon	Asia	Russia	ERR1448002
RU496	-	-	1a	20	Reindeer	Asia	Russia	ERR1448001
489	-	-	1a	20	Reindeer	Asia	Russia	ERR1447987
RU488	-	42	1a	20	Salmon	Asia	Russia	ERR1448003
N912	-	14	2b	-	Rabbit	Asia	China	ERR1447986
H722-36/88	1986	11	6	8	Dog	Asia	Japan	ERR024920
IP33177	-	26	1	-	Cabbage	Asia	Russia	ERR1447974
DC356	-	33	1b	6	Cat	Asia	Japan	ERR1447962
Pa3606	-	2	1b	40	Human	Asia	Japan	ERR1447976
H-1	-	2	1b	12	Human	Asia	Russia	ERR1447977
H-158	-	-	1b	12	Human	Asia	Russia	ERR1447978
H-2212	-	2	1b	12	Human	Asia	Russia	ERR1447979
H1647	-	-	1b	12	Human	Asia	Russia	ERR1448019

H416	-	2	1b	12	Human	Asia	Russia	ERR1448021
Pa3597	-	-	1b	40	Human	Asia	Japan	ERR1448022
8011-3	-	33	1b	15	Human	Asia	Japan	ERR1448013
IP31758	1966	37	1b	12	Human	Asia	Russia	NC_009708
Gifu-liver	-	38	1b	13	Monkey	Asia	Japan	ERR1448026
MW145-2	-	89	1b	40	Freshwater	Asia	Japan	ERR1448014
H1746	-	2	1b	12	Small mammal	Asia	Russia	ERR1447980
H2517	-	-	1b	12	Small mammal	Asia	Russia	ERR1447981
H404	-	-	1b	12	Small mammal	Asia	Russia	ERR1448020
Wla352	-	-	1b	37	Racoon dog	Asia	Japan	ERR1447975
1231	1985	2	4b	-	Small mammal	Asia	Russia	ERR024910
Soil-4	-	33	1b	15	Environment	Asia	Japan	ERR1447973
MW Taniguci	-	-	1b	37	Freshwater	Asia	Japan	ERR1448015
SP93422	1993	1	15	-	Human	Asia	Korea	ERR027412
D1040	-	44	2b	25	Dog	Asia	Japan	ERR1448036
Wla708	-	33	1b	40	Duck	Asia	Japan	ERR1447970
K22	-	40	5a	23	Human	Asia	Japan	ERR1447990
79136	-	88	1b	16	Human	Asia	Korea	ERR1448017
Uematu289	-	-	1b	2	Human	Asia	Japan	ERR1448028
Chigamatsu	-	-	1b	4	Human	Asia	Japan	ERR1448032
GS95	-	43	3	1	Human	Asia	China	ERR1448040
IP33250	-	32	3	-	Human	Asia	Russia	ERR024921
PC94-72	-	44	4a	25	Pig	Asia	Japan	ERR1447989
TP1039	-	-	1b	3	Pig	Asia	Japan	ERR1448027
PT682	1987	52	2b	37	Pig	Asia	Japan	ERR024908
PC504	-	44	2c	2	Pig	Asia	Japan	ERR1448037
S106	-	-	1b	1	Rabbit	Asia	China	ERR1447972

J51	-	47	13	31	Rabbit	Asia	China	ERR1447991
R103-2	-	45	5b	7	Rabbit	Asia	China	ERR1448042
R104-2	-	46	5b	14	Rabbit	Asia	China	ERR1448043
OK5586	1990	62	3	-	Racoon dog	Asia	Japan	ERR024907
OK6088	1990	18	10	5	Racoon dog	Asia	Japan	ERR027411
CN2	-	49	1c	21	Small mammal	Asia	China	ERR1447988
DD362	-	-	1b	29	Dog	Asia	Japan	ERR1447984
PC708	-	3	1b	6	Pig	Asia	Japan	ERR1447985
T-469-1	-	-	1b	29	Pig	Asia	Japan	ERR1448025
RD20	-	31	1b	37	Racoon dog	Asia	Japan	ERR1447971
TE-93181	-	3	1b	29	Racoon dog	Asia	Japan	ERR1448033
2814/1998	1998	-	1a	20	Hare	Europe	Finland	ERR1448038
3822/2000	2000	-	1a	20	Hare	Europe	Finland	ERR1448068
IH111554	-	42	1a	-	Cat	Europe	Finland	ERR1447982
103	-	42	1a	20	Small mammal	Europe	Italy	ERR1448016
PB1	1960	68	1a	10	Small mammal	Europe	England	ERR1448064
2886	-	42	1a	10	Hare	Europe	Italy	ERR1448029
2809/1998	1998	-	1a	20	Small mammal	Europe	Finland	ERR1448030
15193/74	1974	-	1a	20	Human	Europe	Finland	ERR1448031
Rollier	-	42	1a	20	Human	Europe	Belgium	ERR024925
H655-36/87	1987	42	1a	20	Human	Europe	Germany	ERR1448048
Y.PT/8	-	42	1a	20	Human	Europe	Belgium	ERR1447966
921/93	1993	-	1a	20	Human	Europe	Sweden	ERR1448063
11J	-	42	-	20	Human	Europe	France	ERR024923
IP32953	-	-	1b	42	Human	Europe	France	ERR024927
2800/1998	1998	-	1a	41	Jack daw	Europe	Finland	ERR024926
104	-	42	1a	-	Pigeon	Europe	Italy	ERR024930

504/72	1972	42	1a	20	Duck	Europe	Italy	ERR1448035
Y722	1988	19	1	35	Human	Europe	Germany	ERR1448060
25418L	-	9	1a	11	Canary	Europe	Denmark	ERR1448057
H943-36/89	1989	9	1a	11	Hare	Europe	Germany	ERR1448010
St.1	-	-	1a	11	Human	Europe	Germany	ERR1447963
H892-36/87	1987	12	1a	39	Human	Europe	Italy	ERR1448069
H942-36/89	1989	9	1a	11	Human	Europe	Germany	ERR1448062
2895	-	43	1b	22	Bird	Europe	Italy	ERR1448056
2817/1998	1998	-	1b	22	Hare	Europe	Finland	ERR1448061
3858/2000	2000	-	1b	38	Hare	Europe	Finland	ERR1448044
H749-36/89	1989	43	1b	22	Duck	Europe	Germany	ERR024903
2887	-	43	1b	38	Hare	Europe	Italy	ERR1448045
H938-36/89	1989	43	1b	38	Hare	Europe	Germany	ERR1448049
2497	-	43	1b	38	Hare	Europe	Italy	ERR1448052
3876/2001	2001	-	1b	22	Hare	Europe	Finland	ERR028208
2812/79	1979	-	1b	38	Human	Europe	Finland	ERR1447993
866/81	1981	-	1b	38	Human	Europe	Finland	NC_010634
36/83	1984	-	1b	38	Human	Europe	Finland	ERR1448008
Tytgat	-	43	1b	38	Human	Europe	Belgium	ERR1448047
Y.PT/7	-	43	1b	38	Human	Europe	Belgium	ERR1447959
42/00	2000	-	1b	22	Human	Europe	Sweden	ERR1447996
G2/77/2	1977	43	1b	-	Pet bird	Europe	Denmark	ERR1447999
G798/82/1	1982	43	1b	38	Pet bird	Europe	Denmark	ERR1448004
8597L	-	43	1b	38	Pet bird	Europe	Denmark	ERR1448066
IP32670	-	43	1	-	Pig	Europe	UK	ERR1447954
2812/1998	1998	-	1b	22	Pigeon	Europe	Finland	NC_006155
2889	-	43	1b	22	Turkey	Europe	Italy	ERR1448046

IP33290	-	-	1	38	-	Europe	France	ERR1447992
IP32463	-	16	5	-	Small mammal	Europe	Switzerland	ERR024894
IP32921	-	16	2	-	Hare	Europe	France	ERR1448005
IP32881	-	16	2	-	Monkey	Europe	Switzerland	ERR1447997
IP33054	-	14	2	-	Human	Europe	Spain	ERR1447961
2884	-	14	2a	2	Rabbit	Europe	Italy	ERR1448000
1435/8/2004	2004	-	1b	28	Carrot	Europe	Finland	ERR1447998
2161/13/2006	2006	-	1b	28	Environment	Europe	Finland	ERR1448023
Marsu	1980	14	3	9	Small mammal	Europe	Finland	ERR1448050
7616/84	1984	-	1a	9	Human	Europe	Finland	ERR1448053
2874/2003	2003	-	1b	28	Environment	Europe	Finland	ERR1447983
2484/2006	2006	-	1b	28	Environment	Europe	Finland	ERR1447965
3623/13/2004	2004	-	1b	28	Small mammal	Europe	Finland	ERR1448011
5456/85	1985	-	1b	12	Human	Europe	Finland	ERR1448024
Y718	1986	2	1b	15	Human	Europe	Germany	ERR1448054
677/82	1982	-	1b	12	Human	Europe	Finland	ERR1447957
260	-	42	1a	-	Human	North America	Canada	ERR024902
Y716	-	19	1a	35	Small mammal	North America	USA	ERR024914
283	-	14	1b	28	Human	North America	Canada	ERR1447968
284	-	14	1b	4	Human	North America	Canada	ERR1447969
G5137	-	42	1a	20	Cow	Oceania	Australia	ERR024929
BB1152	-	42	1a	10	Deer	Oceania	New Zealand	ERR024917
No.93	-	42	1a	10	Goat	Oceania	New Zealand	ERR1447995
No.21	-	86	1a	11	Cattle	Oceania	New Zealand	ERR1447994
H305-36/89	1989	43	1b	22	Deer	Oceania	Australia	ERR1448007
IP33038	-	43	1	-	Small Mammal	Oceania	Australia	ERR1448006
No.5	-	54	2b	-	Goat	Oceania	New Zealand	ERR1448012

BF-1	-	-	1a	35	Buffalo	South America	Brazil	ERR1447967
IP32938	-	19	3	-	Cattle	South America	Argentina	ERR024928

Table 3.1. *Y. pseudotuberculosis* genomes used in this study.

Strains underlined in the table were sequenced as part of a previously published study (Reuter *et al*., 2014). All other strains listed were isolated and sequenced as described in section 3.2 and were published recently (Seecharran *et al*., 2017).

*ST: multilocus sequence type according to MLST databases at the University of Warwick, Warwick Medical School (http://mlst.ucc.ie/mlst/dbs/Ypseudotuberculosis/GetTableInfo_html) (Laukkanen-Ninios *et al*., 2011).

3.3. Results and Discussion

3.3.1. Phylogeographic structure of *Y. pseudotuberculosis* indicates an Asian ancestry for the species

A maximum-likelihood SNP-based phylogenetic tree was reconstructed from a core genome alignment of 134 *Y. pseudotuberculosis* genome sequences built using Parsnp (Treangen *et al*., 2014). This analysis revealed that the phylogeny of the *Y. pseudotuberculosis* population is characterised by clades of varying genomic diversity (Fig. 3.1). There is a clade containing high levels of diversity, indicated by long branch lengths and a clade containing noticeably lower levels of diversity, as shown by much shorter branch lengths. Inclusion of a selection of *Y. similis* and *Y. wautersii* strains (Appendix 2) on the tree as an outgroup confirmed the highly diverse clade to be the ancestral clade for *Y. pseudotuberculosis*, as indicated by the tree topology (Appendix 3). Geographical source of isolation is available as part of the metadata for the majority of strains in the population. The phylogenetic tree was annotated with the continent of origin for each strain, where available, as illustrated by coloured bars at the branch tips in Figure 3.1. With regards to most of the strains isolated from Russia (Table 3.1), it was not entirely clear what specific geographical region these isolates were derived from. However, with the presence of Siberian isolates in the data set (e.g. strain IP33177), all strains obtained from Russia were accordingly categorised as being isolated from Asia. Annotation of continent of origin data revealed a very clear geographic split in the phylogeny, with the ancestral highly diverse clade containing predominantly Asian origin isolates and the second low diversity clade containing predominantly European origin isolates. A small transitional cluster of isolates originating from South Africa, North and South America, and Europe can be seen between the two distinct clades. The phylogenetic structure of globally dispersed isolates demonstrated in this chapter is consistent with an Asian origin for *Y. pseudotuberculosis*, with two separate migrations into Africa and the Americas, and more recently into Europe. An Asian ancestry of the *Y. pseudotuberculosis* species is in line with the postulated ancestry of *Y. pestis* (Morelli *et al*., 2010; Achtman *et al*., 1999), which is a clone that recently evolved from *Y. pseudotuberculosis* within the last 3,000 to 6,000 years, shortly before the first known human plague pandemics originating from central Asia. The data from the present study, however, appear to show that the greatest genetic variation occurs in Japan, not China. This does not appear to be an artefact of the geographical sampling in this chapter, however, it cannot be discounted that a more thorough and dense genomic sampling may provide a different result. Migration of isolates into Europe is consistent with a bottleneck event leading to the successful establishment of a small number of clones, as demonstrated by the low genetic diversity exhibited by the European clade of isolates.

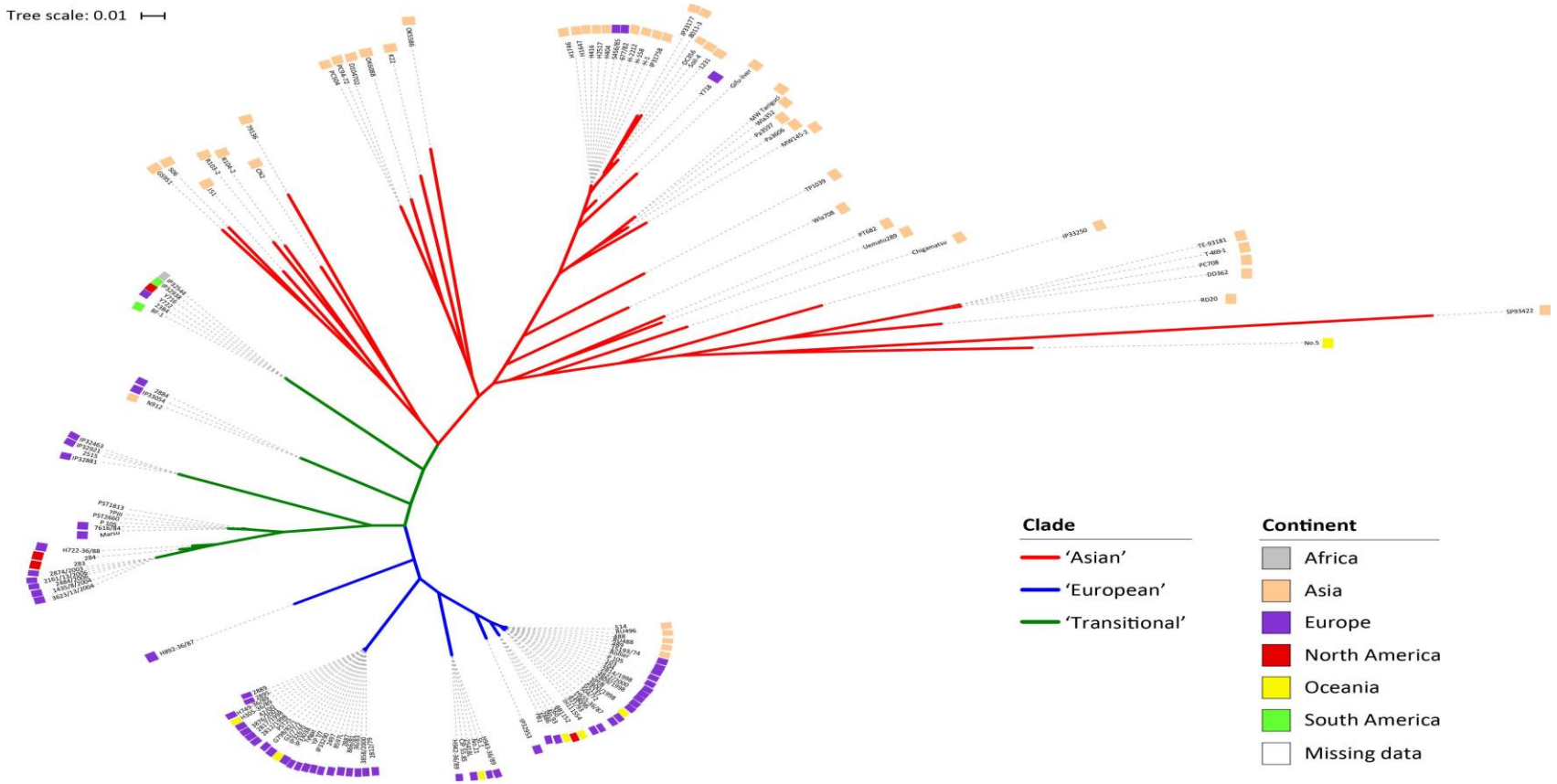


Fig. 3.1. Maximum-likelihood phylogenetic tree of 134 *Y. pseudotuberculosis* isolates. The phylogeny is derived from a core genome alignment (2,947,945 bp, 1,042,987 SNPs) constructed using Parsnp and the tree was visualised using iTOL. The scale bar corresponds to the number of nucleotide substitutions per site. Continent of origin is superimposed on the tree as coloured bars. Geographic clades are defined by tree branch colouring. There is a clear phylogeographic split in the population, with the ancestral, highly diverse clade containing primarily Asian isolates and a second low diversity clade containing primarily European isolates. A small transitional cluster of African, American, and European isolates exists between the two clades.

3.3.2. Phylogenetic clusters within *Y. pseudotuberculosis* associate with discrete CRISPR cassette patterns

CRISPR spacer data were available as part of the metadata for the majority of *Y. pseudotuberculosis* strains analysed in this chapter. The CRISPR loci for these strains were identified previously by Katja Koskela (University of Helsinki) using BLASTN (Koskela *et al*., 2015). Bayesian clustering of the presence/absence of all 2,969 known *Y. pseudotuberculosis* CRISPR spacers present in the data set (performed by Jukka Corander, University of Oslo) identified a total of 33 distinct sequence clusters of CRISPR cassettes (Table 3.1, section 3.2). The CRISPR cluster data were used initially to determine any obvious genotypic traits associated with the phylogeographic split observed in the *Y. pseudotuberculosis* population. Each identified CRISPR cluster was annotated on to the *Y. pseudotuberculosis* maximum-likelihood phylogenetic tree as coloured bars, using iTOL (Fig. 3.2).

This analysis revealed the striking observation that CRISPR clusters form phylogenetically distinct groups within the *Y. pseudotuberculosis* population. The highest diversity of CRISPR clusters was found in the 'Asian' phylogenetic clade, with at least 23 of the 33 identified CRISPR clusters represented in this clade. Conversely, the 'European' clade exhibited the lowest diversity of CRISPR clusters, with isolates belonging primarily to clusters 10, 11, 20, 22, and 38. This observation indicates a correlation between diversity of CRISPR clusters and genomic diversity within each geographic clade. The phylogenetic distribution of CRISPR cluster cassettes observed here may suggest that this clustering could be a result of strains isolated in the same short time span or localised source. However, closer examination of the CRISPR cluster pattern to isolation year and geographical source of isolation (Table 3.2) suggests that this is not the case. For strains with isolation dates available, the temporal sampling indicates that strains of these CRISPR clusters were generally isolated in separate time periods. Furthermore, there are instances where strains from clusters 9, 11, 12, 20, 22, and 38 have a temporal separation of at least a decade or more (Table 3.2).

To confirm the global geographical distribution of strains belonging to each of the 33 identified CRISPR clusters, prevalence of all CRISPR clusters were mapped to their respective countries of isolation (Fig. 3.3). The global map shows that CRISPR clusters are widely distributed across the world with some correlation to the phylogeographic split observed earlier in Figure 3.1. The highest diversity in CRISPR clusters occurs in the Far East of Asia, which is consistent with an Asian ancestry of *Y. pseudotuberculosis* (Achtman *et al*., 1999). The map also illustrates a clear coexistence of multiple CRISPR cluster-type strains in a number of different countries, suggesting that the formation of phylogenetically distinct CRISPR clades cannot be attributed to the

geographical separation of these strains. Given the previous observation that CRISPR evolution in the Enterobacteriaceae, in particular, is controlled by vertical and not horizontal evolution (Kupczok, Landan and Dagan, 2015), it is possible that the CRISPR-associated phylogenetic lineages observed in this study represent descent rather than due to CRISPRs restricting gene flow between strains. Some vertical descent may occur for independent reasons so, while CRISPR correlates with phylogeny, this does not necessarily mean that it is causal. This warranted further investigation into the formation of these CRISPR clades through phylogenetic dating and analysis of gene sharing in sections 3.3.4 and 3.3.5, respectively.

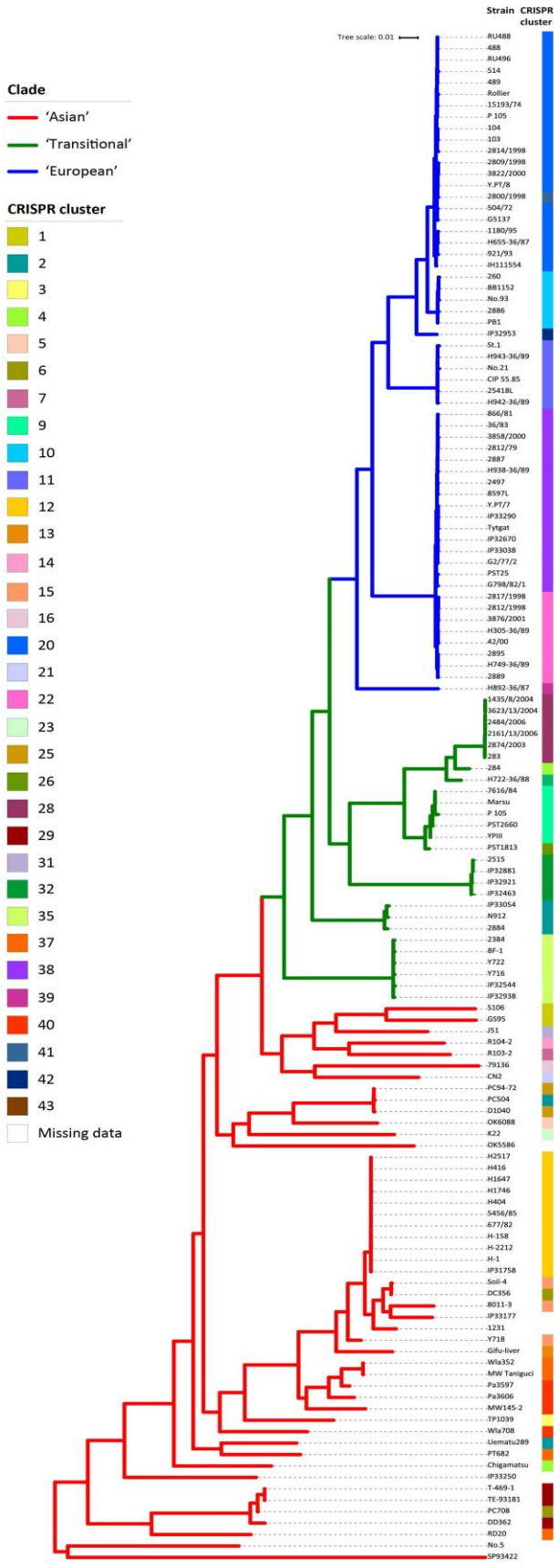


Figure 3.2. Maximum likelihood phylogenetic tree of 134 *Y. pseudotuberculosis* isolates annotated with the 33 identified CRISPR clusters.

CRISPR clusters are determined by Bayesian clustering of concatenated CRISPR spacer sequence arrays and are annotated on the phylogenetic tree as coloured bars using iTOL. The tree is rooted by midpoint rooting.

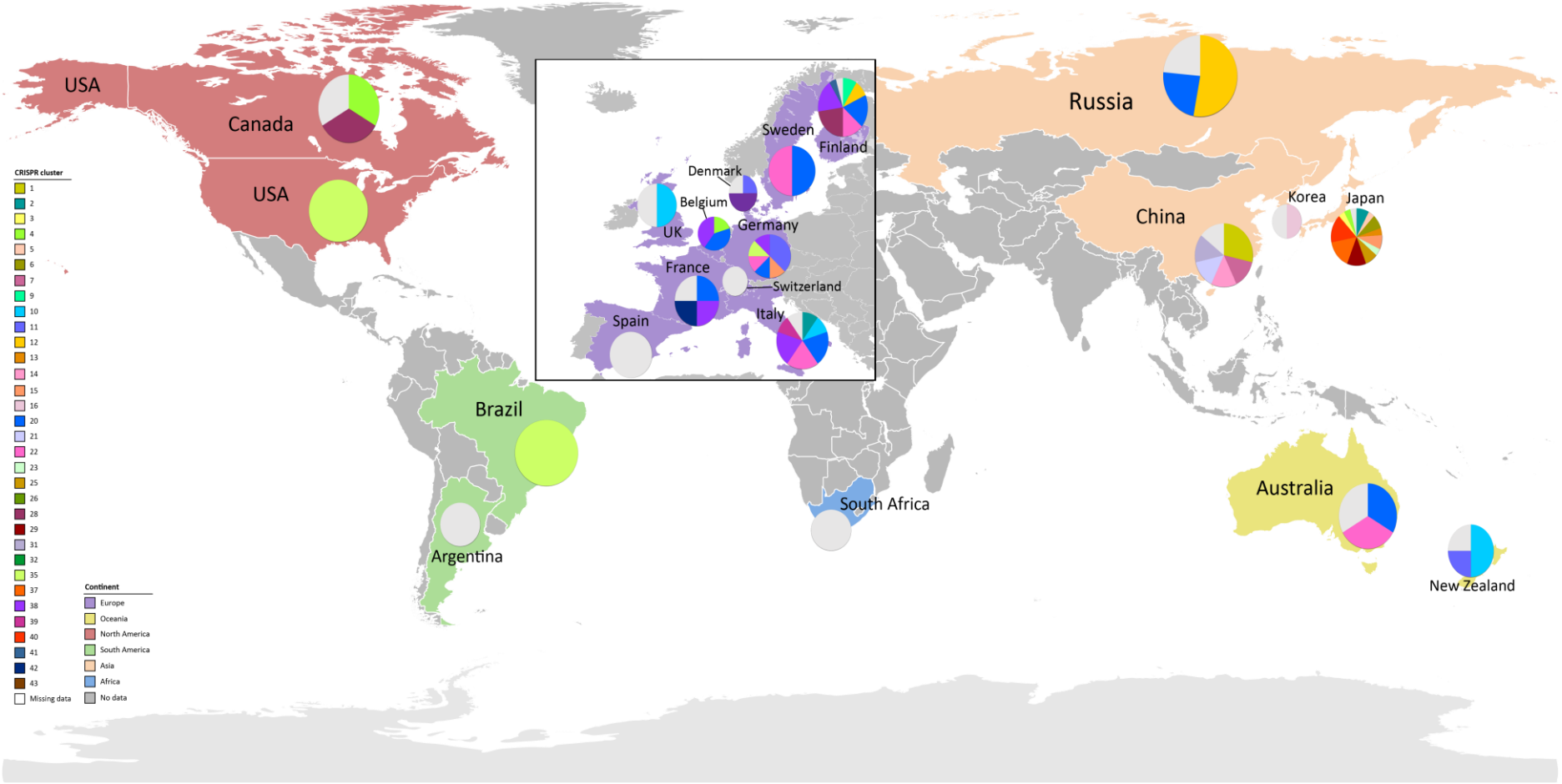


Figure 3.3. Global map showing the geographical sources of isolation of strains belonging to each of the 33 identified CRISPR clusters.

Thirty-three unique CRISPR cluster types were identified in the population. The pie charts represent the proportions of each CRISPR cluster type prevalent in each country. The map demonstrates a clear coexistence of multiple CRISPR cluster-type strains in a number of countries.

Table 3.2. CRISPR cluster-type *Y. pseudotuberculosis* strains with known isolation dates.

Strain name	CRISPR cluster	Country	Year
OK6088	5	Japan	1990
H722-36/88	8	Belgium	1988
7616/84	9	Finland	1984
Marsu	9	Finland	1980
P 105	9	-	1990
PB1	10	England	1960
H942-36/89	11	Germany	1989
H943-36/89	11	Germany	1989
CIP 55.85	11	-	1952
677/82	12	Finland	1982
5456/85	12	Finland	1985
IP31758	12	Russia	1966
Y718	15	Germany	1986
2809/1998	20	Finland	1998
3822/2000	20	Finland	2000
2814/1998	20	Finland	1998
15193/74	20	Finland	1974
H655-36/87	20	Germany	1987
504/72	20	Italy	1972
921/93	20	Sweden	1993
1180/95	20	-	1995
H305-36/89	22	Australia	1989
3876/2001	22	Finland	2001
2817/1998	22	Finland	1998
2812/1998	22	Finland	1998
H749-36/89	22	Germany	1989
42/00	22	Sweden	2000
2874/2003	28	Finland	2003
2161/13/2006	28	Finland	2006
2484/2006	28	Finland	2006
1435/8/2004	28	Finland	2004
3623/13/2004	28	Finland	2004
Y722	35	Germany	1988
PT682	37	Japan	1987
G798/82/1	38	Denmark	1982
3858/2000	38	Finland	2000
2812/79	38	Finland	1979
866/81	38	Finland	1981
36/83	38	Finland	1984
H938-36/89	38	Germany	1989
H892-36/87	39	Italy	1987
2800/1998	41	Finland	1998

For strains with isolation dates available, strains belonging to the same CRISPR cluster were generally isolated across different time periods, and in some cases, from separate geographical locales.

3.3.3. Cryptic ecology suggests that *Y. pseudotuberculosis* is a host generalist

In addition to the CRISPR cluster data, metadata concerning the serotype, country of origin, and host species/source of isolation were available for the majority of the *Y. pseudotuberculosis* population. With regards to O-antigen serotyping of these strains, a large proportion of the population comprise the O:1a, O:1b, and O:3 serotypes, which is consistent with previous studies indicating that the majority of *Y. pseudotuberculosis* strains isolated from human cases belong to these serotypes (Williamson *et al*., 2016; Laukkanen-Ninios *et al*., 2011). Other serotypes present in the population include serotypes O:2, O:4, O:5, and O:6, although these were lower in prevalence. All additional metadata were annotated onto the phylogenetic tree as coloured bars and the 'Asian' and 'European' clades were defined by tree branch colouring (Fig. 3.4). This analysis revealed that the 'European' clade is further split into distinct serotype O:1a and serotype O:1b phylogenetic clusters, further demonstrating the low diversity of this clade and indicating the predominance of these lineages among human-clinical and non-human environments in Europe. The 'Asian' clade is predominantly composed of strains belonging to serotype O:1b, whilst exhibiting the highest diversity of serotypes with O:2, O:4, O:5, and O:6 also present in this clade. With regards to the geographical distribution of serotypes, this is parallel to a previous study that revealed geographic heterogeneity between East Asian and Western European strains of *Y. pseudotuberculosis* (Fukushima *et al*., 2001). It was observed that almost all human-clinical strains from Europe belonged exclusively to serotypes O:1a and O:1b, whilst those from the Far East of Asia belonged to serotypes O:1b and a variety of subtypes of O:2–6 (Fukushima *et al*., 2001). Skurnik and co-authors (2000) suggested that the cryptic O-antigen gene cluster of *Y. pestis* biovar Orientalis showed that *Y. pestis* is most closely related to, and has evolved directly from, a *Y. pseudotuberculosis* serotype O:1b strain, isolated from a patient in Japan. This suggestion, in conjunction with the high diversity of O:1b strains originating from the Far East observed in this chapter (Fig. 3.4), would provide further evidence to support an ancestry of the *Y. pseudotuberculosis* species associated with Asia and the Far East, in particular.

Annotation of host species/isolate source on the phylogenetic tree did not reveal any clustering of strains within specific host groups (Fig. 3.4). This is indicative of the widespread ecology of *Y. pseudotuberculosis* across the globe, with isolates obtained from various sources including livestock and domesticated animals, wild animals and birds, fish, vegetables, and the environment, in addition to human disease cases. Isolates from all non-human sources were distributed throughout the phylogenetic tree and did not cluster separately to the human isolates, and the human sourced isolates did not cluster either. Although enteropathogenic *Y. pseudotuberculosis* and the closely related *Y. enterocolitica* are both aetiological agents for

human yersiniosis, there are notable differences in the ecologies of both organisms. Infections caused by *Y. enterocolitica* mainly originate from swine (Virtanen *et al*., 2013), but fresh produce, such as iceberg lettuce and raw carrots, has been the primary source for widespread *Y. pseudotuberculosis* outbreaks within recent decades (Kangas *et al*., 2008; Jalava *et al*., 2006). Despite the frequent presence of *Y. pseudotuberculosis* in environmental samples, its reservoir is considered to be wildlife (Niskanen, Fredriksson-Ahomaa and Korkeala, 2002). This would suggest that while *Y. enterocolitica* is more adapted to its ecological niche of swine, *Y. pseudotuberculosis* is more of a host generalist pathogen and has an ability to thrive and multiply in habitats such as the environment, from which it could easily disseminate into the food chain as a contaminant. Additionally, it would suggest that *Y. pseudotuberculosis* is capable of frequent host switching, and this is demonstrated in Figure 3.4, by strains belonging to the same serotype and CRISPR cluster that are isolated from multiple host species. *Y. pseudotuberculosis* is also common in pork meat, and in this study, isolates from swine samples are also prevalent in the population. Several strains obtained from swine samples (PC94-72, T-469-1, IP32544, and IP32670) have clustered together with strains from other domesticated and wild animal samples, in each of the 'Asian', 'Transitional', and 'European' clades of the population (Fig. 3.4). This is consistent with a recent study that has demonstrated through comparative genomic hybridisation analysis that European *Y. pseudotuberculosis* strains from swine cluster together with strains from human and wildlife samples (Jaakkola, Somervuo and Korkeala, 2015). This is further supported by a previous study by Niskanen *et al*. (2002), who reported on the homogeneity of *Y. pseudotuberculosis* strains isolated from swine samples based on pulsed-field gel electrophoresis analysis. Of importance in the population of the current study is perhaps the prevalence of human-clinical isolates from multiple countries among each geographic clade. This may suggest that human *Y. pseudotuberculosis* infection occurs sporadically, affecting multiple nations in different time periods. Given that *Y. pseudotuberculosis* is widespread among different ecological niches, the distribution of human strains across the phylogeny may suggest that human populations, through domestication of animals and movement across the globe, have provided a vector for dissemination allowing *Y. pseudotuberculosis* to reach new habitats in Europe and the West. Although the distribution of *Y. pseudotuberculosis* is worldwide, it is clear from the data set that most strains originate from regions of the Northern hemisphere, such as Europe, North America, Russia, China, and Japan. This may be indicative of under-sampling from countries of the Southern Hemisphere, but the emergence and dispersal of successful lineages in Europe and North America would have been dependent on the species' ability to adapt to significantly different conditions compared to those of its Asian origin, implying that *Y. pseudotuberculosis* has evolved to become a cold-tolerant species.

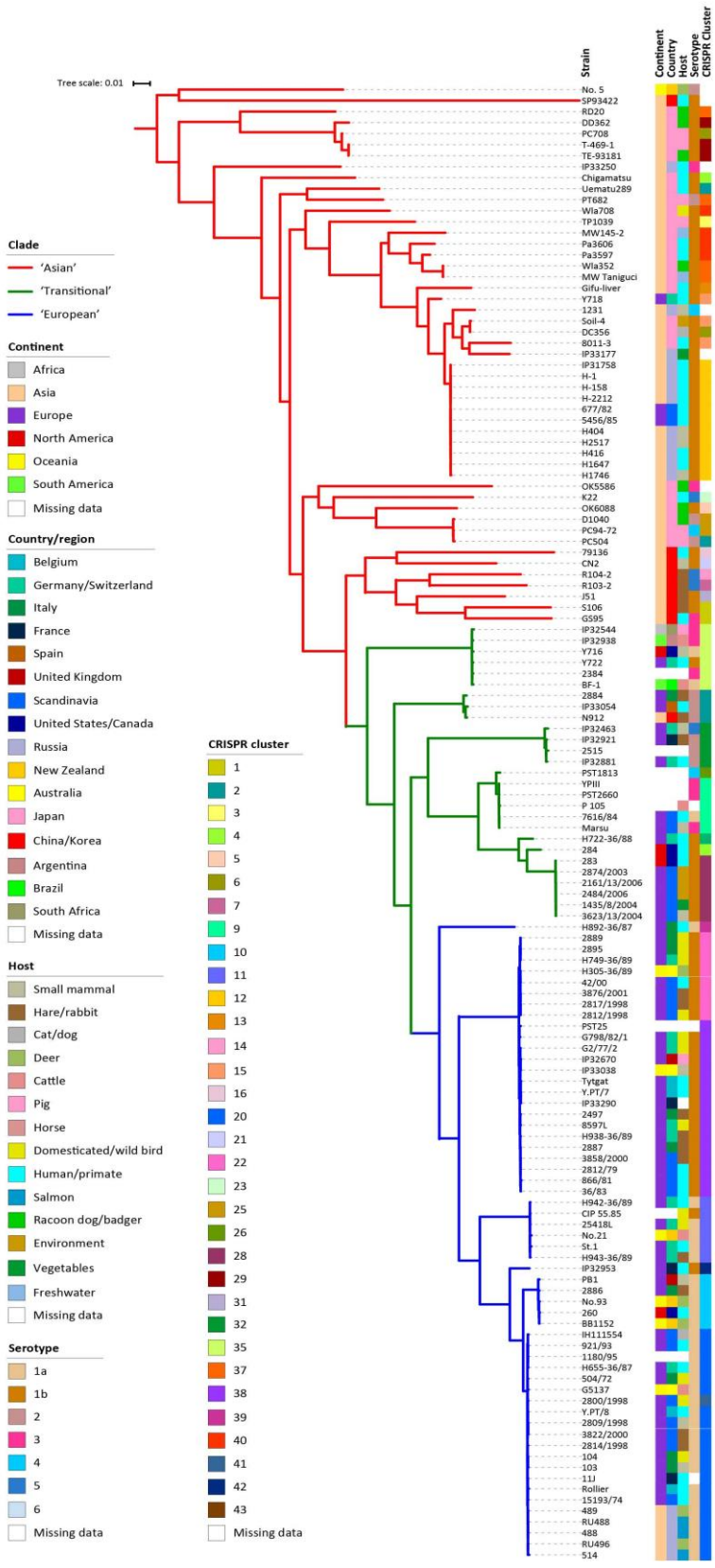


Figure 3.4. Maximum-likelihood phylogenetic tree of 134 *Y. pseudotuberculosis* isolates annotated with all available metadata. Additional metadata regarding host species/source, serotype, and country of origin were annotated on the tree as coloured bars using iTOL. No phylogenetic grouping is associated with ecological patterns (host species/isolate source). The 'European' clade is split into serotype 1a and serotype 1b clusters.

3.3.4. Phylogenetic dating suggests recent geographical divergence

Of the 134 *Y. pseudotuberculosis* genomes sequenced, isolation dates were available for 46 isolates representing the full diversity of the phylogeny. To date the evolutionary split of the 'European' clade of isolates from the 'Asian' clade, the BEAST 2 program (Bouckaert *et al*., 2014) was run by Alan McNally (University of Birmingham), analysing only the 46 strains for which isolation dates were available. The resulting dated maximum clade credibility (MCC) tree (Fig. 3.5A) suggests a time to the most recent common ancestor (TMRCA) for the data set of 33,591 years before the present. Error bars for each node, representing the upper and lower values within the 95% HPD (highest probability density) from the BEAST analysis, are displayed in the MCC tree of Appendix 4. The tree also suggests that the divergence of the 'European' and 'Asian' clades occurred approximately 12,500 years ago, which in the context of the TMRCA for the data set represents a recent phylogeographic split (Fig. 3.5A). With regards to human history, this period coincides with the end of the last ice age, during the transitionary period between the Neolithic and Mesolithic eras (Achtman, 2017). Migration of ancestral *Y. pseudotuberculosis* from Asia to Europe, and then further dissemination towards the West, would also appear to correlate with the beginning of livestock domestication and wheat and barley farming, as human populations migrated across the globe during this time period, generally favouring a system of nomadic agriculture. Running the BEAST 2 analysis also produced a Bayesian Skyline reconstruction of the *Y. pseudotuberculosis* population (Fig. 3.5B). The Skyline plot suggests that the population size remained stable over time but it also supports the possibility of a strong bottleneck occurring in the population within the 'European clade', in the very recent past (i.e. in the last few hundred years), owing to the estimated population size reduction indicated in the figure. This is consistent with the reduction in diversity and establishment of only a small number of clones in Europe.

We also sought to determine a TMRCA for the CRISPR-associated clades with isolation dates available. The CRISPR cluster patterns identified in Figure 3.2 were annotated onto the dated MCC tree (Fig. 3.5A) allowing an estimation of the time for the emergence of these clusters. The most recent of these clusters has a TMRCA to the rest of the population of approximately 5,222 years before present, suggesting that this clustering is not a recent phenomenon, nor is it due to any temporal artefacts of the sampling. This corroborates the comparison of the CRISPR cluster pattern to isolation years in Table 3.2, and it also indicates that these CRISPR-associated clades have existed over a considerable period of time, with the oldest of these clusters being of predominantly Asian origin and the more recently emerged clusters largely of European origin.

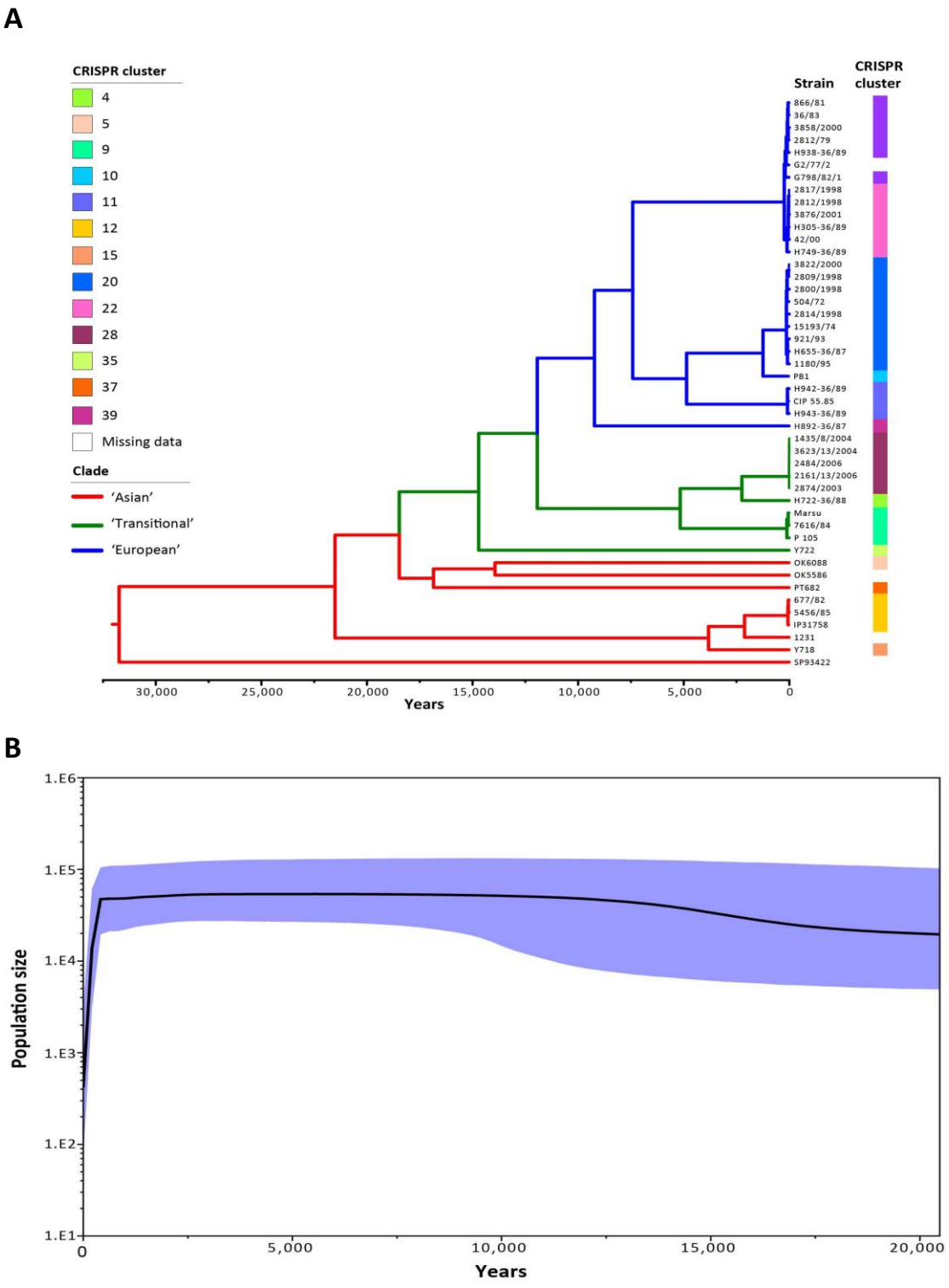


Figure 3.5. (A) Dated maximum clade credibility (MCC) tree produced from BEAST 2 analysis performed on the 46 *Y. pseudotuberculosis* strains for which isolation dates are available. (B) Bayesian Skyline reconstruction from BEAST 2 analysis. The dating analysis was performed by Alan McNally (University of Birmingham), using the BEAST 2 program with all known dates of isolation used to date individual taxa. The MCC tree (A) was produced and visualised using FigTree. The tree is annotated with CRISPR clusters as determined by Bayesian clustering of concatenated CRISPR sequence arrays. The Skyline plot (B) shows the estimated effective population size through time, as inferred by the Skyline demographic model.

3.3.5. CRISPR-associated phylogenetic clusters are associated with patterns of accessory gene conservation and core genome recombination

Phylogenetic distinctions were observed between the CRISPR clusters of the *Y. pseudotuberculosis* population analysed in this chapter, based on core genome SNP analysis. Further comparative genomic analyses, taking consideration of the accessory and pan-genomes of these strains, would provide a high-resolution assessment of the extent of homogeneity between strains of the same CRISPR cluster. Considering the role of CRISPR in generating acquired immunity to foreign DNA across bacteria, an investigation was carried out to determine whether the CRISPR clusters within *Y. pseudotuberculosis* were associated with any signature of gene sharing. Firstly, a pan-genome matrix was created for all 134 genomes using the LS-BSR pipeline (Sahl *et al*., 2014), and then the accessory genomes were extracted from the pan-genome matrix by running the post-matrix Python script filter_BSR_variome.py, which is provided by LS-BSR. Genes prevalent in > 90% of strains, and also those found in fewer than 5 strains, were excluded as they were not of interest for this analysis. The resulting accessory genome matrix was visualised by generating a heat map using the ggplot2 package of the R statistical software (http://www.R-project.org/; R Core Team, 2015; Wickham, 2016), allowing the presence and absence of every distinct genetic locus in the *Y. pseudotuberculosis* accessory genome to be displayed. The accessory gene presence/absence matrix was then used to annotate the core phylogenetic tree alongside the CRISPR cluster patterns (Fig. 3.6). From this analysis, it is noticeable that there are clear patterns within the accessory genome profiles which are concordant with the pattern of CRISPR clusters on the phylogenetic tree. Some of the most discernible patterns are highlighted on the phylogenetic tree, which include CRISPRs 9, 10, 11, 12, 20, 22, 28, and 35. This observation would suggest that for the majority of CRISPR clusters, *Y. pseudotuberculosis* strains have conserved unique combinations of accessory genes according to the CRISPR cluster they belong to. Despite coexisting with other CRISPR cluster-type strains in the same ecological niches and in close geographical proximity, where opportunities for gene sharing to occur would be frequent, it suggests that horizontal gene transfer between different CRISPR cluster-type strains is highly restricted.

Another interesting observation was made in the accessory gene profiles with respect to the clusters of strains representing CRISPRs 22 and 35. These strains exhibit much smaller accessory genome profiles when compared to the rest of the population. This observation is exemplified by these clusters of strains lacking a number of genes that are present in the majority of the population, which can be seen from ~100–300 on the accessory loci index (x–axis). Assembly quality analysis of these strains indicate that they are good quality and are complete sequenced genomes, with an average N50 of 116,376 bp and average genome size and GC content of 4.5

Mbp and 47.5%, respectively, which is within the range of what is expected for *Y. pseudotuberculosis* (genome sizes range from 4.3–4.8 Mbp and GC content ranges from 47–48%). Furthermore, no genomes had dropped-out; all genomes were within the average range for genome size, N50 and GC content. The disparity in accessory genome content when compared to the rest of the population cannot therefore be due to poor assembly quality. Bacterial pathogens often have smaller genomes and fewer genes than their non- or less-pathogenic relatives. The same pattern is seen within some bacterial species, with repeated transitions to higher virulence associated with reductive genome evolution (Weinert and Welch, 2017). With this in mind, the smaller accessory genomes exhibited by the strains of CRISPR clusters 22 and 35 may suggest that these strains have undergone significant accessory gene-loss events in comparison to the rest of the population, perhaps indicating an evolutionary shift to more pathogenic lineages.

To further investigate the congruence between CRISPR cassette clusters and accessory gene patterns, the Python script compare_BSR.py, of LS-BSR (Sahl *et al*., 2014), was used to identify genes unique to any given CRISPR cluster, as well as genes unique to either the 'European' clade of strains or the 'Asian' clade of strains. Attempts made to identify any unique genes in each CRISPR cluster and geographic clade were largely unsuccessful. Only one unique coding sequence (CDS) was detected in the 'European' clade of strains and no unique CDSs in the 'Asian' clade. Of CRISPR clusters that contained more than one representative strain, CRISPR 22 contained nine unique CDSs relative to all other CRISPR clusters within the population. Other clusters include CRISPR 28, with two unique CDSs, and CRISPRs 1, 9, and 11 each with one unique CDS, when compared to the rest of the population. Unique genes could not be identified in the remainder of CRISPR clusters within the population. Overall, this analysis suggests that each distinct CRISPR cluster contains a combination of accessory genes that are unique to that cluster, rather than unique individual genes within each cluster. To confirm this, the average accessory genome dissimilarity matrix for all detected CRISPR clusters containing more than one strain (18 out of 33 clusters) was calculated by Jukka Corander (University of Oslo). A standard permutation test was then performed to assess the significance of the observed dissimilarity pattern. This analysis confirmed that in 12 out of 18 CRISPR clusters, strains have significantly more similar gene profiles to strains in the same cluster than to strains in other clusters ($p < 0.05$ based on 10,000 random permutations). Based on this evidence, it would suggest that gene sharing between strains in the *Y. pseudotuberculosis* population is largely restricted to within individual CRISPR clades.

To investigate this further, BratNextGen analysis was run on core genome alignments of the *Y. pseudotuberculosis* population to detect recombination events between the core genomes of

all strains. The detected recombination events are displayed against the proportion of shared ancestry (PSA) tree annotated with CRISPR cluster patterns (Fig. 3.7). This analysis identified a distribution of core genome recombination events which is highly concordant with the pattern of CRISPR clustering. Despite very high levels of recombination being detected across the data set, the recombination occurring is not eroding the CRISPR cluster signal. This would suggest that inter-cluster horizontal transfer of genetic material is largely inhibited, or occurs at very low frequency compared to intra-cluster recombination events, consistent with the CRISPR cluster signature observed in the accessory gene profiles (Fig. 3.6). To support this, the amounts of intra-cluster and inter-cluster recombination detected from the BratNextGen analysis were quantified (Supplementary Data, doi: 10.1099/mgen.0.000133), and it was revealed that intra-cluster recombination events generally outweigh the amount of recombination occurring between strains of different CRISPR clusters (Seecharran *et al*., 2017). For example, higher numbers of intra-cluster recombination events were detected for various CRISPR clusters when compared to the average inter-cluster recombination occurring for those clusters. This includes, but is not limited to, CRISPR clusters 12 (intra-cluster = 311, average inter-cluster = 18), 20 (intra-cluster = 282, average inter-cluster = 14), and 28 (intra-cluster = 274, average inter-cluster = 16). Another interesting observation can be made from the BratNextGen analysis (Fig. 3.7), where no observable core genome recombination events were detected for a European cluster of strains represented by CRISPR clusters 22 and 38. This would suggest that strains of these clusters have perhaps 'locked' their genomes from transferring and receiving genetic material to/from other CRISPR cluster-type strains.

Overall, determining the accessory gene profiles and core genome recombination events for this globally and temporally distributed population of *Y. pseudotuberculosis* has demonstrated a strong association between CRISPR and the restriction of both accessory and core gene exchange between different CRISPR clusters. The distinct phylogroup structure of *Y. pseudotuberculosis* is therefore maintained throughout the population over a considerable period of time. This is consistent with a previous study which demonstrated that CRISPR systems play an important role in shaping the accessory genomes of the model antibiotic-refractory pathogen *Pseudomonas aeruginosa* (van Belkum *et al*., 2015). Furthermore, phylogenetic analysis based on CRISPR sequences of *Shigella* genomes, carried out by Yang and co-workers (2015), revealed a correlation between CRISPR loci and phylogenetic structure, as CRISPR sequences were conserved among subtypes of this genus. This is parallel to the observation of CRISPR-associated phylogenetic clades of *Y. pseudotuberculosis* revealed in the present study. Additionally, homology analysis of spacers showed that CRISPR might be involved in the regulation of virulence transmission (Yang *et al*., 2015), which would suggest that CRISPR

analysis is applicable for investigation of evolutionary relationships of bacterial pathogens and the formation of distinct microbial lineages, and this warrants further study across bacteria.

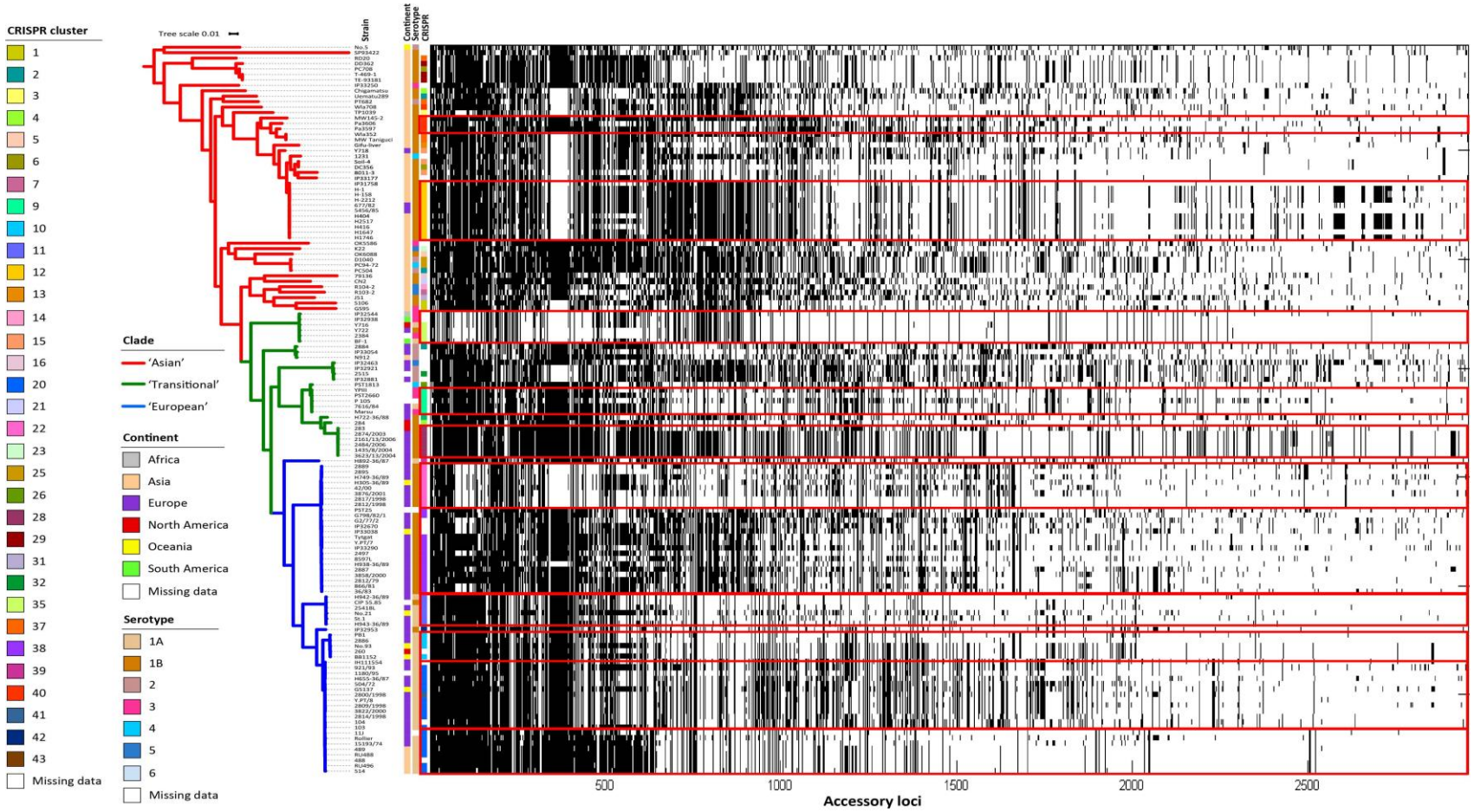


Figure 3.6. Distribution of accessory gene profiles for 134 isolates of *Y. pseudotuberculosis*. The genes (x-axis) have been sorted by their presence/absence pattern (black, present; white, absent) across strains (y-axis), which have been sorted according to the maximum likelihood phylogenetic tree, shown on the right. CRISPR clusters, continent of origin and serotype are annotated on the tree as coloured bars. The 'Asian' and 'European' clades are defined by tree branch colouring.

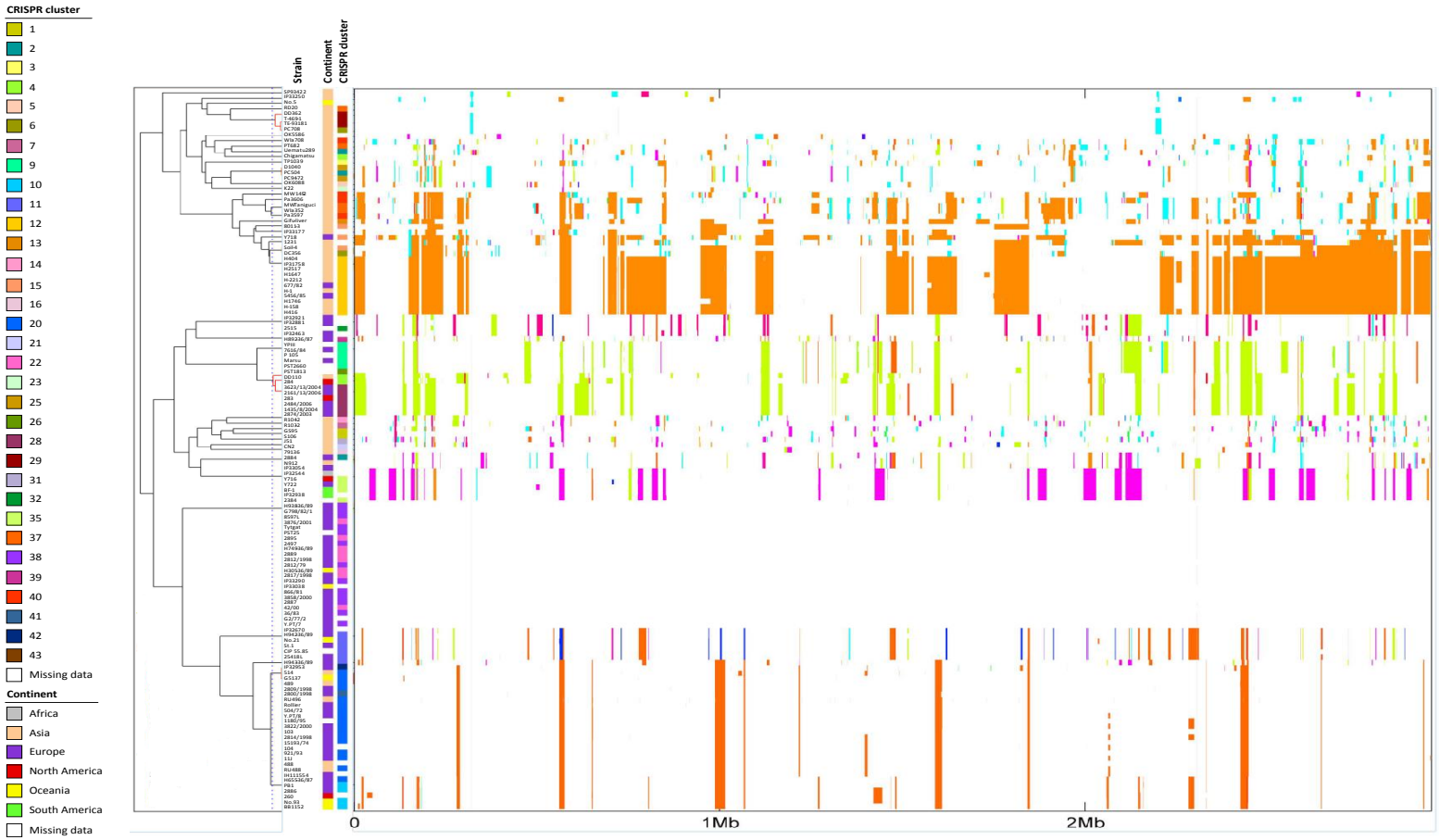


Figure 3.7. BratNextGen analysis of core genome recombination events for 134 isolates of *Y. pseudotuberculosis*. BratNextGen was run by Alan McNally. The PSA tree of 134 isolates is shown on the left. Horizontal coloured bars show the indicated recombination events for each strain (y-axis), at the relative base pair positions (x-axis). Where segments are the same colour in vertically overlapping positions, this indicates recombination events that are shared between those strains. The colours of the detected segments indicate the cluster in which the segment is most prevalent. CRISPR clusters and continent of origin are annotated on the tree as coloured bars.

3.4. Conclusions

The genus *Yersinia* has acted as a model for developing our understanding of microbial pathogenesis, molecular microbiology, and microbial ecology and evolution (McNally *et al*., 2016b). *Yersinia* was the first bacterial genus to have all representative species sequenced, allowing fine-scale analysis of how pathogenesis evolved in the three human pathogenic members of the genus (Reuter *et al*., 2014). This analysis revealed a striking degree of parallelism in how human pathogenesis evolved in pathogenic *Yersinia* (Reuter *et al*., 2014). However, further fine-scale evolutionary genomic studies of *Y. pestis* and *Y. enterocolitica* have shown very distinct mechanisms of intra-species evolution. *Y. pestis* is a recently evolved clone of *Y. pseudotuberculosis*, which is globally disseminated and host-restricted with very low levels of diversity, allowing fine-scale transmission events to be successfully reconstructed (Morelli *et al*., 2010). In complete contrast to this, pathogenic *Y. enterocolitica* have evolved from a non-pathogenic ancestor and have split into ecologically distinct clades, which move rapidly across host species (Reuter *et al*., 2015). To enhance our understanding of the population structure, ecological dissemination, and evolutionary events that define the model bacterial species *Y. pseudotuberculosis*, it was imperative to carry out purpose-designed, large-scale global population genomic analyses of *Y. pseudotuberculosis* in this study.

By sequencing a globally and temporally distributed set of 134 *Y. pseudotuberculosis* genomes, isolated from a wide range of hosts and environments, it can be shown that evolution in this species is driven by completely different mechanisms from those seen in the other pathogenic *Yersiniae*. This study revealed that *Y. pseudotuberculosis* is the only pathogenic *Yersinia* species which shows a clear phylogeographic split in its population. This was once postulated to be the case for *Y. enterocolitica* (Wang *et al*., 2011) with Old World and New World strains, however, comprehensive population genomics have shown this is not the case (Reuter *et al*., 2015; Reuter *et al*., 2014). The indication of Asian ancestry for *Y. pseudotuberculosis* is consistent with an Asian ancestry of *Y. pestis* (Morelli *et al*., 2010; Achtman *et al*., 1999), though the greatest amount of genetic variation was found in Japan rather than China. This is interesting because of the fact that a sub-clade of *Y. pseudotuberculosis* exists which causes Far East scarlet-like fever and is associated with Japan and tropical South-East Asia (Eppinger *et al*., 2007), suggesting larger variation in this region and a potential focus of ancestry for the species. Although it is difficult to accurately date the phylogeny of a data set with only a relatively small number of isolation dates available, the TMRCA for the entire *Y. pseudotuberculosis* data set is in the same range (10,000 – 40,000 years before present) as that estimated for the emergence of *Y. pestis* (Achtman *et al*., 1999). Based on the evidence from this dating analysis, it is inviting

to speculate that the emergence of *Y. pestis* coincided with a larger population dispersal event across *Y. pseudotuberculosis*.

Previous work by our research group analysed patterns of accessory gene sharing between serotype-specific clades of *Y. enterocolitica*, and concluded that the species is composed of ecologically distinct phylogroups (Reuter *et al*., 2015). This inference was made on the basis that the limited inter-clade sharing of genes could not be due to steric hindrance by the O-antigens nor genetic exclusion, as no such mechanisms existed. Data from the present study also identify clearly distinct phylogenetic subgroups within the geographic clades of *Y. pseudotuberculosis*. These phylogroups have unique combinations of accessory genes with little variation in their accessory genomes, and a very similar pattern of core genome homologous recombination. Similar to *Y. enterocolitica*, it is highly unlikely that this might be a result of some factor which prevents physical contact between strains, given the limited variety of serotypes present in *Y. pseudotuberculosis* (Savin *et al*., 2014). Rather, analysis from this study presents a strong case for the role of CRISPR clusters in the formation of these phylogroups. The primary evidence for the active role of CRISPR in mediating this genetic restriction is the fact that different CRISPR cluster-type strains coexist in the same geographical locations. Given that *Y. pseudotuberculosis* is ubiquitous in nature, no active barrier precluding recombination would exist between strains occupying the same habitat. Therefore, it would be expected that the signal that identifies each CRISPR cluster to be eroded relatively quickly over time (Sheppard *et al*., 2008), resulting in a lack of clear phylogroup signatures (Dearlove *et al*., 2016; Sheppard *et al*., 2008). This would especially be the case given the high levels of recombination detected in the core genome of *Y. pseudotuberculosis*. However, as the CRISPR-associated phylogenetic clusters have coexisted in locations around the world, for approximately 5,000 years or more, and continue to display a clear signature of within-cluster similarity, the data suggest that the CRISPR system is strongly associated with restricting both accessory and core gene exchange between clusters and maintenance of the distinct *Y. pseudotuberculosis* phylogroups.

CRISPR has been shown to play a role in shaping the accessory genomes of *Pseudomonas aeruginosa* (van Belkum *et al*., 2015), and CRISPR analysis correlated with phylogenetic structure in a study of *Shigella* genomes (Yang *et al*., 2015). However, it is likely that the present study provides the first evidence of a possible causative link between CRISPR cassettes and the evolution of distinct phylogroups in bacterial pathogens. Bayesian analysis of core genome recombination events suggests that the influence of CRISPR is exerted at the level of horizontal gene transfer. Data from a previous study have shown that CRISPR evolution in bacteria, particularly in the Enterobacteriaceae, is controlled by vertical and not horizontal evolution (Kupczok, Landan and Dagan, 2015). When taken together, the data from the current study

create a hypothesis for *Y. pseudotuberculosis* evolution, whereby large population perturbations lead to the emergence of geographically isolated clones. During the early formation of these clones, exposure to geographically localised exogenous DNA – such as genetic elements present in plasmids and phage – creates a CRISPR array of immunity, which then tightly regulates the repertoire of genetic material that can be transferred and acquired from the gene pool. This can be explained by the CRISPR mechanism, through which, new spacers derived from the genome of the invading virus are incorporated into the CRISPR array by an unknown mechanism (Barrangou *et al*., 2007). During the crRNA biogenesis stage, a CRISPR precursor transcript is processed by Cas endoribonucleases within repeat sequences to generate small crRNAs (Brouns *et al*., 2008). During the targeting/interference stage, the match between the crRNA spacer and target sequences (complementary protospacer) specifies the nucleolytic cleavage of the invading nucleic acid. As each of these nascent clones then globally disseminate, they encounter other clones of *Y. pseudotuberculosis* and coexist together in geographical isolation, but genetic transfer between different clones is restricted according to the CRISPR array of each clone. Clones comprising the same CRISPR array can acquire and transfer genes without any restriction. However, transfer of genetic material between different clones cannot occur at levels required to erode the clonal phylogenetic structure within the population, and consequently, distinct phylogroups of *Y. pseudotuberculosis* persist in the population.

In reference to the phylogeographic split observed in the *Y. pseudotuberculosis* population, the highest diversity of serotypes and CRISPR clusters was detected among 'Asian' strains, in contrast to the noticeably lower levels of diversity observed among strains of the 'European' clade. This would suggest that as the species migrated towards Europe, from its Asian origin, a bottleneck event occurring in the recent past would have resulted in the successful establishment of a small number of clones in new ecosystems, and thus lead to subsequent dispersal of these clonal lineages into Europe and the rest of the world. Analysis of all metadata available for the *Y. pseudotuberculosis* population, indicated that clones which have successfully disseminated into Europe belong almost exclusively to serotypes O:1a and O:1b, consistent with a previous study (Fukushima *et al*., 2001), and in addition, these clones are largely associated with only a small number of CRISPR loci. A larger sample size of isolates would be required in future work, to determine whether similar levels of diversity in geographic clades are maintained in the wider population of *Y. pseudotuberculosis*. With regards to the ecology of *Y. pseudotuberculosis*, this study did not reveal any phylogenetic clustering of strains within specific host groups or sources. Isolates from all non-human sources were distributed throughout the phylogeny and did not cluster separately from human isolates. This is indicative

of the widespread ecology of *Y. pseudotuberculosis* and its ubiquitous nature in many different non-human hosts and environments. The absence of genetic patterns associated with the ecology of *Y. pseudotuberculosis*, revealed in this chapter, is in contrast to previous studies that have reported ecological barriers to gene flow and recombination, which exist within populations of important bacterial pathogens. Such examples would include *Y. enterocolitica* (Reuter *et al*., 2015), *E. coli* (McNally *et al*., 2013; Luo *et al*., 2011), and *C. jejuni* (Sheppard *et al*., 2014), for which, the formation of distinct, ecologically separated lineages have been demonstrated. Contrary to these species, which comprise ecotypes formed through restricted recombination due to ecological barriers, *Y. pseudotuberculosis* is a host generalist species, capable of frequent host switching and occupation of multiple ecological niches, from which it could easily become a contaminant of the food chain. The cryptic ecology of *Y. pseudotuberculosis* may explain the ability of this species to colonise multiple hosts, suggesting that host generalism can also be a successful ecological strategy for bacterial pathogens. By analysing a set of geographically and temporally dispersed *Y. pseudotuberculosis* genomes in this chapter, the results indicate that the observed phylogenetic structure of the *Y. pseudotuberculosis* population is driven by factors other than those that preclude physical contact. The data from this study highlight how CRISPR can be used to infer the evolutionary trajectory of bacterial lineages, and they show that the evolution and ecology of *Y. pseudotuberculosis* differs from that seen previously in the other two human pathogenic species of the genus *Yersinia*.

CHAPTER 4

Defining the population structure of *Escherichia coli* from non-human sources

4.1. Introduction

Although *Escherichia coli* were once thought of as clonal organisms, they are now recognised as a bacterial species of extreme heterogeneity – a result of very high levels of recombination within the accessory genome (Wirth *et al*., 2006). *E. coli* are generally categorised into three main groups: commensal *E. coli*, intestinal pathogenic *E. coli*, and extraintestinal pathogenic *E. coli* (ExPEC). From this level of classification, *E. coli* are traditionally further subdivided into pathotypes on the basis of their isolation source and the possession of certain 'virulence-associated genes' (VAGs). Intestinal infections, characterised mainly by severe diarrhoea, are caused by pathotypes such as enteropathogenic *E. coli* (EPEC). Within ExPEC, the pathotype of *E. coli* which causes urinary tract infections (UTIs) are termed uropathogenic *E. coli* (UPEC) on the basis that they were isolated from the bladder and possess UPEC-specific VAGs (Janke *et al*., 2001). *E. coli* causing disease in birds are termed avian pathogenic *E. coli* (APEC), and those causing neonatal meningitis are termed 'NMEC'.

For many years, the accepted dogma for *E. coli* sub-classification was that the combination of virulence genes possessed by an organism influences pathogenic potential (Vejborg *et al*., 2011). Some studies have shown that there is little correlation between the disease caused by *E. coli* and its genotype. Strains that belong to different phylogroups can occupy distinct ecological niches and display diverse properties or ability to cause infections (Clermont *et al*., 2013). Due to the era of enhanced phylo-typing and sequence-typing methods, we now know that *E. coli* strains, even within a single pathotype, can vary immensely in terms of their evolutionary descent, which in turn can affect pathogenic potential and fitness in an infection (Wirth *et al*., 2006; Picard *et al*., 1999). Extraintestinal pathogenic *E. coli* infections are the most common cause of hospital-acquired infections in the UK and also cause significant levels of community-acquired infections (Woodford *et al*., 2004). Human ExPEC strains have often been characterised at the sequence type (ST) level and a small number of STs, namely ST69, ST73, ST95, and ST131, were found to be the most predominant among cases of UTIs and bloodstream infections (Kallonen *et al*., 2017; Riley, 2014; Alhashash *et al*., 2013; Croxall *et al*., 2011b). These clonal groups are known to constitute highly antimicrobial-resistant strains. Multidrug resistance (MDR) mediated by ESBLs is of increasing global concern in *E. coli*, because strains tend to harbour various resistance genes, in particular CTX-M-15 and 14 (Nicolas-Chanoine *et al*., 2008). These genes are primarily plasmid-borne, which facilitates the transfer of resistance determinants to other strains, species, or genera (Woodford, Turton and Livermore, 2011). Moreover, infections caused by multidrug-resistant bacteria can lead to inadequate or delayed antimicrobial therapy and increased costs associated with treatment (Magiorakos *et al*., 2012).

Using both phenotypic and molecular methods, many previous studies have demonstrated that both ExPEC and MDR strains of *E. coli* can be detected in various environments other than the human intestinal tract. These would include rivers and other water sources, soils, various wild and domesticated animals, including food animals, as well as raw meat and poultry (Gomi *et al*., 2017b; Johnson *et al*., 2017; Muller, Stephan and Nuesch-Inderbinen, 2016; Vincent *et al*., 2010; Jakobsen *et al*., 2010). Although most environmental and non-human *E. coli* strains are thought to be harmless, some strains have demonstrated pathogenic potential, and thus contamination of surface waters and retail meat by such strains can increase the risk of waterborne and foodborne diseases. Also of major concern is the prevalence of antimicrobial-resistant and MDR *E. coli* detected in these non-human reservoirs (Gomi *et al*., 2017b; Johnson *et al*., 2017). The hypothesis that human ExPEC and MDR *E. coli* strains may have a non-human reservoir in surface waters and food animals has been an area of study by various research groups worldwide (Kappell *et al*., 2015; Dolejska *et al*., 2011b; Dolejska *et al*., 2011a). Experimental studies reporting shared pathogenicity between human ExPEC and avian pathogenic *E. coli* (APEC) suggests that these extraintestinal *E. coli* strains may share a common ancestry and evolutionary roots with APEC (Logue *et al*., 2017). Several previous studies have consistently reported detection of specific human ExPEC strains in poultry or retail poultry meat products, but rarely in other meat products. This would appear to support the hypothesis that a poultry reservoir for human ExPEC may exist. These studies have regularly identified *E. coli* ST131 and other human-clinical associated ExPEC lineages, such as ST69, ST394, ST95, ST10 and ST117, in cases of human extraintestinal infection as well as in retail meat products and other food animals (Bergeron *et al*., 2012; Vincent *et al*., 2010). In Sweden, *E. coli* sequence types ST69, ST10, and ST117 represented half of the ESBL-producing *E. coli* isolated from retail chicken meat in this country. Moreover, there are several global environmental studies reporting the presence of MDR *E. coli* and potential pathogenic ExPEC strains in rivers and surface waters, which would support the hypothesis that water-related environments may be considered one of the important non-human reservoirs of human ExPEC (Gomi *et al*., 2017b; Muller, Stephan and Nuesch-Inderbinen 2016; Kappell *et al*., 2015). It has been suggested that contamination of surface waters by *E. coli* strains belonging to clinically important clonal groups may increase the risk of waterborne disease, because surface waters are used for sources of drinking water, irrigation, and recreational purposes (Gomi *et al*., 2017b).

Despite evidence from previous studies suggesting that surface waters and retail poultry may provide a potential non-human reservoir for human ExPEC, few of these studies have been able to determine the prevalence of potential pathogenic human ExPEC and MDR *E. coli*, in the context of the wider population of *E. coli* found in these sources. Information regarding the fine-

scale phylogeny, clonal composition and resistance determinant profiles of an unbiased sample of *E. coli* isolated from non-human environments, such as surface waters and retail poultry, is lacking in the current literature. The true population structure of the non-human population of *E. coli* in the wider environment has therefore not been adequately characterised. Several studies have been conducted addressing the potential for foodborne and waterborne transmission of ExPEC, but the majority of these studies are focussed on MDR *E. coli*, and more specifically, ESBL-producing ExPEC (Manges 2016; Lazarus *et al*., 2015; de Been *et al*., 2014). Numerous environmental studies enrich for antimicrobial-resistant isolates, using selective media prior to phenotypic and genotypic characterisation, thus their sampling strategies may bias towards MDR and ExPEC strains. Studies of ESBL-positive ExPEC lineages from surface waters and retail poultry therefore tend to be over-represented in the literature. To date, only one large-scale population study, based on whole-genome analysis of the prevalence of antimicrobial-resistant and extraintestinal pathogenic *E. coli* strains in river water, has been conducted (Gomi *et al*., 2017b). Similarly, a previous study by de Been and colleagues (de Been *et al*., 2014) implemented WGS analyses to study the relatedness of cephalosporin-resistant *E. coli* from humans, retail chicken meat, poultry and pigs. This study provided high resolution differentiation between human and poultry-associated isolates and suggested that there is little or no overlap between resistant *E. coli* isolates of human and poultry origin. A review of the current global evidence, implicating poultry meat as a potential reservoir for human ExPEC and MDR *E. coli*, has suggested the need for more whole-genome-based and comparative genomic analyses of *E. coli* populations recovered from food animals and retail meat products with human-clinical strains (Manges, 2016).

4.1.1. Aim and objectives

In order to infer ecological patterns of *E. coli* from non-host-associated habitats, such as the environment and food products, more phylogeny-based population genomic analyses must be conducted, similar to what was implemented recently in *Y. pseudotuberculosis* (Seecharran *et al*., 2017) and the *E. coli* ST131 lineage (McNally *et al*., 2016a). The current study employs an unbiased sampling procedure, by not selectively isolating antimicrobial-resistant strains. Coupling this strategy with whole-genome analysis of *E. coli* isolated from river water and retail chicken meat in the Nottingham area, it allows a snapshot to be constructed of the relative abundance of MDR *E. coli* and ExPEC strains in the wider non-human population of *E. coli*, which is largely unknown. This would provide great insight into determining the true population structure of non-human *E. coli* in aquatic environments and in the food chain. By determining the population structure of non-human *E. coli* isolated specifically from the Greater Nottingham

area in this chapter, it would allow for a geographically constrained comparison with the well-characterised human-clinical population of *E. coli*, to be conducted in chapter 5. There are comprehensive phenotypic and genotypic data available for human-clinical *E. coli* isolates collected in this region over the past decade (Alhashash *et al*., 2016; Alhashash *et al*., 2013; Croxall *et al*., 2011b; Croxall *et al*., 2011a), and therefore, Nottingham provides the ideal ecosystem for generating a non-human *E. coli* data set for comparison with the human-clinical population of *E. coli*.

Specific objectives of this chapter were:

- To identify and isolate a population of *E. coli* from river water and retail chicken meat sampled in Nottingham and sequence the whole genomes of a non-biased representative proportion of the population.
- To assess the level of diversity within the non-human population of *E. coli*, as determined by *in silico* multilocus sequence typing analysis of whole-genome sequence data.
- To reconstruct the phylogeny of the non-human population of *E. coli* and determine the population structure and genotypic diversity, with regards to the prevalence of *E. coli* phylogroups.
- To detect antimicrobial resistance determinants within bacterial strains isolated from river water and retail chicken samples, in order to create a snapshot of the prevalence of potential multidrug-resistant strains within the non-human population of *E. coli*.
- To determine the prevalence of human ExPEC strains within the non-human population of *E. coli*, as defined by the presence of ExPEC-specific virulence-associated genes.

4.2. Materials and Methods

The key methods, culture media, culturing conditions, and bioinformatics tools and scripts used in this chapter were described previously in sections 2.3 – 2.6 of chapter 2. The publicly available *E. coli* reference genomes used in this chapter are listed in Table 4.1.

Table 4.1. *E. coli* reference genomes used for phylogrouping analysis of non-human *E. coli* strains.

Strain	Phylogroup	ST	Country	Sample	Reference
TW10509	C-I	747	India	Human	(Kaas *et al*., 2012)
TW15838	C-I	-	Australia	Environment	(Luo *et al*., 2011)
B1147	C-II	-	Australia	Bird	(Walk *et al*., 2009)
TW09276	C-III	-	United States	Environment	(Luo *et al*., 2011)
TW11588	C-IV	2	Puerto Rico	Environment	(Luo *et al*., 2011)
TW14182	C-IV	-	United States	Environment	(Luo *et al*., 2011)
TW09308	C-V	-	United States	Environment	(Luo *et al*., 2011)
P12b	A	10	-	-	(Ratiner, 1985)
DH1	A	1060	-	-	(Kaas *et al*., 2012)
IAI1	B1	1128	France	Human	(Kaas *et al*., 2012)
SE11	B1	156	Japan	Human	(Kaas *et al*., 2012)
55989	B1	678	CAR*	Human	(Kaas *et al*., 2012)
SE15	B2	131	Japan	Human	(Kaas *et al*., 2012)
JJ1886	B2	131	United States	Human	(Owens *et al*., 2011)
536	B2	127	-	Human	(Kaas *et al*., 2012)
UTI89	B2	95	-	Human	(Kaas *et al*., 2012)
MS_85-1	C	88	United States	Human	(Kaas *et al*., 2012)
TW14425	C	23	United States	Human	(Kaas *et al*., 2012)
042	D	414	Peru	Human	(Kaas *et al*., 2012)
UMN026	D	597	United States	Human	(Kaas *et al*., 2012)
EDL933	E	11	United States	Food	(Kaas *et al*., 2012)
Sakai	E	11	Japan	Human	(Kaas *et al*., 2012)
IAI39	F	62	France	Human	(Kaas *et al*., 2012)

Publicly available genomes were downloaded from the National Center for Biotechnology Information (NCBI) website (https://www.ncbi.nlm.nih.gov/genome/). These reference genomes belonging to known phylogenetic groups were used to assign *E. coli* strains from non-human samples to one of 7 *E. coli* phylogroups (A, B1, B2, C, D, E, F) or one of 5 *Escherichia* cryptic clades (C-I, C-II, C-III, C-IV, C-V). Reference strains were selected from the genomes analysed in previous studies, indicated in the 'Reference' column.

*CAR: Central African Republic

4.3. Results and Discussion

4.3.1. Prevalence of *E. coli* isolated from river water and retail chicken meat

Nine river water samples were collected in July 2015 from 4 geographically separate locations within the Trent River basin (Nottinghamshire and Derbyshire) and were tested for the presence of *E. coli*. Samples were taken from upstream and downstream sites of waste water and sewage treatment plants as well as upstream and downstream sites of agricultural land – specifically cattle farms. A total of 20 British whole retail chickens, obtained from 6 major supermarket outlets in the Greater Nottingham area in October 2015, were also sampled for *E. coli* (detailed information on sampling is provided in sections 2.3.1 and 2.3.2 of chapter 2). *E. coli* present in the collected samples were identified via a process involving subculture onto differential chromogenic agar (CLED agar with Andrade's indicator and HiCrome™ UTI agar) and biochemical identification tests (as described in section 2.3.3). *E. coli* were identified in 6 out of 9 total river water samples and 11 out of 20 total retail chicken samples that were processed. Isolates were selected from the initial CLED plates based on colony morphology which resembled that of *E. coli* (as described in Table 2.3). From the 504 river water isolates that were selected from CLED plates, a total of 82 isolates (16%) were formally identified as *E. coli*. The majority of isolates from river water samples selected for identification testing (83%) were represented by other species of the Enterobacteriaceae, which were therefore excluded from further analyses in this study. Conversely, a high proportion (88%) of the 416 retail chicken isolates selected from CLED plates were confirmed as *E. coli*, and thus only a minority (12%) represented other species of Enterobacteriaceae. Because of the high prevalence of *E. coli* in retail chicken samples, the number of isolates to be included in the study population was limited to 148. This resulted in a study population of 230 isolates, across 29 river water and retail chicken samples, that were identified as *E. coli* with ≥ 80% confidence, according to the API 20E biochemical test system. Specific numbers of *E. coli* isolates identified from CLED plates for each respective sample of river water and retail chicken are given in Table 4.2 and Table 4.3, respectively.

The presence of *E. coli* in freshwater, such as river water, can indicate contamination by animal or human faeces. This is because *E. coli* colonise the gastrointestinal tracts of a wide range of wild and domesticated animals, especially animals raised for human consumption, such as chickens. Contamination of retail chicken with pathogens such as *E. coli* can occur at multiple steps along the food chain, including production, processing, distribution, or packaging and retail marketing. A previous study had reported that a large proportion of chicken meat imported into Sweden that was contaminated with ESBL-producing *E. coli*, primarily ST10,

ST131, and ST69, was found to have spread from imported parent broilers to broiler meat (Egervarn *et al*., 2014). This indicates that the occurrence of these drug-resistant ExPEC lineages in chicken meat is likely due to faecal contamination at the slaughterhouse, which is at a very early stage of the retail meat production process. In terms of *E. coli* isolated from river water in this study, 70% of the population were obtained upstream and downstream of wastewater treatment plants at Giltbrook and Pinxton. This would indicate that treated effluents from sewage treatment plants would have a high influence on freshwater contamination by faecal pathogens, more so than rural runoff from farmland. A possible explanation for the difference in *E. coli* prevalence observed between retail chicken and river water samples could be due to the effectiveness of wastewater treatment processes, in Nottinghamshire and Derbyshire, to remove faecal contaminants before releasing effluents into streams and rivers. On the other hand, during the process of preparing British chickens for human consumption, from farm and abattoir to retail outlet, it would appear that contamination by faecal organisms is much harder to control.

Table 4.2. Total number of *E. coli* isolates identified from river water samples.

Sample name	Number of *E. coli* isolates selected from CLED plates
Giltbrook upstream	35
Giltbrook downstream	14
Erewash Pinxton downstream	2
Erewash Pinxton upstream	6
East Leake downstream	0
East Leake upstream	24
Keyworth upstream	0
Keyworth downstream	0
Keyworth cattle field puddle	1
Total	**82**

The numbers presented in the table correspond to the final count of *E. coli* isolates that were selected from CLED agar plates for each river water sample. A total of 82 *E. coli* isolates were obtained from river water samples. The majority of isolates were obtained from the Giltbrook and Erewash Pinxton samples. Giltbrook and Erewash Pinxton are sample sites near to wastewater treatment plants, whilst East Leake and Keyworth are sample sites near to cattle farms.

Table 4.3. Total number of *E. coli* isolates identified from retail chicken samples.

Sample name	Number of *E. coli* isolates selected from CLED plates
Tesco sample 1	18
Tesco sample 2	0
Tesco sample 3	10
Tesco free range sample	10
Sainsbury's sample 1	0
Sainsbury's sample 2	10
Sainsbury's free-range sample	10
Asda sample 1	0
Asda free range sample 1	10
Asda free range sample 2	10
Aldi sample 1	0
Aldi sample 2	0
Iceland sample 1	20
Iceland sample 2	20
Iceland sample 3	0
Iceland sample 4	0
Morrisons sample 1	0
Morrisons sample 2	15
Morrisons sample 3	15
Morrisons sample 4	0
Total	**148**

The numbers presented in the table correspond to the final count of *E. coli* isolates that were obtained from CLED agar plates for each retail chicken sample. Due to the large number of *E. coli* being obtained from retail chicken samples, a final reduced count of 148 isolates were selected for subsequent analyses.

4.3.2. Molecular detection of β-lactamase genes in non-human *E. coli*

Due to the reported high prevalence of ESBL-producing *E. coli* in retail meat products and environmental waters, and the hypothesis that these sources could represent reservoirs for human ExPEC and MDR *E. coli* (Gomi *et al*., 2017b; Manges 2016; Vincent *et al*., 2010), preliminary work on prevalence of β-lactamase and ESBL genes among *E. coli* isolated from river water and retail chicken samples was carried out in this chapter. This involved performing multiplex PCR assays to screen for the presence of the β-lactamase genes bla_{TEM}, bla_{SHV}, $bla_{CTX\text{-}M}$ and bla_{OXA} (Fig. 4.1). To put the prevalence of β-lactamase genes in non-human *E. coli* in the context of human-clinical *E. coli*, percentage prevalence data for β-lactamase and ESBL gene carriage in 415 *E. coli* strains, isolated from human-clinical cases in Nottingham, were included in the analysis for comparison (Fig. 4.2). One hundred and fifty of these strains were isolated from cases of urinary tract infection and urosepsis in elderly patients (Croxall *et al*., 2011b); 140 and 125 strains were isolated from bacteraemia patients and urinary samples, respectively, as part of a separate Nottingham-based study (Alhashash *et al*., 2013). The overall prevalence of the β-lactamase gene SHV (0.4%) and the ESBL gene CTX-M (5.2%), detected in *E. coli* from non-human samples, was significantly lower than the prevalence of those genes in *E. coli* from human-clinical samples (8.9% and 19.5% respectively; $p < 0.0001$, two-tailed Fisher's test). The most commonly detected gene in non-human *E. coli* was the TEM β-lactamase (38.7%), although prevalence of this gene is still significantly higher in human-clinical isolates of *E. coli* (52%; $p < 0.001$, two-tailed Fisher's test). The high prevalence of the TEM β-lactamase gene in non-human and human clinical samples is in agreement with observations that TEM is the most frequently encountered β-lactamase in clinical Enterobacteriaceae, as it accounts for around 80% of all plasmid-encoded β-lactamases (Bajpai *et al*., 2017; Paterson and Bonomo, 2005). The OXA ESBL gene went undetected in the non-human population, whereas this gene was detected in 9.9% of human-clinical *E. coli* isolates ($p < 0.0001$, two-tailed Fisher's test). The rationale for carrying out molecular detection of β-lactamase and ESBL genes, prior to whole-genome sequencing, is that a preliminary snapshot of the prevalence of MDR *E. coli* in the non-human population could be compared to that of the well-characterised human-clinical population. In addition, this work gave an insight into the potential prevalence of MDR plasmids within the population, which would inform whether *in silico* analyses of plasmid DNA from WGS data should be carried out. Considering the very low prevalence of ESBL genes, particularly of the CTX-M type, it is likely that the prevalence of MDR plasmids circulating within the non-human population is also very low. Based on this observation, it was decided that comparative plasmid analyses for the human-clinical and non-human populations of *E. coli* would not be carried out in this study.

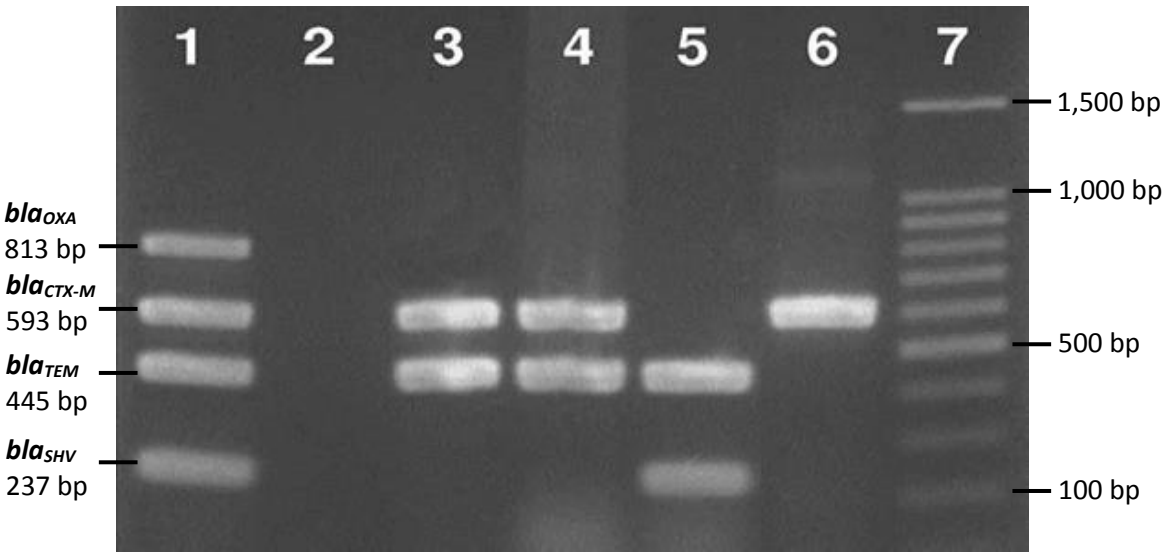


Figure 4.1. Electrophoresis gel showing PCR amplicons of β-lactamase genes detected in *E. coli* isolated from non-human samples.

The genes encoding the SHV, TEM, CTX-M, and OXA β-lactamases were amplified from *E. coli* isolates from river water and retail chicken samples by multiplex PCR, using previously published primers (Table 2.4). PCR amplicons were electrophoresed on a 2% agarose gel and fragment sizes were checked against a 100 bp DNA ladder (New England Biolabs). Lane 1 contains the positive control strain: *Klebsiella pneumoniae* UTI448, bla_{SHV}^{+}, bla_{TEM}^{+}, $bla_{CTX\text{-}M}^{+}$, bla_{OXA}^{+}. Lane 2 contains the negative control. Lanes 3–6 contain a selection of retail chicken isolates from this study: AFR-6 (lane 3), SFR-9 (lane 4), I2-20 (lane 5), and M3-22 (lane 6). Lane 7 contains the 100 bp molecular weight marker.

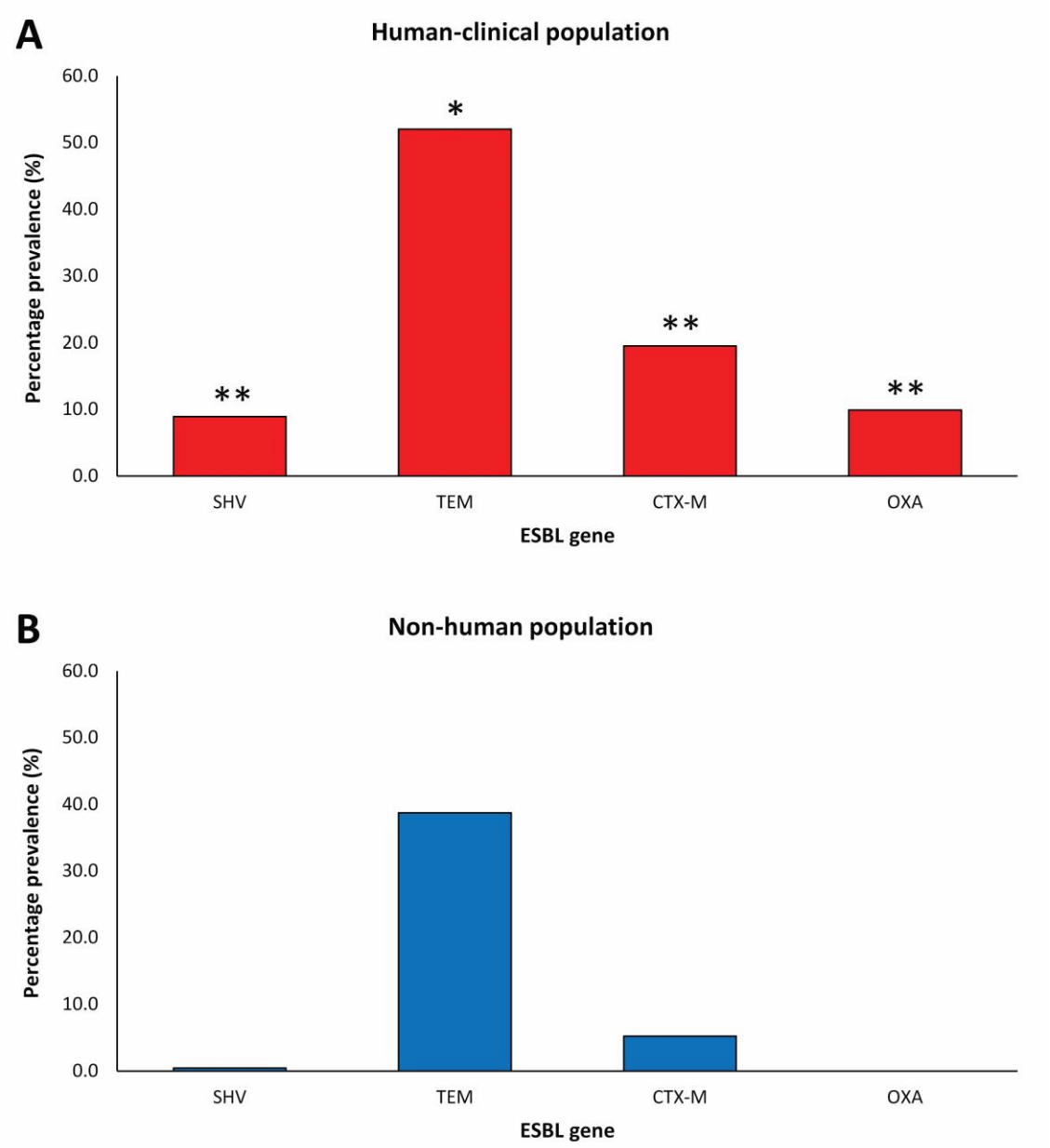


Figure 4.2. Percentage prevalence of β-lactamase genes *bla*$_{TEM}$, *bla*$_{SHV}$, *bla*$_{CTX-M}$, and *bla*$_{OXA}$ in *E. coli* isolates from human-clinical (A) and non-human (B) samples collected in Nottingham.

The human-clinical population of *E. coli* presented here (n = 415) comprise 150 isolates from cases of UTIs in 150 elderly patients (Croxall *et al*., 2011b), as well as 140 isolates from cases of bacteraemia and 125 isolates from urine samples across diverse patient groups (Alhashash *et al*., 2013). The non-human population (n = 230) represent 82 *E. coli* isolates from river water samples and 148 isolates from retail chicken samples obtained in this study. The prevalence of SHV, CTX-M, and OXA genes in *E. coli* isolated from human-clinical samples was significantly higher than the prevalence of those genes in *E. coli* from non-human samples (**human-clinical vs non-human; $p < 0.0001$, two-tailed Fisher's test). Although TEM was the most commonly detected β-lactamase gene in non-human *E. coli* isolates, the prevalence of this gene is still significantly higher in human-clinical isolates of *E. coli* (*human-clinical vs non-human; $p < 0.001$, two-tailed Fisher's test).

4.3.3. Whole-genome sequencing of non-human *E. coli* isolates

To provide insight into the previously undefined population structure of non-human *E. coli*, as well as confirm the presence of putative resistance genes and identify ExPEC-associated virulence determinants within the population, whole-genome sequencing was performed on *E. coli* isolated from non-human sources in this study. Due to budget constraints, 180 *E. coli* isolates were sequenced from the population of 230 *E. coli* sampled from non-human sources, so as to represent the full diversity of the study population with regards to sample source. Indexed and paired-end libraries were prepared using the Nextera XT DNA Library Preparation Kit and the libraries were sequenced using the Illumina MiSeq. *De novo* assemblies of raw reads into contigs and scaffolds were performed using the SPAdes assembler. Assembly statistics were obtained from running the QUAST quality assessment tool and details of N50 and L50 values, genome size, GC content, and number of Ns per 100 kbp are provided in Appendix 5. N50 values indicate the length for which the collection of all contigs of that length or longer covers at least half an assembly. To eliminate any incomplete assemblies from further genomic analyses, N50 values of at least 1,900 bp and genome sizes of at least 4.3 Mbp were considered as the minimum criteria for assembled contigs to be used in genomic analyses. The first 12 assemblies highlighted in red in Appendix 5 indicate genome assemblies that were excluded from the study population. In addition to excluding incomplete sequenced genomes, all redundant isolates (i.e. identical isolates of the same sequence type, from the same sample, with the same antimicrobial resistance gene profiles) were also excluded from further genomic analyses. To determine redundant isolates, a multilocus sequence typing (MLST) script (https://github.com/tseemann/mlst) and the bioinformatics tool ABRicate (Kleinheinz, Joensen and Larsen, 2014) were run on the assembled genomes to assign a sequence type (ST) and generate *in silico* antimicrobial resistance gene profiles for each isolate. The MLST script scans the assembled scaffolds against PubMLST databases and assigns an ST to each isolate. All isolates, barring one, obtained from the same sample source, assigned to the same ST, and defined by the same antimicrobial resistance gene profile were omitted from the study population, as these were multiple isolates of a single strain and were therefore considered redundant for further genome comparative analyses. The remaining isolates to be included in the study population were chosen at random. This resulted in a study population inclusive of 128 non-redundant *E. coli* strains (Table 4.4), isolated from non-human sources, which were to be subjected to further genomic analyses. A summary of the number of *E. coli* isolates at each stage of the investigation is provided in Figure 4.3.

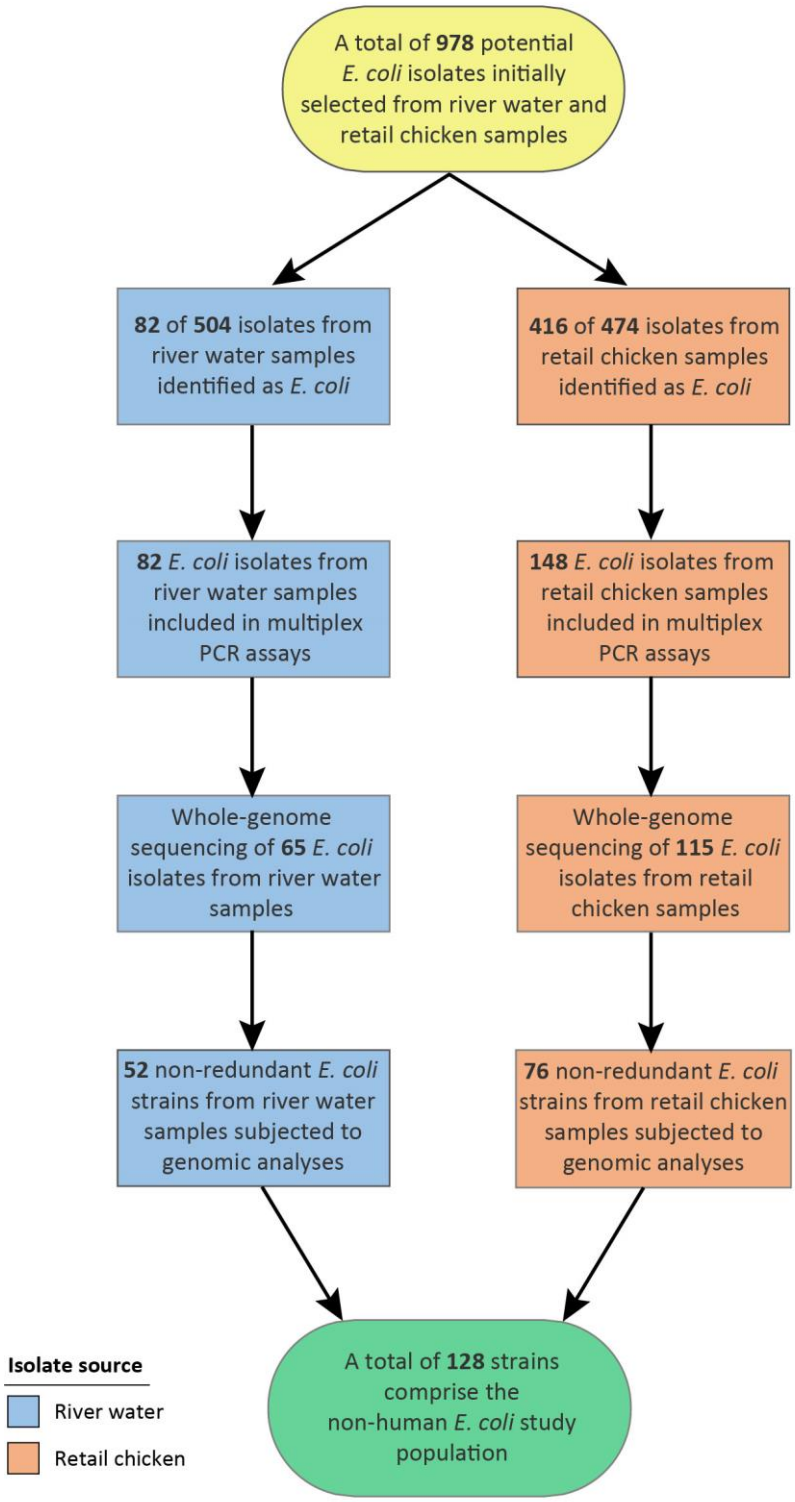


Figure 4.3. Workflow indicating the numbers of *E. coli* isolates from river water and retail chicken samples consolidated at each stage of the investigation.

The flowchart illustrates the respective numbers of *E. coli* isolates from river water and retail chicken samples at the stages of formal identification of *E. coli*, multiplex PCR for β-lactamase genes, whole-genome sequencing, and post-sequencing quality assessment of assemblies. The sample sizes were refined accordingly at each stage, resulting in a non-human *E. coli* study population of 128 strains.

4.3.4. Whole-genome-based multilocus sequence typing (MLST) analysis of the non-human population of *E. coli*

One method of determining the level of genetic heterogeneity within the non-human population of *E. coli*, isolated from river water and retail chicken samples in Nottingham, was to determine the prevalence of sequence types (STs). The study population was subjected to *in silico* MLST analysis and ST designations were obtained for all 128 strains, by running a MLST script which scans the genome assemblies against PubMLST databases, based on the seven *E. coli* housekeeping genes (*adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, *recA*), and assigns an ST to each strain. Closely related STs were then grouped into ST complexes using the PHYLOViZ platform, and a complete minimum spanning tree (MST) of all the STs in the population was constructed (Fig. 4.4). The sequence types were clustered based on sharing allele types rather than weighting SNPs; therefore, the MST illustrates clusters of 'like-organisms' and does not attempt to infer phylogeny.

Sequence typing analysis revealed that the *E. coli* population isolated from non-human samples in this study is genotypically diverse, with a total of 64 STs identified among all 128 strains (Table 4.5). Considering the size of this study population, such a variety of different STs being identified from these samples would suggest considerable genotypic diversity among non-human *E. coli*, from river water and retail chicken samples. The ST designations are grouped into clonal complexes by their similarity to a central allelic profile (genotype). In the study population, 16 different ST complexes were identified, with 50 strains being grouped into one of the known *E. coli* ST complexes. However, the majority of strains (n = 78) were designated STs which are not grouped into larger clonal complexes of closely related STs, further demonstrating the genotypic diversity within the population. Determining the sequence types of bacterial strains can assist in the production of a population map, which can be used to analyse the genetic relatedness of a population of bacteria. The MST (Fig. 4.4) reveals that the river water and retail chicken populations of *E. coli* demonstrate relatively similar levels of diversity with regards to ST prevalence, with 38 STs identified among strains isolated from river water samples and 31 STs from retail chicken samples. Only 5 of the STs identified (ST10, ST93, ST746, ST752, and ST1551) were present in both river water and retail chicken samples, whereas 33 STs were exclusive to the river water population and 26 STs were exclusive to the retail chicken population. This illustrates a genotypic difference between *E. coli* strains isolated from the two sources, and it suggests that separate populations of *E. coli* exist in each environmental source (freshwater versus the food chain).

The population map (Fig. 4.4) also suggests that the non-human population of *E. coli* is not dominated by the major STs/ST complexes that are largely responsible for extraintestinal infections in human-clinical populations of *E. coli* (Kallonen *et al*., 2017; Alhashash *et al*., 2013; Croxall *et al*., 2011b; Lau *et al*., 2008). Rather, the non-human population of *E. coli* would appear to be composed largely of a wide variety of different STs, with the ST10 clonal complex emerging as the major central genotype, along with closely related STs such as ST93, ST746, ST752, and ST1551, which were identified in both types of non-human sources of *E. coli* analysed. This suggests a wider prevalence of these particular STs across the environment and in the food chain. The abundance of the ST10 clonal complex across river water and retail chicken samples observed in this study is expected as several previous studies have reported ST10 as the most prevalent genotype in retail chicken meat, other meat types, and environmental waters (Gomi *et al*., 2017b; Chen *et al*., 2016; Cohen Stuart *et al*., 2012; Overdevest *et al*., 2011). Only one instance of the important pathogenic ExPEC lineage, ST131, was observed in the non-human population of *E. coli* in this study. The dearth of this particular sequence type, as well as the lack of other well-known ExPEC sequence types, such as ST95, ST73, and ST69, provides an indication of the prevalence of human ExPEC strains in the non-human population of *E. coli*, which appears to be very low. It also gives an insight into the phylogenetic structure of the population and suggests that the non-human population of *E. coli* is perhaps predominated by strains that are characteristically commensal or non-pathogenic.

Table 4.4. Sequence type designations for the 128 sequenced non-human *E. coli* genomes.

Strain name	ST	ST complex	Sample source	Sample name
AFR-4	10	10	Retail chicken	Asda free range 1
AFR-6	10	10	Retail chicken	Asda free range 1
AFR-12	10	10	Retail chicken	Asda free range 2
AFR-22	10	10	Retail chicken	Asda free range 2
ELU39	10	10	River water	East Leake upstream
ELU103	10	10	River water	East Leake upstream
GU34	10	10	River water	Giltbrook upstream
I1-16	10	10	Retail chicken	Iceland 1
I1-17	10	10	Retail chicken	Iceland 1
I1-19	10	10	Retail chicken	Iceland 1
I1-24	10	10	Retail chicken	Iceland 1
M2-2	10	10	Retail chicken	Morrisons 2
M2-3	10	10	Retail chicken	Morrisons 2
M2-4	10	10	Retail chicken	Morrisons 2
M2-8	10	10	Retail chicken	Morrisons 2
T1-61	10	10	Retail chicken	Tesco 1
GU53	20	20	River water	Giltbrook upstream
I1-21	48	10	Retail chicken	Iceland 1
I2-1	48	10	Retail chicken	Iceland 2
ELU7	58	155	River water	East Leake upstream
ELU21	58	155	River water	East Leake upstream
EPU62	58	155	River water	Erewash Pinxton upstream
I2-20	69	69	Retail chicken	Iceland 2
M3-27	69	69	Retail chicken	Morrisons 3
T3-3	69	69	Retail chicken	Tesco 3
T3-14	69	69	Retail chicken	Tesco 3
EPU17	93	168	River water	Erewash Pinxton upstream
M3-18	93	168	Retail chicken	Morrisons 3
S2-4	93	168	Retail chicken	Sainsbury's 2
S2-8	93	168	Retail chicken	Sainsbury's 2
TFR-1	93	168	Retail chicken	Tesco free range
GU48	108	None	River water	Giltbrook upstream
M3-24	115	None	Retail chicken	Morrisons 3
M3-28	115	None	Retail chicken	Morrisons 3
M3-30	115	None	Retail chicken	Morrisons 3
M3-34	115	None	Retail chicken	Morrisons 3
M3-36	115	None	Retail chicken	Morrisons 3
SFR-11	117	None	Retail chicken	Sainsbury's free range
T1-27	117	None	Retail chicken	Tesco 1
T1-30	117	None	Retail chicken	Tesco 1
T1-39	117	None	Retail chicken	Tesco 1
GD45	131	131	River water	Giltbrook downstream
GU50	135	None	River water	Giltbrook upstream
GU87	141	None	River water	Giltbrook upstream
T1-35	155	155	Retail chicken	Tesco 1

GU51	201	469	River water	Giltbrook upstream
EPD30	218	10	River water	Erewash Pinxton downstream
I1-11	354	354	Retail chicken	Iceland 1
I1-25	354	354	Retail chicken	Iceland 1
I2-5	354	354	Retail chicken	Iceland 2
I2-6	354	354	Retail chicken	Iceland 2
S2-3	362	None	Retail chicken	Sainsbury's 2
GU15	394	394	River water	Giltbrook upstream
EPU5	399	399	River water	Erewash Pinxton upstream
EPU51	399	399	River water	Erewash Pinxton upstream
GD3	409	None	River water	Giltbrook downstream
GD162	410	23	River water	Giltbrook downstream
GU35	446	446	River water	Giltbrook upstream
ELU87	537	14	River water	East Leake upstream
GU82	546	None	River water	Giltbrook upstream
ELU65	635	399	River water	East Leake upstream
GU1	635	399	River water	Giltbrook upstream
GD46	642	278	River water	Giltbrook downstream
GD109	644	538	River water	Giltbrook downstream
GD49	648	648	River water	Giltbrook downstream
T1-11	665	None	Retail chicken	Tesco 1
GU10	706	None	River water	Giltbrook upstream
ELU122	746	None	River water	East Leake upstream
SFR-6	746	None	Retail chicken	Sainsbury's free range
T1-1	746	None	Retail chicken	Tesco 1
ELU98	752	None	River water	East Leake upstream
T1-5	752	None	Retail chicken	Tesco 1
T1-25	752	None	Retail chicken	Tesco 1
T1-32	752	None	Retail chicken	Tesco 1
T1-53	752	None	Retail chicken	Tesco 1
T1-57	752	None	Retail chicken	Tesco 1
T1-73	752	None	Retail chicken	Tesco 1
T1-52	770	None	Retail chicken	Tesco 1
GU43	906	None	River water	Giltbrook upstream
GU6	929	None	River water	Giltbrook upstream
GU77	929	None	River water	Giltbrook upstream
EPD5	973	None	River water	Erewash Pinxton downstream
I1-5	997	None	Retail chicken	Iceland 1
I1-12	997	None	Retail chicken	Iceland 1
T3-21	1011	None	Retail chicken	Tesco 3
SFR-4	1112	None	Retail chicken	Sainsbury's free range
ELU28	1122	None	River water	East Leake upstream
GU24	1125	None	River water	Giltbrook upstream
ELU88	1276	None	River water	East Leake upstream
SFR-15	1408	None	Retail chicken	Sainsbury's free range
AFR-16	1551	None	Retail chicken	Asda free range sample 2
GU52	1551	None	River water	Giltbrook upstream
M3-22	1551	None	Retail chicken	Morrisons 3
S2-7	1551	None	Retail chicken	Sainsbury's 2

T1-3	1551	None	Retail chicken	Tesco 1
T3-1	1551	None	Retail chicken	Tesco 3
T3-7	1594	None	Retail chicken	Tesco 3
T3-18	1594	None	Retail chicken	Tesco 3
TFR-2	1594	None	Retail chicken	Tesco free range
TFR-15	1716	None	Retail chicken	Tesco free range
T1-7	1800	None	Retail chicken	Tesco 1
GU2	2136	None	River water	Giltbrook upstream
GD93	2178	None	River water	Giltbrook downstream
GU47	2178	None	River water	Giltbrook upstream
M2-1	2309	None	Retail chicken	Morrisons 2
M2-5	2309	None	Retail chicken	Morrisons 2
I2-18	2459	None	Retail chicken	Iceland 2
GU41	2521	None	River water	Giltbrook upstream
GU45	2521	None	River water	Giltbrook upstream
M3-29	2705	None	Retail chicken	Morrisons 3
ELU29	2722	None	River water	East Leake upstream
ELU34	3578	None	River water	East Leake upstream
GU13	4105	None	River water	Giltbrook upstream
GU27	4105	None	River water	Giltbrook upstream
GU31	4105	None	River water	Giltbrook upstream
GU70	4105	None	River water	Giltbrook upstream
GU80	4105	None	River water	Giltbrook upstream
S2-5	4243	None	Retail chicken	Sainsbury's 2
S2-10	4243	None	Retail chicken	Sainsbury's 2
T3-19	4243	None	Retail chicken	Tesco 3
TFR-6	4937	None	Retail chicken	Tesco free range
S2-2	4993	None	Retail chicken	Sainsbury's 2
T1-56	4994	None	Retail chicken	Tesco 1
GD138	5236	None	River water	Giltbrook downstream
TFR-13	5931	None	Retail chicken	Tesco free range
GU5	5995	None	River water	Giltbrook upstream
M2-7	6664	None	Retail chicken	Morrisons 2
T1-49	6664	None	Retail chicken	Tesco 1

Table 4.4. Sequence type designations for the 128 sequenced non-human *E. coli* genomes.

Of the 180 sequenced *E. coli* genomes obtained from river water and retail chicken samples, 128 non-redundant *E. coli* strains (i.e. strains meeting the minimum assembly quality criteria and are distinct in terms of ST or resistance gene profile, if isolated from the same sample) were included in the final study population of non-human *E. coli*. The population represents strains isolated from 5 different river water samples and 11 different retail chicken samples. Sequence types were determined for each strain by *in silico* MLST analysis.

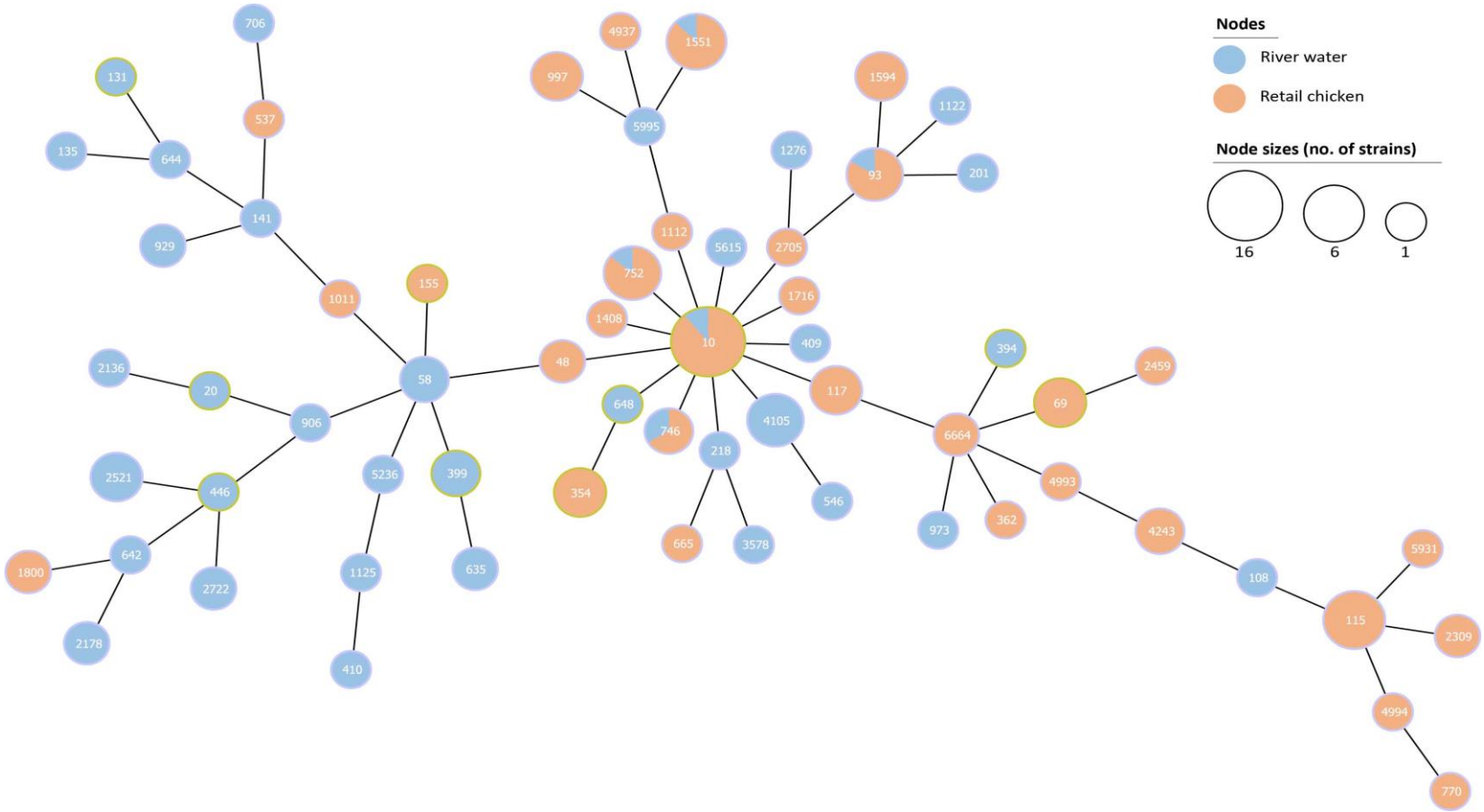


Figure 4.4. Minimum spanning tree (MST) illustrating STs of the non-human *E. coli* population isolated from river water and retail chicken samples.

The MST was produced using Phyloviz v3.0, which uses goeBURST to divide an MLST data set into groups of related isolates and clonal complexes. The size of the nodes reflects the number of strains belonging to each ST. Nodes outlined by a yellow-green ring represent ST complexes present in the population. The sample sources from which the strains were isolated are overlaid onto the diagram, which reveals a diversity of STs prevalent in both sample types.

4.3.5. Defining the phylogeny of the non-human population of *E. coli*

To assign each strain to a phylogenetic group, the core genomes of the 128 strains from the study population were aligned with those of 23 publicly available reference strains (Table 4.1), representing the seven *E. coli* phylogroups (A, B1, B2, C, D, E, and F) and five cryptic *Escherichia* clades (C-I, C-II, C-III, C-IV, and C-V). The core genome SNP-based phylogenetic tree of the non-human *E. coli* population was built using Parsnp, and phylogroup/cryptic clade assignment was made based on clustering with the reference strains on the phylogenetic tree (Fig. 4.5). From this analysis, it was revealed that 119 strains belonged to six of the seven *E. coli* phylogroups, with phylogroup C not being represented in the population. Furthermore, two strains were grouped with the sister *Escherichia* cryptic clade C-I, whereas the remainder of the strains belonged to the more distantly related cryptic clades, C-III (n = 1) and C-V (n = 6). These cryptic clades are novel lineages of the genus *Escherichia*, that are genetically distinct but phenotypically indistinguishable from *E. coli*. C-II and C-IV strains were not identified in the population, so the reference strains representing these clades, in addition to phylogroup C, were removed from the phylogenetic tree in an attempt to improve the resolution and interpretability of the tree. Cryptic clade C-I shares a recent common ancestor with the seven phylogroups comprising the majority of the population, illustrating the genetic similarity between this lineage and the *E. coli* phylogroups. In fact, according to the Clermont *E. coli* phylo-typing method, clade C-I is recognised as the eighth *E. coli* phylogroup (Clermont *et al*., 2013). This indicates how phylogenetically distinct the cryptic clades C-III and C-V are from the rest of the population, and they therefore serve as an outgroup for the population under investigation in this study, whilst allowing the tree to be rooted, as shown in Figure 4.5. Due to the presence of C-III and C-V strains in the population, a very small core genome alignment of 87,981 bp covering the 146 strains was achieved using Parsnp (≤ 5% of the reference genome covered by the alignment), however a total of 59,317 SNPs were identified across the alignment, further demonstrating the extreme genetic diversity that exists in the non-human population of *E. coli*.

With regards to the *E. coli* phylogroups present in the non-human population, it was found that the most highly represented group was phylogroup A, with 49 (38%) out of 128 strains belonging to this group. This was followed by 27 strains (21%) being grouped in phylogroup D, 18 (14%) in phylogroup B1, 10 (8%) in phylogroup E, 9 (7%) in phylogroup B2, and 6 (5%) in phylogroup F. Phylogroup B2 strains generally carry more virulence-associated genes than strains belonging to the other groups do (Picard *et al*., 1999), and strains that cause extraintestinal infections are predominantly associated with phylogroup B2 and, to a lesser extent, phylogroup D (Picard *et al*., 1999). Several studies have reported phylogroups A and B1 as being chiefly composed of commensal strains of *E. coli* (Picard *et al*., 1999, Duriez *et al*., 2001). Based on the phylogroups

represented in the non-human population of *E. coli* in this study, it would appear that the majority of the population constitutes largely commensal and non-pathogenic strains, as illustrated by the higher prevalence of phylogroup A and B1 strains in the population. Phylogroup D was found to be the second-most prevalent group in the population, so the presence of extraintestinal pathogenic *E. coli* (ExPEC) cannot be ruled out; however, due to the lack of phylogroup B2 strains in the population, it would be expected that the prevalence of ExPEC is low among isolates from river water and retail chicken samples. While there was no discernible pattern of phylogenetic grouping associated with source of isolation (i.e. no clear phylogenetic split), it was noticeable that certain phylogroups appeared to be dominated by strains isolated from a particular source (Fig. 4.5). This would include phylogroups B1, B2, and cryptic clades C-III and C-V, which are dominated by river water isolates, and phylogroups D and E, which are largely composed of isolates from retail chicken samples.

ST/ST complex designations representing three or more isolates (as typed by *in silico* MLST analysis) are overlaid onto the phylogenetic tree (Fig. 4.5), in addition to known representatives of ExPEC (ST131 and ST648) and EPEC (ST20) within the population, though these are represented only by a single isolate each. The distribution of STs across the non-human population of *E. coli* would suggest that phylogroup A comprises the highest diversity of strains, with 16 different STs identified within this group; the major sequence types being ST10, ST93, ST746, ST752, and ST1594 among them. Other sequence types of note in the population would include ST115 (n = 7) and ST117 (n = 4), of phylogroup D, which are both largely associated with wild birds and commercial poultry (Cristovao *et al*., 2017), and are commonly shared by APEC and human ExPEC strains (Maluta *et al*., 2014; Oteo *et al*., 2010). Close genetic relations have been detected in ST117 *E. coli* strains of animal and human origin, which have been identified in large poultry producing countries, such as Brazil (Maluta *et al*., 2014) and the USA (Danzeisen *et al*., 2013). Also of relevance in phylogroup D is ST69 (n = 5), which has been reported as a highly virulent strain in some animal models with a high content of resistance determinants (Cristovao *et al*., 2017; Tartof *et al*., 2005), and has also been associated with extraintestinal infections, such as UTIs and bacteraemia, in clinical case studies (Kallonen *et al*., 2017; Alhashash *et al*., 2013; Croxall *et al*., 2011b; Lau *et al*., 2008). Phylogroup F strains are also prevalent within the population, which is a group closely related to the ExPEC-associated phylogroups B2 and D (Clermont *et al*., 2013). Prior to its recognition as a distinct phylogroup, its members were generally categorised under group D, based on a PCR-based phylo-typing assay that delineates only four major *E. coli* phylogroups (Clermont, Bonacorsi and Bingen, 2000). However, an enhanced version of this assay (Clermont *et al*., 2013), and whole-genome-based *in silico* MLST analysis, enable differentiation of phylogroup F from phylogroup D. Within this group, one

instance of ST648 has been identified in the non-human population of *E. coli*. This genotype is reported increasingly as an emerging resistance-associated lineage (Pitout, 2012) and is distributed worldwide, occurring as a pathogen and commensal of humans and animals (whether food-producing, domesticated, or wild), and in the environment (Muller, Stephan and Nuesch-Inderbinen, 2016; Goncalves *et al*., 2016; Sato *et al*., 2014; Kang *et al*., 2013). Though the phylogenetic structure of the non-human population of *E. coli* demonstrates a paucity of pathogenic strains, that are predominantly associated with extraintestinal infections in humans, the presence of lineages such as ST131, ST69, and ST648 in the non-human population would warrant further investigation into the profiles of resistance determinants and virulence-associated genes that define each strain. This would allow for a determination of the prevalence of MDR and ExPEC strains among *E. coli* isolated from foodborne and environmental water sources, which will be compared with that of the human-clinical population of *E. coli* in chapter 5.

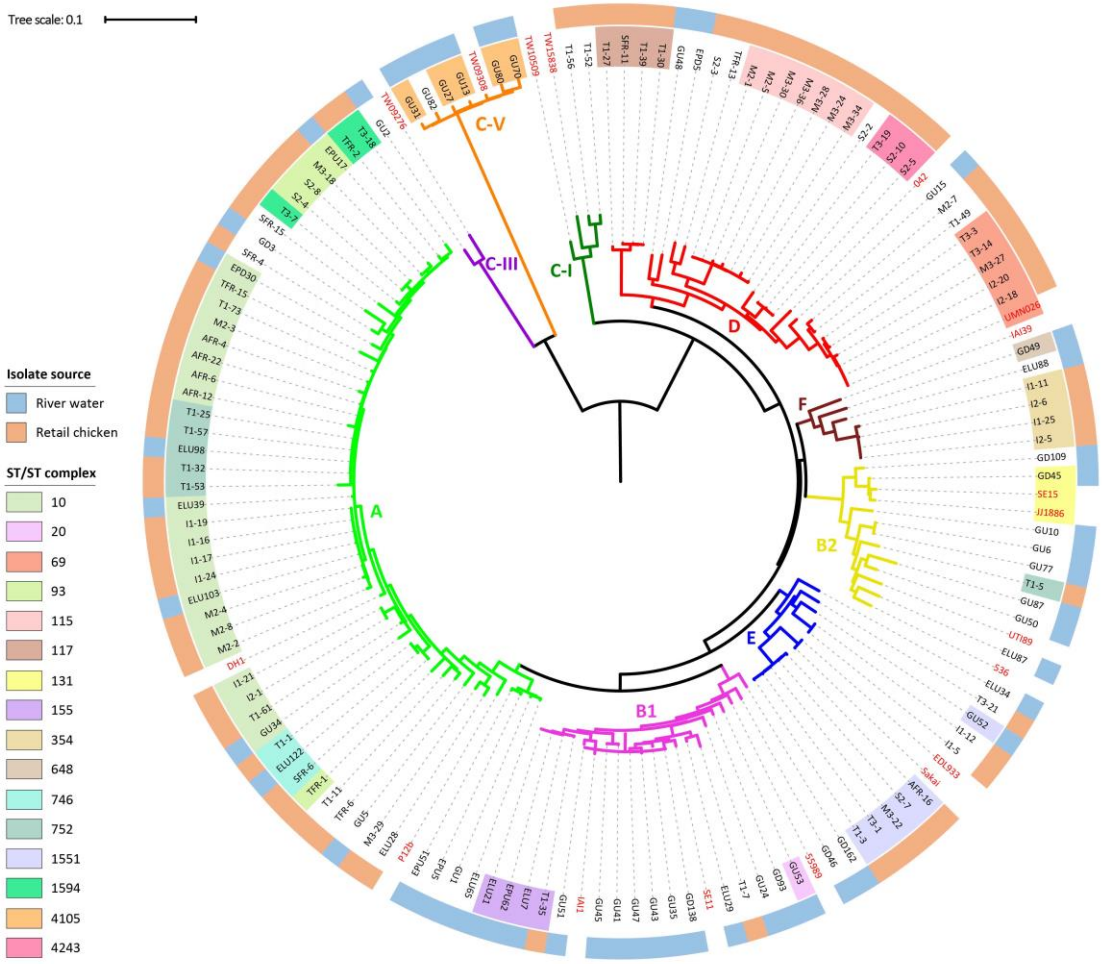


Figure 4.5. Maximum-likelihood phylogenetic tree of 128 *E. coli* strains isolated from river water and retail chicken samples in Nottingham and 18 reference strains.

The phylogeny was inferred from a core genome alignment of the population (87,981 bp, 59,317 SNPs, 146 genomes) constructed using Parsnp. Publicly available reference genomes (Table 4.1 and strain names marked in red on the tree), belonging to the *E. coli* phylogroups (A, B1, B2, D, E, and F) and cryptic *Escherichia* clades (C-I, C-III, and C-V), were included in the alignment to assign strains to a phylogenetic group based on its position on the tree. The phylogenetic tree was visualised and edited using iTOL. Source of isolation is annotated on the tree as coloured bars, as well as STs/ST complexes which define three or more strains. ST20, ST131, and ST648, each represented by a single strain, are also indicated on the tree as shading behind the isolate's name. The phylogenetic clades are defined by branch colouring according to each phylogroup. The tree reveals that six of the seven phylogroups are represented in the population, with the largest proportion of strains belonging to phylogroup A. Strains belonging to cryptic clades C-I, C-III, and C-V were identified in the population. No strains belonged to phylogroup C or clades C-II or C-IV, so these reference strains were removed from the alignment.

4.3.6. Distribution of antimicrobial resistance genes among non-human *E. coli*

All non-human *E. coli* genomes were screened for the presence of acquired antibiotic resistance genes, by running the bioinformatics pipeline ABRicate, which scans the ResFinder database to generate *in silico* antibiotic resistance gene profiles for each isolate (Fig. 4.6). A total of 46 different resistance determinants were identified, corresponding to 8 different antibiotic classes: aminoglycosides, β-lactams, macrolide-lincosamide-streptogramins (MLS), phenicols, rifampicin, sulphonamides, tetracyclines, and trimethoprim. The phylogenetic distribution of resistance genes appears to be relatively consistent across the population, with the exception of a large proportion of phylogroup B1 represented by river water isolates, which demonstrate a clear lack of antimicrobial resistance genes compared to the rest of the population. Among these strains, only *tet*(34) encoding tetracycline resistance was identified, whereas *aadA1* (aminoglycoside), $bla_{TEM\text{-}1B}$ (β-lactamase), and sulphonamide resistance genes (*sul*), though possessed by the majority of the population, could not be detected. This is supported by low level antimicrobial resistance among phylogroup B1 *E. coli* isolated from cattle, which has been observed previously (Bok *et al*., 2014). Noticeably, there is also a lack of resistance determinants present in cryptic clades C-III and C-V, which is consistent with a previous study which reported a low frequency of resistance for the cryptic clades (Ingle *et al*., 2011), but is in contrast with a recent study which has reported antimicrobial resistance among cryptic clade isolates (Blyton *et al*., 2015). Acquired antibiotic resistance genes were most prevalent among strains belonging to phylogroup A and phylogroup D, which is in agreement with previous studies of antimicrobial resistance among *E. coli* phylogroups (Pavlickova, Dolezalova and Holko, 2015; Mosquito *et al*., 2015).

Aminoglycoside resistance genes

Twelve distinct aminoglycoside resistance genes were detected, representing the largest group of antimicrobial resistance determinants in the population. The *aadA1* gene, which confers resistance to streptomycin and spectinomycin, was the most prevalent aminoglycoside resistance gene and was detected in approximately 45% of the population, representing each phylogroup except for cryptic clades C-III and C-V. Other streptomycin resistance genes included *strA* and *strB*, which were also detected at relatively high frequencies, and these two genes were always detected together. Other than streptomycin resistance, the *aac(3)-IId* gene, which confers resistance to gentamycin and tobramycin, was detected in two isolates. Aminoglycoside resistance genes that were unique to retail chicken isolates included *aadA2*, *aadA5*, *aadA12*, *aadA13*, *aac(3)-IId*, *aac(3)-Iva*, *aph(3')-Ic*, and *aph(4)-Ia*, whereas the only unique gene to river water isolates was *aph(3')-Ia*. The high prevalence of

streptomycin resistance genes in this population is consistent with a previous study, which found these genes to be widespread in environmental habitats and often occur on mobile genetic elements, which can easily be acquired by different strains and species of bacteria (van Overbeek *et al*., 2002). Notwithstanding this, resistance to streptomycin does not usually define multidrug resistance, so phenotypic testing for susceptibility to this agent is not essential (Magiorakos *et al*., 2012), and the prevalence of genes conferring resistance to this antibiotic cannot be used as a measure of multidrug resistance in the non-human population of *E. coli*.

β-lactam resistance genes

The *in silico* antibiotic resistance gene analysis of the 128 sequenced *E. coli* genomes of the study population indicated an even lower prevalence of ESBL genes in the population (Fig. 4.6) than what was suggested by the multiplex PCR assays, conducted in section 4.3.2. Among sequenced isolates, a total of 9 different genes encoding β-lactamases were identified, with the most common being $bla_{\text{TEM-1B}}$ (31.5%). The bla_{TEM} family of β-lactamase genes, conferring resistance to the penicillin-like antibiotics such as ampicillin, were prevalent in 35.2% of isolates, representing 5 different variants of the gene ($bla_{\text{TEM-1A}}$, $bla_{\text{TEM-1B}}$, $bla_{\text{TEM-1C}}$, $bla_{\text{TEM-1D}}$, and $bla_{\text{TEM-33}}$). This is consistent with several previous studies that have also reported bla_{TEM} as a commonly encountered class of antibiotic resistance genes (Bajpai *et al*., 2017; Jena *et al*., 2017). Regarding the β-lactamase gene bla_{SHV}, and the ESBL genes $bla_{\text{CTX-M}}$ and bla_{OXA}, a notable disparity was observed between the percentage prevalence data of the *in silico* antibiotic resistance gene analysis when compared to the multiplex PCRs. bla_{SHV} and $bla_{\text{CTX-M}}$ could not be detected in the sequenced isolates, whereas these genes were amplified by PCR in 0.4% and 5.2% of the population, respectively, representing 13 isolates. However, it must be noted that 8 of these 13 isolates were not included in the population of sequenced isolates, thus possibly contributing to the lack of bla_{SHV} and $bla_{\text{CTX-M}}$ genes observed in the non-human *E. coli* population subjected to WGS. Furthermore, bla_{OXA} genes, namely $bla_{\text{OXA-1}}$ and $bla_{\text{OXA-10}}$ encoding carbapenem-hydrolysing oxacillinases, were identified in 1.6% of sequenced isolates; however, no bla_{OXA} genes could be detected by PCR. A possible reason for this disparity could be attributed to the methods used to detect antimicrobial resistance genes. Bioinformatic detection of resistance genes from WGS data is considered to be more effective in determining the full spectrum of antibiotic resistance genes in each isolate, and can generally detect more resistance genes than PCR (Moran *et al*., 2017). In contrast to the low prevalence of ESBL-producing *E. coli* in the population, 5 instances of the AmpC-like β-lactamase gene bla_{CMY} were observed. This gene encodes a cephalomycinase, which confers extended resistance to many β-lactams, including first-, second-, and third-generation cephalosporins, as well as cephamycins such as cefoxitin and ceftriaxone (Zhao *et al*., 2001). The presence of variants of bla_{CMY} in the population was unique to *E. coli* isolated from

retail chicken. This is consistent with previous studies that have reported that bla_{CMY} genes are commonly present in *E. coli* and *Salmonella* isolated from food animals and retail chicken (Zhao *et al*., 2001; Winokur *et al*., 2001), which have also displayed decreased susceptibility to ceftiofur and ceftriaxone. It was noted that a relatively even distribution of β-lactamase genes was observed across river water and retail chicken isolates.

Macrolide-lincosamide-streptogramin B (MLS) resistance genes

Genes conferring resistance to macrolide, lincosamide, and streptogramin B (MLS) antibiotics are widespread in bacteria, including environmental isolates of *E. coli* (Gomi *et al*., 2017b). Several of these genes were identified in 34 isolates of the non-human population of *E. coli*, with the most prevalent being *Inu*(F) (11.7%), and others including *Inu*(B) (8.9%), *mph*(B) (4.7%), and *mef*(B) (3.1%). The erythromycin resistance gene *erm(B)* and *Isa(A)* gene were detected in one isolate, respectively, in addition to *mph*(A), which was identified in two isolates. *Inu*(B) and *Inu*(F) are recognised as members of the *Inu* gene family (Achard *et al*., 2005), which encode lincosamide nucleotidyltransferase enzymes responsible for the mediation of specific resistance to lincosamides, such as lincomycin and clindamycin. The *mph*(A) gene, which was also detected, encodes a macrolide phosphotransferase shown to confer azithromycin resistance in *E. coli* (Howie *et al*., 2010). Macrolide resistance genes were also detected in *E. coli* in a recent environmental study (Gomi *et al*., 2017b). The results of these studies suggest that *E. coli* may represent a major reservoir for macrolide resistance genes which could then be horizontally transferred to other bacteria.

Chloramphenicol resistance genes

Four different chloramphenicol resistance genes (*cml*A1, *cat*A1, *cat*B3, and *floR*) were detected in the population. These same four genes were also detected in chloramphenicol-resistant *E. coli* in two previous studies, which also reported *floR* as the most prevalent chloramphenicol resistance gene (Gomi *et al*., 2017b). This would suggest that chloramphenicol resistance in *E. coli* is typically encoded by these four main genes.

Sulphonamide resistance genes

Resistance to sulphonamides in *E. coli* results from the acquisition of an alternative dihydropteroate synthase gene (*sul*) (Perreten and Boerlin, 2003). There are three known types of *sul* genes (*sul1*, *sul2*, and *sul3*) described in the literature (Zankari *et al*., 2012), and all three were detected in the present study. These genes were identified in roughly 40% of the population of non-human *E. coli*, and thus a higher reported prevalence of sulphonamide resistance genes when compared to the β-lactamase genes. *sul1* and *sul2* were reported with

relatively similar prevalence (25.8% and 21.1%, respectively), whereas *sul3* was detected in only four isolates. While the detection frequency of these three *sul* genes is in agreement with some previous studies (Gomi *et al*., 2017b; Kaper, Nataro and Mobley, 2004), it would seem to differ among other studies (Su *et al*., 2012; Hu *et al*., 2008), demonstrating the diversity of *sul* gene distribution profiles among various regions.

Tetracycline resistance genes

Of the 46 distinct *tet* alleles described to date, three types of tetracycline resistance genes [*tet*(A), *tet*(B), and *tet*(34)] were detected in 113 isolates (88.3% of the population). The *tet*(34) gene was the most prevalent tetracycline resistance determinant in the population (78.9%), followed by *tet*(A) (28.1%) and *tet*(B) (11.7%). A similar observation of higher *tet*(A) prevalence compared to *tet*(B) prevalence in *E. coli* recovered from surface waters was reported in a previous study (Hu *et al*., 2008). In the present study, *tet*(A) was associated with retail chicken isolates more so than river water isolates of *E. coli*, and the *tet*(B) gene was not detected at all in *E. coli* isolated from river water. Although tetracycline is not used to treat *E. coli* infections in humans, resistance to tetracycline is still common among *E. coli* (Dominguez *et al*., 2002). This would also appear to be the case in non-human *E. coli*, as suggested by the high prevalence of tetracycline resistance genes reported in this study.

Trimethoprim resistance genes

Trimethoprim resistance is mainly attributed to the acquisition of a trimethoprim-insensitive dihydrofolate reductase, which is the target enzyme of this agent (Seputiene *et al*., 2010). More than 30 different dihydrofolate reductase (*dfr*) genes have been identified (Zankari *et al*., 2012). Seven of these *dfr* genes were detected in the non-human *E. coli* population of the present study, with the *dfrA1* resistance gene being the most prevalent variant (16.4%). The prevalence of other *dfr* alleles in the population is low, with the *dfrA5* allele detected in only 3 isolates and *dfrA7*, *dfrA16*, *dfrA17*, *dfrB1*, and *dfrB4* present in only single isolates, respectively. *dfrA17* has frequently been found in clinical *E. coli* isolates (Seputiene *et al*., 2010), which correlates well with the high level of resistance against trimethoprim observed for *E. coli* isolates of human-clinical origin in a previous Nottingham-based study (Croxall *et al*., 2011b). In contrast, the lower prevalence of *dfr* genes identified in non-human *E. coli* would suggest much lower levels of resistance to trimethoprim. This demonstrates the inefficacy of trimethoprim in clinical cases, which has traditionally been described for prophylaxis in UTIs.

Although a diverse range of resistance determinants were detected in the population, corresponding to 8 different antibiotic classes, it was not completely determined what proportion of the population explicitly expressed a multidrug-resistance phenotype (non-

susceptibility to antimicrobial drugs belonging to 3 or more classes). Carriage of certain antimicrobial resistance genes does not necessarily mean that isolates in possession of those genes would express the corresponding resistance phenotype. It is suggested that phenotypic antimicrobial susceptibility testing should be performed as part of future work on this study population, as such testing was not carried out in this study due to time constraints. The low prevalence of ESBL genes detected in the population would indicate that expression of multidrug resistance to the extended spectrum of β-lactam antibiotics likely occurs at a very low frequency, if at all. Producing an antibiogram for the entire population would provide further insight into whether a correlation exists between resistance genotypes and resistance phenotypes, and would further our understanding of the prevalence of multidrug resistance in the wider non-human population of *E. coli*. Comparisons between the antibiotic resistance gene profiles that characterise non-human and human-clinical *E. coli* will be made in chapter 5.

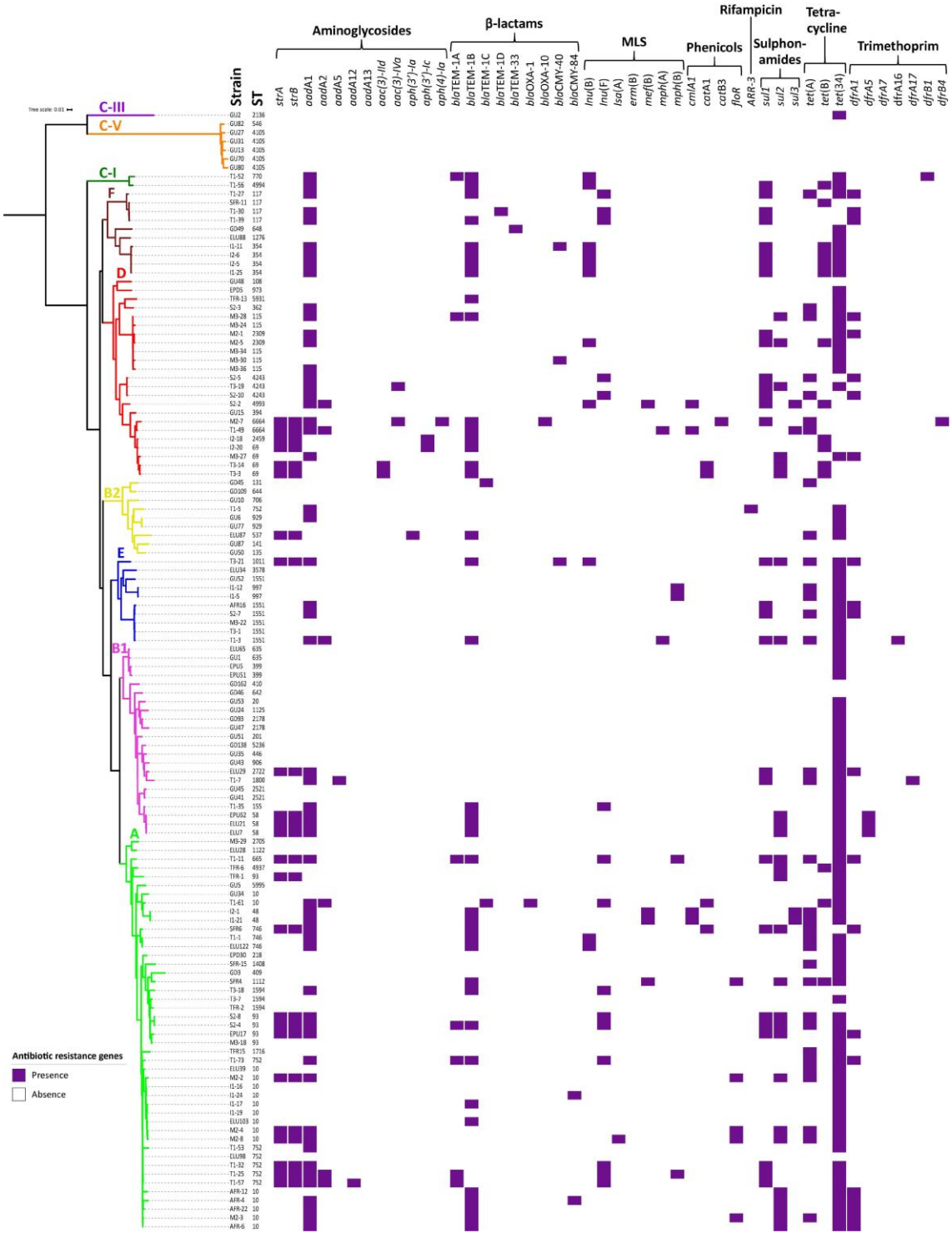


Figure 4.6. Distribution of antibiotic resistance gene profiles across the population of 128 *E. coli* strains isolated from river water and retail chicken samples in Nottingham.

The genomes of 128 *E. coli* strains were mass screened for antibiotic resistance gene carriage, by running the ABRicate bioinformatics tool against the ResFinder database. Presence of resistance determinants are shown on the phylogenetic tree as purple-coloured bars and are grouped by antibiotic class. ST designations, as determined by *in silico* MLST, are also indicated on the tree. The most prevalent antimicrobial resistance genes in the population were identified as *tet*(34), *aadA1*, $bla_{TEM\text{-}1B}$, and the *sul* genes. Isolates belonging to cryptic clades C-I and C-V, as well as a number of isolates of phylogroup B1, revealed much lower carriage of these genes in comparison to the rest of the population.

4.3.7. Determining the prevalence of ExPEC strains in the non-human population of *E. coli*

Multiple research groups around the world have reported a consistent observation of specific human ExPEC lineages in poultry or poultry products (Johnson *et al*., 2017; Jakobsen *et al*., 2010), as well as in river and surface waters (Gomi *et al*., 2017b; Muller, Stephan and Nuesch-Inderbinen, 2016). This would apparently provide evidence to support the hypothesis that there may be a poultry and environmental reservoir for human ExPEC. For this reason, an investigation was carried out to determine the prevalence of ExPEC strains among *E. coli* isolated from retail chicken and river water samples in this study. This involved screening all genomes of the study population for the presence of specific virulence-associated genes (VAGs), by running the bioinformatics pipeline ABRicate, which scans the Virulence Factors Database (VFDB) to generate *in silico* VAG profiles for each isolate (Fig. 4.7). The definition of ExPEC used in this study is based on the presence of five virulence markers, as implemented in previous studies (Gomi *et al*., 2017b; Johnson and Stell, 2000). The ExPEC pathotype was defined by the presence of two or more of *papA* and/or *papC*, *afa/dra*, *kpsMT II*, *iutA*, and *sfa/foc*, so therefore only these VAGs were considered in the analysis. The most frequently detected ExPEC virulence marker in the population (Fig. 4.7) was the iron acquisition gene *iutA* (61/128), which was more commonly detected in retail chicken isolates (90.2%) than in river water isolates of *E. coli* (9.8%). Gene clusters for the S fimbrial adhesin (*sfa*) and F1C fimbriae (*foc*) were less frequently detected in the population (9/128). Similarly, subunits of the *pap* operon *papA* and *papC*, which are associated with adhesion to the upper urinary tract, were detected together in 8 strains, with the exception of one strain (AFR-12) which possessed only *papC*. The *afa/dra* operons were present in only one isolate, whilst the type II capsule marker *kpsMT II* was not detected at all.

Contrary to previous studies, the ExPEC virulence gene profiles generated in this study indicate that the prevalence of ExPEC strains in the non-human population of *E. coli* is very low; only 11 of the 128 non-human *E. coli* isolates (8.6%) were classified as ExPEC, based on their VAG profiles. These ExPEC strains are phylogenetically distinct and are distributed across four different phylogroups, with 2 strains belonging to phylogroup D and 3 strains belonging to each of phylogroups A, B1, and B2, respectively. Additionally, ExPEC strains were identified in both river water (n = 7) and retail chicken (n = 4), indicating a presence of potentially pathogenic strains in the environment as well as the food chain. Consistent with a recent environmental study by Gomi *et al*. (2017b), non-human isolates of ExPEC exhibited clonal distribution with ST10, STC14, ST69, STC168, ST115, ST131, ST141, STC155 and ST4937 prevalent among these isolates. Clinically important clonal groups among these isolates would include ST10, STC14, ST69, and ST131. Although the sample size of ST131 isolates found and analysed in the present

study is very small (exactly one isolate), this strain did not carry any CTX-M-type ESBLs, which may be an early indication that environmental ST131 and human-clinical ST131 strains may be different in terms of their phylogenetic distribution, as suggested by a previous study using a similar study setting (Gomi *et al*., 2017b). Further genomic analyses between clinically important strains of non-human and human origin will be carried out in chapter 5. It was found that ST58 isolates, belonging to the ST155 complex of phylogroup B1, were the most commonly encountered of all ExPEC isolates detected (27.3%). These isolates were positive for the VAGs *pacA*, *papC*, *iutA*, and *sfa/foc*, exhibiting a similar virulence gene profile to the ST131 strain. STC-(sequence type complex)-155 was recently reported as a clonal group of animal origin that is spreading in humans and is highly drug-resistant (Skurnik *et al*., 2016). It is therefore interesting to note in the present study that STC155 isolates were prevalent in river water as well as retail chicken samples, with ExPEC strains of STC155 being exclusive to river water. This is supported by previous studies which have also reported STC155 strains in surface waters (Gomi *et al*., 2017b; Muller, Stephan and Nuesch-Inderbinen, 2016). Noticeably, however, genes encoding multidrug resistance were not identified in STC155 strains isolated in the current study, suggesting little clinical relevance. Contamination of surface waters with ExPEC strains belonging to clinically important STs would therefore be of little or no concern to human health due to the occurrence of these strains at such low frequencies in the environment.

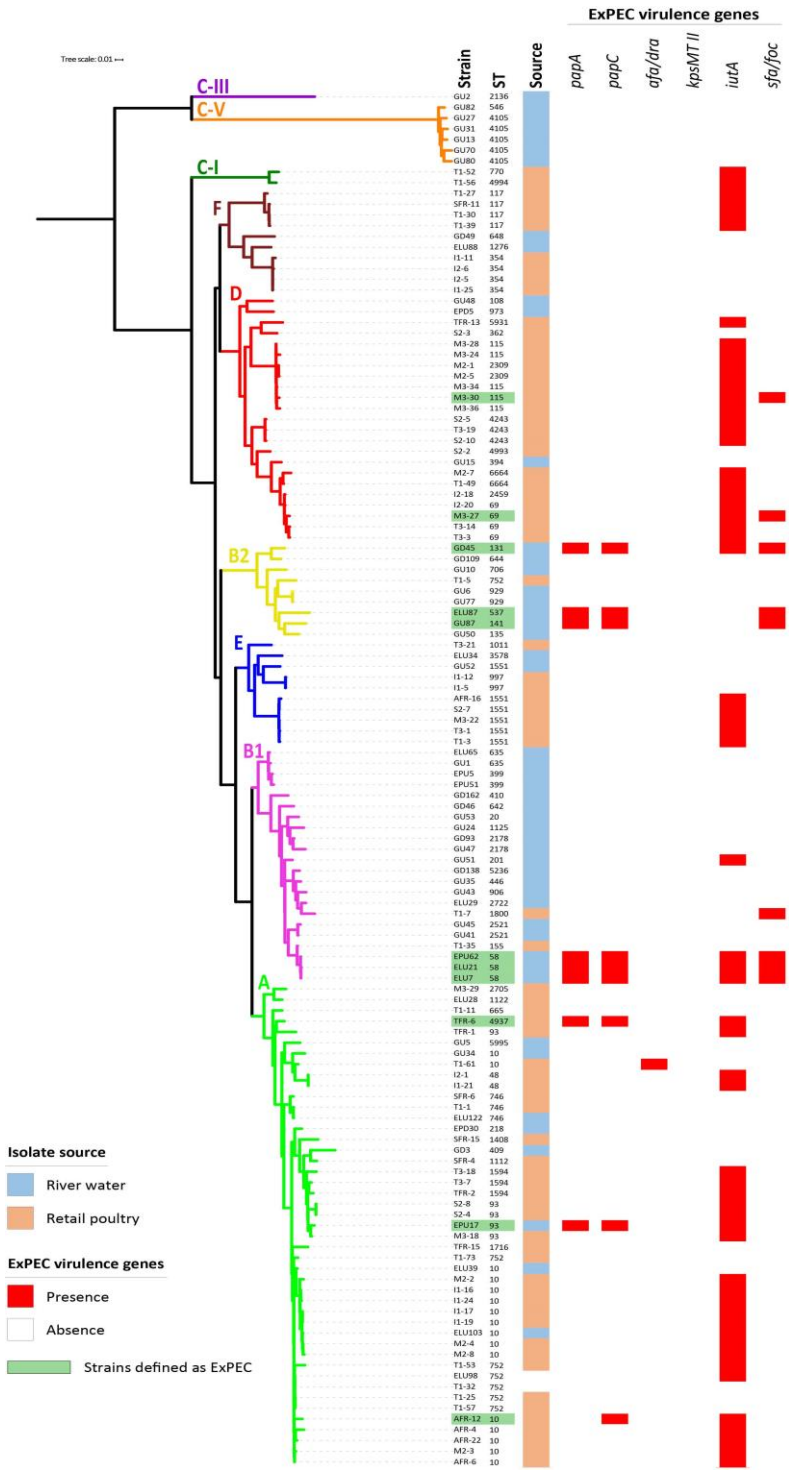


Figure 4.7. Distribution of ExPEC virulence-associated genes among the population of 128 *E. coli* strains isolated from river water and retail chicken samples in Nottingham. The genomes were mass screened for VAG carriage, by running the ABRicate bioinformatics tool to scan the Virulence Factors Database (VFDB). The presence of two or more of the following VAGs were used to define the ExPEC pathotype: *papA* and/or *papC*; *afa/dra*; *kpsMT II*; *iutA*; and *sfa/foc*. Presence of VAGs is annotated onto the maximum-likelihood phylogenetic tree as red-coloured bars and the strains identified as ExPEC are highlighted in green. ST designations, as determined by *in silico* MLST, and isolate sample source are also indicated on the tree. The prevalence of ExPEC strains among river water and retail chicken samples is low (8.6%) and these pathotypes are distributed among phylogroups B2 and D, as well as phylogroups B1 and A.

4.4. Conclusions

In the present study, *E. coli* obtained from river water and retail chicken samples in the Greater Nottingham area, were successfully characterised by whole-genome sequencing. From the initial sampling, *E. coli* were identified in 6 out of the total 9 river water samples collected, and 82 of all isolates selected from CLED plates (16%) were formally identified as *E. coli*, through a combination of microbial culture on chromogenic media, biochemical testing, and confirmation with the API 20E test system. *E. coli* were also isolated from 11 out of 20 retail chicken samples processed. A markedly higher prevalence of *E. coli* (88%) was observed from the 416 isolates selected from retail chicken sample plates. The presence of *E. coli* in freshwater and surface waters is a strong indication of recent human sewage or animal waste contamination. *E. coli* is a natural coloniser of the gastrointestinal tracts of a wide range of wild and domesticated animals, particularly those raised for human consumption, such as chickens. The retail chickens sampled in this study may have been contaminated at any of the multiple steps along the food chain, from production and processing at the abattoir to distribution and retail marketing in Nottingham. The presence of ExPEC lineages in retail chicken samples may indicate some early faecal contamination of the chicken meat occurred at the slaughterhouse. The difference in *E. coli* prevalence observed between retail chicken and river water samples in this study could be attributed to the wastewater treatment processes within the Trent River basin, which would appear to keep the release of effluents under relative control, and thus reducing the contamination of streams and rivers by faecal pathogens. Conversely, the numbers of *E. coli* isolated in this study would suggest that contamination of British retail chicken by faecal organisms is much harder to control, during the process of preparing chicken for human consumption.

Whole-genome sequences for the study population of 128 non-human *E. coli* strains were analysed by *in silico* multilocus sequence typing (MLST), and the population was found to be extremely clonally diverse, consisting of 64 different sequence types. It was evident from this analysis that the non-human population of *E. coli* is not dominated by the ST complexes that are commonly associated with urinary tract and bloodstream infections, such as the ST131, ST95, ST73, and ST69 complexes (Kallonen *et al*., 2017; Alhashash *et al*., 2013; Croxall *et al*., 2011b; Lau *et al*., 2008). The non-human population of *E. coli* analysed in this study comprised a wide variety of different STs, with the majority of strains not grouping into larger clonal complexes, demonstrating the genotypic diversity within the population. The ST10 clonal complex was the most frequently encountered clonal group, representing approximately 15% of the Nottingham non-human *E. coli* population analysed in this study, and was prevalent in both river water and retail chicken samples. This is an indication that the ST10 complex is widespread across the

environment and in the food chain, consistent with the findings of several previous studies (Gomi *et al*., 2017b; Chen *et al*., 2016; Cohen Stuart *et al*., 2012; Overdevest *et al*., 2011). Although current literature has reported the presence of well-known ExPEC sequence types in environmental waters and retail poultry, this study pinpoints a lack of such strains in the Nottingham population of non-human *E. coli*, suggesting the overall prevalence of human ExPEC strains in populations of non-human and commensal *E. coli* is very low.

Whole-genome analysis enabled the detection of 46 different resistance determinants among all isolates of the study population. The most frequently encountered resistance determinants in the non-human population of *E. coli* were *tet*(34), *aadA1*, *sul*, and $bla_{\text{TEM-1B}}$, which confer resistance to tetracycline, streptomycin/spectinomycin, the sulphonamides and penicillin-like antibiotics, respectively. The high prevalence of these resistance genes noted in the current study is in agreement with the current literature, which suggests that resistance conferred by these genes is widespread among *E. coli* isolated from environmental habitats (Gomi *et al*., 2017b; Bajpai *et al*., 2017; Kaper, Nataro and Mobley, 2004). Many previous studies have reported the presence of MDR *E. coli* strains in surface waters and water-related environments, as well as in retail poultry and food animals, on the basis of phenotypic resistance testing and detection of several resistance determinants (Gomi *et al*., 2017b; Johnson *et al*., 2017; Muller, Stephan and Nuesch-Inderbinen, 2016; Vincent *et al*., 2010; Jakobsen *et al*., 2010; Johnson *et al*., 2005a). Multidrug resistance in *E. coli* is mediated by extended-spectrum β-lactamases (ESBLs), mainly of the CTX-M family, particularly CTX-M-15 and 14, and less frequently of the SHV and OXA families (Nicolas-Chanoine *et al*., 2008; Lau *et al*., 2008). Considering that antimicrobial susceptibility testing was not performed in this study, it could not be determined what proportion of the population explicitly expressed a multidrug-resistance phenotype (non-susceptibility to antimicrobial drugs belonging to 3 or more classes). However, *in silico* antibiotic resistance gene analysis enabled a determination of the prevalence of the β-lactamase genes bla_{SHV}, bla_{TEM}, $bla_{\text{CTX-M}}$, and bla_{OXA}, more accurately than through multiplex PCR assays. Many environmental studies reporting isolates that carry ESBL genes often target resistant isolates in their sampling procedures, and therefore do not report an accurate representation of ESBL prevalence with regard to the wider non-human and environmental population of *E. coli*. In this study, however, an unbiased sampling strategy was employed, in combination with whole-genome sequencing of isolates, to reveal that a paucity of ESBL genes exists in the non-human population of *E. coli*, suggesting that multidrug resistance occurs at levels much lower than described in human-clinical populations of *E. coli*.

Reconstruction of the phylogenetic structure of the population revealed that six of the seven known *E. coli* phylogroups were represented in the population, as well the cryptic *Escherichia*

clades C-I, C-III, and C-V, demonstrating the full spectrum of genomic diversity of *E. coli* in the non-human *E. coli* study population. The majority of strains (approximately 52%) belonged to phylogroups A and B1, which are usually dominated by commensal and non-pathogenic strains (Duriez *et al*., 2001; Picard *et al*., 1999), whereas roughly 28% of the population were assigned to phylogroups B2 and D, which are typically associated with strains that cause extraintestinal infections (Johnson and Stell 2000; Picard *et al*., 1999). This suggests that the non-human population of *E. coli* constitutes a high proportion of commensal strains, with a much lower prevalence of ExPEC strains. This was confirmed by *in silico* VAG profiling of each strain for genes that are used to define ExPEC. This analysis revealed that only 8.6% of the non-human *E. coli* population could be classified as ExPEC. Moreover, these strains exhibited clonal diversity and only a small number of clinically important clonal groups were identified among these isolates, with very low prevalence. Although a small-scale representation of human ExPEC lineages, such as ST131, ST69, and ST648, has been determined through phylogenetic analysis in the present study, it cannot be overlooked that previous studies consistently reporting the presence of such clonal groups in the environment and the food chain would appear to support the hypothesis that these sources may serve as reservoirs for human ExPEC infection. However, further investigation using methods with higher resolving power would be required to determine the genomic relatedness between clinically important *E. coli* STs, isolated from non-human and human-clinical sources, as MLST and phylogenetic analyses only take into account variation within sections of the core genome. Any variations within the accessory genome that may contribute to virulence would not be detected. This will be addressed in chapter 5, through comprehensive comparative genomics, which may offer the discrimination necessary to determine if a non-human reservoir of human ExPEC exists and address whether it contributes to the burden of human extraintestinal infections.

CHAPTER 5

Comparative population genomics of *Escherichia coli* from human-clinical and non-human sources

5.1. Introduction

Multilocus sequence typing (MLST) analysis in chapter 4 demonstrated that the non-human population of *E. coli* in Nottingham is clonally diverse. A very low prevalence of ExPEC strains was identified in this population, and additionally, the occurrence of extended-spectrum β-lactamase (ESBL) genes (and thus potential for multidrug resistance) was negligible. These results contrasted with previous studies reporting the presence of specific human extraintestinal pathogenic *E. coli* (ExPEC), and multidrug-resistant (MDR) lineages, in poultry (Johnson *et al*., 2017; Jakobsen *et al*., 2010) and surface waters (Gomi *et al*., 2017b; Coleman *et al*., 2013). Several previous studies have attempted to attribute transmission of ExPEC in humans to poultry or other environmental sources, but the majority of these studies usually selectively culture for antimicrobial-resistant bacteria, rather than culturing all bacteria and then quantifying resistant strains within that population. These studies, therefore, do not present an accurate snapshot of the prevalence of multidrug-resistant *E. coli* in the environment and food-chain (Manges 2016; Lazarus *et al*., 2015). Many of these studies have used traditional, low-resolution typing methods to deduce that these *E. coli* strains, and genes such as those encoding antibiotic resistance and virulence factors, can spread from food-producing animals, via the food-chain, to humans (Platell *et al*., 2011b; Dolejska *et al*., 2011a), and additionally from environmental sources (Jang *et al*., 2013; Dolejska *et al*., 2011b). However, these methods may not have provided sufficient resolution to reliably assess the relatedness of these strains isolated from non-human and human-clinical sources. This highlights the importance of applying whole-genome comparative analysis, with the aim of distinguishing between seemingly related strains of bacteria.

E. coli isolated from different geographical regions and ecosystems have highly heterogeneous genomes and may vary in size by up to 1 Mbp (Bergthorsson and Ochman, 1998). The diversity between *E. coli* genomes can be attributed to the deletion or acquisition of mobile genetic elements by horizontal gene transfer. In a previous study by Lawrence and Ochman (1998), it was found that ∼18% of all open reading frames (ORFs) of the *E. coli* strain MG1655 are horizontally acquired, which have conferred properties permitting *E. coli* to colonise otherwise unreachable ecological niches. Horizontal gene transfer is largely responsible for the evolution of different *E. coli* pathotypes, as many virulence-associated genes (adhesins, toxins, invasins, and others) and antibiotic resistance genes may be encoded on mobile genetic elements, such as pathogenicity islands, plasmids, and transposons. Genes conferring other selective advantages, such as niche adaptation and fitness, also make up part of the dispensable (accessory) genome and can be readily transferred between strains via methods of horizontal gene transfer, such as conjugation and transduction. Whole-genome sequencing has provided a

method by which to comprehensively characterise the genetic diversity and evolution of large populations of related strains, which had proven very difficult prior to the advent of whole-genome analysis (Metzker, 2010). The *E. coli* species, which can range from harmless commensal to versatile pathogen, is a model organism for such whole-genome based studies (Tenaillon *et al*., 2010). An application of the comparative genomics approach would include the investigation of disease outbreaks and diagnosis of infectious disease agents. One example is the use of genome sequencing of environmental *E. coli* to expand the understanding of the ecology and speciation of this model organism, which was achieved in a study by Luo and co-authors (2011). Genomic comparisons were conducted between pathogenic/commensal *E. coli* and environmental strains that are phenotypically and taxonomically indistinguishable from typical *E. coli* (commensal or pathogenic). It was found that the commensal genomes, which encode for more functions that are important for fitness in the human gut, do not exchange genetic material with their environmental counterparts. Due to the high discriminatory ability of comparative genomics, it was revealed that genetic exchange between emergent ecologically distinct phylogenetic clades of *E. coli* may not be as pronounced or prolonged as would be expected (Luo *et al*., 2011).

Phylogenetic analysis of whole-genome sequence data has transformed our understanding of the evolution and expansion of many important bacterial lineages, due to the high-resolution view it provides. However, many of these analyses do not consider the potential role of the accessory genome when inferring evolutionary paths. To be able to accurately determine and compare the entire gene contents of multiple genomes, the pan-genome approach was developed (Tettelin *et al*., 2008). The pan-genome is the entire gene set of all strains of a species. It consists of a core genome, which represents the genes present in all strains of the species, and also a dispensable or variable (accessory) genome, which refers to genes that are not present in all of the strains; these include genes present in two or more strains, or even genes unique to only single strains (Tettelin *et al*., 2008). Therefore, on inclusion of every new genome in a pan-genome analysis, new strain-specific genes are added and thus, the size of the pan-genome increases. The core genome typically includes housekeeping genes for cell envelope or regulatory functions, while the accessory genome comprises genes which encode for the species diversity and provides selective advantages for strains, such as niche adaptation, antibiotic resistance, and virulence factors. Pan-genome analysis is therefore beneficial for comparing the population structures and mechanisms of adaptation and evolution for different bacterial populations, as well as providing targets for vaccines and antibiotic treatment (Tettelin *et al*., 2008). Pan-genome analysis can also be used to determine the gene pool of a given species (Tettelin *et al*., 2005), as well as from different species (Gordienko, Kazanov and Gelfand, 2013).

The pan-genome approach can also provide a method with which to contrast the gene content of strains from the same species, but of different pathotypes or clonal groups.

A recent study by de Been and colleagues (2014) implemented whole-genome sequencing (WGS) analyses to study the relatedness of cephalosporin-resistant *E. coli* from humans, retail chicken meat, poultry and pigs. This analysis demonstrated significant heterogeneity between human and poultry-associated isolates and the study failed to provide evidence for recent clonal transmission of cephalosporin-resistant *E. coli* strains from poultry to humans, as had been suggested previously based on traditional, low-resolution typing methods. Although this study suggested that there is little or no overlap between *E. coli* isolates of human and poultry origin, the study had focussed primarily on ESBL-producing *E. coli*. Similarly, several previous studies have been conducted attempting to address the question of potential for waterborne transmission of *E. coli* to humans from freshwater sources (Ojer-Usoz, Gonzãlez and Vitas, 2017; Zhang, Gao and Chang, 2016). However, these studies are usually biased towards ESBL-producing and MDR *E. coli* because of selectively culturing for resistant isolates in their sampling procedure, and thus the relative abundance of these isolates is largely unknown.

5.1.1. Aim and objectives

There is a need for whole-genome-based comparative analyses of *E. coli* populations from poultry and the environment and humans, in which ESBL-producing *E. coli* are not the sole focus, but instead investigated in proportion to their frequency. To address this, comparative genomic analysis of WGS data for non-human and human-clinical populations of *E. coli*, obtained from the same geographical region, is performed in this chapter, in order to provide sufficient resolution to determine the level of genetic heterogeneity between these populations. The two populations compared in this chapter include an unbiased sample of isolates, with regards to antimicrobial resistance, obtained from river water and retail chicken meat in the Nottingham area (i.e. the non-human population of *E. coli* defined in chapter 4). Representing the human-clinical population of Nottingham is a collection of blood and urine-derived isolates of *E. coli*, obtained from hospital- and community-acquired extraintestinal infections. Nottingham provides the ideal regional ecosystem for comparison of such populations of *E. coli*, because of the comprehensive phenotypic and genotypic data available for human-clinical ExPEC strains, collected by the NTU Pathogen Research Group over the past decade (Alhashash *et al*., 2016; Alhashash *et al*., 2013; Croxall *et al*., 2011b; Croxall *et al*., 2011a). Inclusion of human-clinical strains isolated from these Nottingham-based studies allows for a geographically constrained comparison between the two populations. By applying a pan-genome approach, in conjunction with core genome phylogenetic analyses, this chapter aims to provide high-resolution genomic

comparison to determine the extent of genetic overlap between the non-human and human-clinical populations of *E. coli* in Nottingham. These comprehensive comparative genomic analyses aim to provide sufficient discrimination, to address whether the non-human reservoir of *E. coli* contributes to the burden of hospital- and community-acquired human extraintestinal infections in this region.

Specific objectives of this chapter were:

- To compare the population structures of human-clinical and non-human *E. coli*, with regards to the prevalence of clinically important STs, as determined by *in silico* multilocus sequence typing.
- To determine the relatedness of strains from the human-clinical and non-human populations of *E. coli*, by constructing a SNP-based core genome phylogenetic tree.
- To compare the prevalence of antimicrobial resistance determinants and the prevalence of human ExPEC strains between the human-clinical and non-human populations of *E. coli*.
- To construct phylogenetic trees from core genome alignments, in order to determine the relatedness of two non-human representative strains of clinically important STs, ST131 (GD45) and ST648 (GD49), to the wider populations of the ST131 and ST648 lineages, obtained from multiple hosts.
- To perform comparative genomic analysis of all non-human and human-clinical strains of *E. coli*, using a pan-genome approach to identify the proportions of genomic loci that are unique to either population or present in both populations.
- To determine the extent of gene movement between closely related strains of the human-clinical and non-human populations of *E. coli*, by comparing the pan-genomes and detected core genome recombination events between strains of ST69 and ST10 from both populations.

5.2. Materials and Methods

The key bioinformatics tools, scripts, and methods used in this chapter were described previously in sections 2.6.2 – 2.6.7 of chapter 2. The *E. coli* strains listed in Table 5.1 represent the human-clinical population in Nottingham, which were used for comparative genomic analyses with the non-human population of *E. coli*. The *E. coli* strains used for comparative phylogenetic analyses of ST131 and ST648 populations are detailed in Table 5.2 and Table 5.3, respectively.

Table 5.1. One hundred and thirty-six sequenced human-clinical *E. coli* genomes isolated from Nottingham.

Strain name	ST	ST complex	Sample source	Disease type	Year of isolation
B3	131	131	Blood	Bacteraemia	2011
B5	131	131	Blood	Bacteraemia	2011
B9	10	10	Blood	Bacteraemia	2011
B10	73	73	Blood	Bacteraemia	2011
B14	73	73	Blood	Bacteraemia	2011
B16	131	131	Blood	Bacteraemia	2011
B18	73	73	Blood	Bacteraemia	2011
B20	10	10	Blood	Bacteraemia	2011
B22	131	131	Blood	Bacteraemia	2011
B26	131	131	Blood	Bacteraemia	2011
B29	73	73	Blood	Bacteraemia	2011
B31	69	69	Blood	Bacteraemia	2011
B33	69	69	Blood	Bacteraemia	2011
B34	95	95	Blood	Bacteraemia	2011
B36	73	73	Blood	Bacteraemia	2011
B37	131	131	Blood	Bacteraemia	2011
B38	95	95	Blood	Bacteraemia	2011
B40	73	73	Blood	Bacteraemia	2011
B44	131	131	Blood	Bacteraemia	2011
B46	131	131	Blood	Bacteraemia	2011
B47	131	131	Blood	Bacteraemia	2011
B48	131	131	Blood	Bacteraemia	2011
B51	131	131	Blood	Bacteraemia	2011
B54	131	131	Blood	Bacteraemia	2011
B58	131	131	Blood	Bacteraemia	2011
B65	131	131	Blood	Bacteraemia	2011
B71	131	131	Blood	Bacteraemia	2011
B72	73	73	Blood	Bacteraemia	2011
B73	73	73	Blood	Bacteraemia	2011
B75	131	131	Blood	Bacteraemia	2011
B77	131	131	Blood	Bacteraemia	2011
B83	196	None	Blood	Bacteraemia	2011

B84	73	73	Blood	Bacteraemia	2011
B87	58	155	Blood	Bacteraemia	2011
B89	131	131	Blood	Bacteraemia	2011
B91	73	73	Blood	Bacteraemia	2011
B94	131	131	Blood	Bacteraemia	2011
B95	131	131	Blood	Bacteraemia	2011
B102	73	73	Blood	Bacteraemia	2011
B104	131	131	Blood	Bacteraemia	2011
B107	38	38	Blood	Bacteraemia	2011
B116	133	None	Blood	Bacteraemia	2011
B125	131	131	Blood	Bacteraemia	2011
B132	131	131	Blood	Bacteraemia	2011
B133	131	131	Blood	Bacteraemia	2011
B134	73	73	Blood	Bacteraemia	2011
B150	131	131	Blood	Bacteraemia	2011
U1	73	73	Urine	UTI	2011
U2	131	131	Urine	UTI	2011
U5	131	131	Urine	UTI	2011
U7	73	73	Urine	UTI	2011
U12	131	131	Urine	UTI	2011
U18	3451	None	Urine	UTI	2011
U19	10	10	Urine	UTI	2011
U21	73	73	Urine	UTI	2011
U22	95	95	Urine	UTI	2011
U24	73	73	Urine	UTI	2011
U30	73	73	Urine	UTI	2011
U36	73	73	Urine	UTI	2011
U42	73	73	Urine	UTI	2011
U44	131	131	Urine	UTI	2011
U48	73	73	Urine	UTI	2011
U50	73	73	Urine	UTI	2011
U58	91	None	Urine	UTI	2011
U60	95	95	Urine	UTI	2011
U64	69	69	Urine	UTI	2011
U67	69	69	Urine	UTI	2011
U76	73	73	Urine	UTI	2011
U79	131	131	Urine	UTI	2011
U80	131	131	Urine	UTI	2011
U92	131	131	Urine	UTI	2011
U102	38	38	Urine	UTI	2011
U104	3452	None	Urine	UTI	2009
UTI18	131	131	Urine	UTI	2009
UTI24	131	131	Urine	UTI	2009
UTI32	131	131	Urine	UTI	2009
UTI62	131	131	Urine	UTI	2009
UTI188	131	131	Urine	UTI	2009
UTI226	131	131	Urine	UTI	2009
UTI306	131	131	Urine	UTI	2009
UTI423	131	131	Urine	UTI	2009

UTI587	131	131	Urine	UTI	2009
F14W091968	10	10	MSU	UTI	2014
F14W127020-13	10	10	MSU	UTI	2014
F14W127020-20	10	10	MSU	UTI	2014
F14W131166-20	10	10	MSU	UTI	2014
M14W080122	12	12	MSU	UTI	2014
M14W098595-18	12	12	MSU	UTI	2014
M14W098595-31	12	12	MSU	UTI	2014
M14W101150	12	12	MSU	UTI	2014
M14W102050	12	12	MSU	UTI	2014
M14W107589	12	12	MSU	UTI	2014
F14W098435	58	155	MSU	UTI	2014
F14W125408	69	69	MSU	UTI	2014
F14W138284	69	69	MSU	UTI	2014
M14W071194-25	69	69	MSU	UTI	2014
M14W071194-2	69	69	MSU	UTI	2014
F14W108540	73	73	MSU	UTI	2014
F14W113875-18	73	73	MSU	UTI	2014
M14W127066	73	73	MSU	UTI	2014
M14W118794	80	568	MSU	UTI	2014
F14W114148-3	88	23	MSU	UTI	2014
F14W114148-7	88	23	MSU	UTI	2014
F14W131166-2	95	95	MSU	UTI	2014
M14W138421	95	95	MSU	UTI	2014
F13W143423	131	131	MSU	UTI	2014
F14W104167-24	131	131	MSU	UTI	2014
F14W104167-30	131	131	MSU	UTI	2014
F14W104167-31	131	131	MSU	UTI	2014
F14W104462-19	131	131	MSU	UTI	2014
F14W104462-28	131	131	MSU	UTI	2014
F14W118623	131	131	MSU	UTI	2014
F14W141832	131	131	MSU	UTI	2014
M14W073874	131	131	MSU	UTI	2014
M14W108795	131	131	MSU	UTI	2014
M14W113876	131	131	MSU	UTI	2014
M14W125435	131	131	MSU	UTI	2014
M14W131103-35	131	131	MSU	UTI	2014
M14W131103-5	131	131	MSU	UTI	2014
F14W071693	355	None	MSU	UTI	2014
F14W080037	404	14	MSU	UTI	2014
F14W108313	404	14	MSU	UTI	2014
F14W113464-2	648	648	MSU	UTI	2014
F14W113464-40	648	648	MSU	UTI	2014
M14W114085	648	648	MSU	UTI	2014
M14W140076	681	None	MSU	UTI	2014
1980_EC	95	95	Blood	Neonatal sepsis	2015
1982_EC	2622	None	Blood	Neonatal sepsis	2015
1983_EC	538	538	Blood	Neonatal sepsis	2015
1984_EC	73	73	Blood	Neonatal sepsis	2015

1985_EC	73	73	Blood	Neonatal sepsis	2015
2113_EC	95	95	Blood	Neonatal sepsis	2015
2114_EC	95	95	Blood	Neonatal sepsis	2015
2286_EC	69	69	Blood	Neonatal sepsis	2015
2297_EC	120	None	Blood	Neonatal sepsis	2015
2300_EC	458	73	Blood	Neonatal sepsis	2015

Table 5.1. One hundred and thirty-six sequenced human-clinical *E. coli* genomes isolated from Nottingham.

One hundred and thirty-six *E. coli* genomes, from the Nottingham Trent University (NTU) Pathogen Research Group strain collection, were included in this chapter for comparative genomic analyses with the Nottingham non-human population of *E. coli*, isolated in chapter 4. These strains were previously isolated from human-clinical samples (blood and urine cultures), obtained from the QMC hospital in Nottingham. Genome sequences and annotations were provided as FASTA files and GFF files from separate PhD studies at NTU: Gemma Clark, 2009; Fahad Alhashash, 2011; Ruqaiah Bedawai, 2014; Mohamed Saad, 2015.

ST: sequence type as confirmed by *in silico* multilocus sequence typing analysis; MSU: midstream sample of urine; UTI: urinary tract infection.

Table 5.2. Two hundred and forty-two sequenced ST131 *E. coli* genomes used for comparative phylogenetic analysis in this chapter.

Strain name	Source	Sample	Year	Country	Reference/ Accession number
JIE186	Human	UTI	2005	Australia	ERR537636
B36EC_81	Human	UTI	2007	Australia	(Petty *et al.*, 2014)
MS2481_77	Human	Bacteraemia	2007	Australia	(Petty *et al.*, 2014)
MS2493_59	Human	Bacteraemia	2007	Australia	(Petty *et al.*, 2014)
S104EC_75	Human	UTI	2008	Australia	(Petty *et al.*, 2014)
S105EC_73	Human	UTI	2008	Australia	(Petty *et al.*, 2014)
S98EC_75	Human	Asymptomatic	2008	Australia	(Petty *et al.*, 2014)
S100EC_53	Human	Unknown	2009	Australia	(Petty *et al.*, 2014)
S101EC_81	Human	Unknown	2009	Australia	(Petty *et al.*, 2014)
S108EC_61	Human	Neutropenia	2009	Australia	(Petty *et al.*, 2014)
S109EC_69	Human	Asymptomatic	2009	Australia	(Petty *et al.*, 2014)
S110EC_77	Human	UTI	2009	Australia	(Petty *et al.*, 2014)
S111EC_69	Human	UTI	2009	Australia	(Petty *et al.*, 2014)
S112EC_73	Human	UTI	2009	Australia	(Petty *et al.*, 2014)
S113EC_75	Human	Unknown	2009	Australia	(Petty *et al.*, 2014)
S65EC_79	Human	Unknown	2009	Australia	(Petty *et al.*, 2014)
S79EC_75	Human	Unknown	2009	Australia	(Petty *et al.*, 2014)
S99EC_49	Human	Asymptomatic	2009	Australia	(Petty *et al.*, 2014)
S102EC_51	Human	Unknown	2010	Australia	(Petty *et al.*, 2014)
S103EC_73	Human	Pyuria	2010	Australia	(Petty *et al.*, 2014)
S107EC_75	Human	Bacteraemia	2010	Australia	(Petty *et al.*, 2014)
S77EC_77	Human	UTI	2010	Australia	(Petty *et al.*, 2014)
S114EC_77	Human	UTI	2011	Australia	(Petty *et al.*, 2014)
S115EC_75	Human	UTI	2011	Australia	(Petty *et al.*, 2014)
19770	DA	Cat	2009	Austria	ERR264251
S121EC_81	Human	UTI	2000	Canada	(Petty *et al.*, 2014)
S123EC_81	Human	UTI	2001	Canada	(Petty *et al.*, 2014)
S125EC_83	Human	SWI	2002	Canada	(Petty *et al.*, 2014)
S126EC_79	Human	UTI	2002	Canada	(Petty *et al.*, 2014)
S127EC_59	Human	UTI	2002	Canada	(Petty *et al.*, 2014)
S131EC_75	Human	UTI	2002	Canada	(Petty *et al.*, 2014)
S122EC_81	Human	UTI	2003	Canada	(Petty *et al.*, 2014)
S124EC_77	Human	UTI	2003	Canada	(Petty *et al.*, 2014)
S128EC_79	Human	Primary sepsis	2004	Canada	(Petty *et al.*, 2014)
S129EC_79	Human	UTI	2004	Canada	(Petty *et al.*, 2014)
S130EC_83	Human	UTI	2004	Canada	(Petty *et al.*, 2014)
S132EC_77	Human	UTI	2005	Canada	(Petty *et al.*, 2014)
S133EC_73	Human	UTI	2005	Canada	(Petty *et al.*, 2014)
S134EC_81	Human	UTI	2005	Canada	(Petty *et al.*, 2014)
S135EC_77	Human	BTI	2005	Canada	(Petty *et al.*, 2014)
S120EC_63	Human	BTI	2009	Canada	(Petty *et al.*, 2014)
WCE266	Human	Ascites fluid	2005	China	LSFO00000000
WCE296	Human	Pleural effusion	2005	China	LSEZ00000000
WCE307	Human	Bacteraemia	2005	China	LSGU00000000

WCE208	Human	UTI	2006	China	LSEX00000000
WCE233	Human	Bacteraemia	2006	China	LSEY00000000
E4	Human	Asymptomatic	2011	China	LRXB00000000
J09	Human	UTI	2011	China	LSEG00000000
J21	Human	Asymptomatic	2011	China	LSEH00000000
M8	Human	Asymptomatic	2011	China	LSEK00000000
K0178B	Avian	Cormorant	2007	CZR	LSEI00000000
HP47	Avian	Rook	2010	CZR	LRXE00000000
27678	Avian	Rook	2011	CZR	ERR264279
27679	Avian	Rook	2011	CZR	ERR264280
27683	Avian	Rook	2011	CZR	ERR264281
27684	Avian	Rook	2011	CZR	ERR264282
27685	Avian	Rook	2011	CZR	ERR264283
27686	Avian	Rook	2011	CZR	ERR264284
27690	Avian	Rook	2011	CZR	ERR264285
17530	DA	Dog	2008	Denmark	ERR264240
18582	DA	Dog	2009	Denmark	ERR264245
22233	DA	Dog	2010	Denmark	ERR264266
18572	DA	Dog	2009	France	ERR264244
23942	DA	Dog	2010	France	ERR264273
12520	Avian	Chicken	2007	Germany	ERR264278
12556	DA	Dog	2008	Germany	ERR264289
18342	DA	Dog	2009	Germany	ERR264242
18982	DA	Dog	2009	Germany	ERR264246
19001	DA	Dog	2009	Germany	ERR264247
19336	DA	Dog	2009	Germany	ERR264249
19529	DA	Dog	2009	Germany	ERR264250
21176	Human	Unknown	2010	Germany	ERR264257
21177	Human	Unknown	2010	Germany	ERR264258
21178	Human	Unknown	2010	Germany	ERR264259
21181	Human	Unknown	2010	Germany	ERR264260
21182	Human	Unknown	2010	Germany	ERR264261
21201	Human	Unknown	2010	Germany	ERR264262
21514	DA	Dog	2010	Germany	ERR264263
22022	Avian	Bird	2010	Germany	ERR264264
22186	DA	Cat	2010	Germany	ERR264265
22239	DA	Dog	2010	Germany	ERR264267
22372	DA	Cat	2010	Germany	ERR264268
23511	DA	Dog	2010	Germany	ERR264269
23778	DA	Dog	2010	Germany	ERR264271
23927	DA	Cat	2010	Germany	ERR264272
24089	DA	Dog	2010	Germany	ERR264274
24510	DA	Dog	2010	Germany	ERR264275
852	Human	UTI	2011	Germany	LNPW00000000
972	Human	UTI	2011	Germany	LNOM00000000
1011	Human	UTI	2011	Germany	LRHP00000000
1019	Human	UTI	2011	Germany	LRHQ00000000
1039	Human	UTI	2011	Germany	LRHR00000000
1087	Human	UTI	2011	Germany	LRHS00000000

1223	Human	UTI	2011	Germany	LRHT00000000
1351	Human	UTI	2011	Germany	LRHU00000000
1366	Human	UTI	2011	Germany	LRVN00000000
1369	Human	UTI	2011	Germany	LRVO00000000
1380	Human	UTI	2011	Germany	LRVP00000000
1389	Human	UTI	2011	Germany	LRVQ00000000
1402	Human	UTI	2011	Germany	LRVR00000000
2963	Human	Asymptomatic	2011	Germany	LRVS00000000
2999	Human	UTI	2011	Germany	LRVU00000000
3019	Human	UTI	2011	Germany	LRVT00000000
3020	Human	UTI	2011	Germany	LRVV00000000
3134	Human	UTI	2011	Germany	LRVW00000000
3140	Human	UTI	2011	Germany	LRVX00000000
24790	DA	Dog	2011	Germany	ERR264276
24839	DA	Dog	2011	Germany	ERR264277
26368	Livestock	Cattle	2011	Germany	ERR264288
IR18E_63	Human	UTI	2009	India	(Petty *et al*., 2014)
IR49_69	Human	UTI	2009	India	(Petty *et al*., 2014)
IR65_69	Human	UTI	2009	India	(Petty *et al*., 2014)
IR68_79	Human	UTI	2009	India	(Petty *et al*., 2014)
MB1074	Human	Asymptomatic	2012	Ireland	LSEL00000000
MB14972	Human	Asymptomatic	2012	Ireland	LSEM00000000
MB17684	Human	Asymptomatic	2012	Ireland	LSEN00000000
MB3298	Human	Asymptomatic	2012	Ireland	LSEO00000000
MB3323	Human	Asymptomatic	2012	Ireland	LSEP00000000
19017	DA	Dog	2009	Italy	ERR264248
19801	DA	Dog	2009	Italy	ERR264252
20130	DA	Dog	2009	Italy	ERR264253
20441	DA	Cat	2010	Italy	ERR264255
20936	DA	Dog	2010	Italy	ERR264256
18570	DA	Dog	2009	NDL	ERR264243
20402	DA	Cat	2010	NDL	ERR264254
23736	DA	Dog	2010	NDL	ERR264270
S92EC_51	Human	Bacteraemia	2009	NZ	(Petty *et al*., 2014)
S93EC_79	Human	Bacteraemia	2009	NZ	(Petty *et al*., 2014)
S94EC_63	Human	Bacteraemia	2009	NZ	(Petty *et al*., 2014)
S95EC_75	Human	Bacteraemia	2009	NZ	(Petty *et al*., 2014)
S96EC_53	Human	Bacteraemia	2010	NZ	(Petty *et al*., 2014)
S97EC_73	Human	Bacteraemia	2010	NZ	(Petty *et al*., 2014)
K0198B	Avian	Cormorant	2007	Serbia	LSEJ00000000
HS115	Avian	Rook	2010	Serbia	LRXF00000000
27702	Avian	Rook	2011	Serbia	ERR264286
27703	Avian	Rook	2011	Serbia	ERR264287
17898	DA	Dog	2008	Spain	ERR264241
HVM1147_73	Human	Abscess	2010	Spain	(Petty *et al*., 2014)
HVM1299_69	Human	Abscess	2010	Spain	(Petty *et al*., 2014)
HVM1619_79	Human	SWI	2010	Spain	(Petty *et al*., 2014)
HVM1997_61	Human	UTI	2010	Spain	(Petty *et al*., 2014)
HVM2044_75	Human	UTI	2010	Spain	(Petty *et al*., 2014)

HVM2289_73	Human	UTI	2010	Spain	(Petty *et al.*, 2014)
HVM277_81	Human	UTI	2010	Spain	(Petty *et al.*, 2014)
HVM3017_57	Human	UTI	2010	Spain	(Petty *et al.*, 2014)
HVM3189_79	Human	UTI	2010	Spain	(Petty *et al.*, 2014)
HVM5_47	Human	UTI	2010	Spain	(Petty *et al.*, 2014)
HVM52_83	Human	UTI	2010	Spain	(Petty *et al.*, 2014)
HVM826_63	Human	RTI	2010	Spain	(Petty *et al.*, 2014)
HVM834_75	Human	UTI	2010	Spain	(Petty *et al.*, 2014)
HVR2496_77	Human	UTI	2010	Spain	(Petty *et al.*, 2014)
HVR83_61	Human	UTI	2010	Spain	(Petty *et al.*, 2014)
P53EC_53	Human	Asymptomatic	2011	Spain	(Petty *et al.*, 2014)
P56EC_47	Human	Asymptomatic	2011	Spain	(Petty *et al.*, 2014)
P146EC_71	Human	Asymptomatic	2011	Spain	(Petty *et al.*, 2014)
P189EC_55	Human	Asymptomatic	2011	Spain	(Petty *et al.*, 2014)
P50EC_77	Human	Asymptomatic	2011	Spain	(Petty *et al.*, 2014)
EC958	Human	UTI	2001	UK	CAFL01000001
S39EC_79	Human	Unknown	2004	UK	(Petty *et al.*, 2014)
S43EC_81	Human	Unknown	2004	UK	(Petty *et al.*, 2014)
S47EC_79	Human	Unknown	2004	UK	(Petty *et al.*, 2014)
S53EC_79	Human	Unknown	2004	UK	(Petty *et al.*, 2014)
S1EC_81	Human	UTI	2007	UK	(Petty *et al.*, 2014)
S2EC_61	Human	UTI	2007	UK	(Petty *et al.*, 2014)
S30EC_81	Human	UTI	2007	UK	(Petty *et al.*, 2014)
S31EC_81	Human	UTI	2007	UK	(Petty *et al.*, 2014)
S32EC_79	Human	UTI	2007	UK	(Petty *et al.*, 2014)
S5EC_75	Human	UTI	2007	UK	(Petty *et al.*, 2014)
S6EC_73	Human	UTI	2007	UK	(Petty *et al.*, 2014)
UTI18	Human	UTI	2008	UK	ERR062284
UTI188	Human	UTI	2008	UK	ERR062289
UTI226	Human	UTI	2008	UK	ERR062293
UTI24	Human	UTI	2008	UK	ERR062292
UTI306	Human	UTI	2008	UK	ERR062295
UTI32	Human	UTI	2008	UK	ERR062294
UTI423	Human	UTI	2008	UK	ERR062296
UTI587	Human	UTI	2008	UK	ERR062291
UTI62	Human	UTI	2008	UK	ERR062297
S10EC_77	Human	UTI	2009	UK	(Petty *et al.*, 2014)
S11EC_83	Human	UTI	2009	UK	(Petty *et al.*, 2014)
S12EC_79	Human	UTI	2009	UK	(Petty *et al.*, 2014)
S15EC_83	Human	UTI	2009	UK	(Petty *et al.*, 2014)
S19EC_73	Human	UTI	2009	UK	(Petty *et al.*, 2014)
S21EC_59	Human	UTI	2009	UK	(Petty *et al.*, 2014)
S22EC_83	Human	UTI	2009	UK	(Petty *et al.*, 2014)
S24EC_75	Human	UTI	2009	UK	(Petty *et al.*, 2014)
S26EC_69	Human	UTI	2009	UK	(Petty *et al.*, 2014)
S34EC_75	Human	UTI	2009	UK	(Petty *et al.*, 2014)
S37EC_83	Human	UTI	2009	UK	(Petty *et al.*, 2014)
S116EC_77	Human	BTI	2011	UK	(Petty *et al.*, 2014)
S117EC_49	Human	BTI	2011	UK	(Petty *et al.*, 2014)

S118EC_69	Human	UTI	2011	UK	(Petty *et al*., 2014)
S119EC_71	Human	Bacteraemia	2011	UK	(Petty *et al*., 2014)
B104	Human	Bacteraemia	2012	UK	LRWU00000000
B125	Human	Bacteraemia	2012	UK	LRWX00000000
B132	Human	Bacteraemia	2012	UK	LRWY00000000
B133	Human	Bacteraemia	2012	UK	LRWZ00000000
B150	Human	Bacteraemia	2012	UK	LRXA00000000
B16	Human	Bacteraemia	2012	UK	LRWA00000000
B22	Human	Bacteraemia	2012	UK	LRWB00000000
B26	Human	Bacteraemia	2012	UK	LRWC00000000
B3	Human	Bacteraemia	2012	UK	LRVY00000000
B37	Human	Bacteraemia	2012	UK	LRWD00000000
B44	Human	Bacteraemia	2012	UK	LRWE00000000
B46	Human	Bacteraemia	2012	UK	LRWF00000000
B47	Human	Bacteraemia	2012	UK	LRWG00000000
B48	Human	Bacteraemia	2012	UK	LRWH00000000
B5	Human	Bacteraemia	2012	UK	LRVZ00000000
B51	Human	Bacteraemia	2012	UK	LRWI00000000
B54	Human	Bacteraemia	2012	UK	LRWJ00000000
B58	Human	Bacteraemia	2012	UK	LRWK00000000
B65	Human	Bacteraemia	2012	UK	LRWL00000000
B71	Human	Bacteraemia	2012	UK	LRWM00000000
B75	Human	Bacteraemia	2012	UK	LRWN00000000
B77	Human	Bacteraemia	2012	UK	LRWO00000000
B89	Human	Bacteraemia	2012	UK	LRWR00000000
B94	Human	Bacteraemia	2012	UK	LRWS00000000
B95	Human	Bacteraemia	2012	UK	LRWT00000000
U12	Human	UTI	2012	UK	LSEQ00000000
U2	Human	UTI	2012	UK	LSER00000000
U44	Human	UTI	2012	UK	LSES00000000
U5	Human	UTI	2012	UK	LSET00000000
U79	Human	UTI	2012	UK	LSGT00000000
U80	Human	UTI	2012	UK	LSEU00000000
U92	Human	UTI	2012	UK	LSEV00000000
F13W143423	Human	UTI	2014	UK	Unpublished
F14W118623	Human	UTI	2014	UK	Unpublished
F14W141832	Human	UTI	2014	UK	Unpublished
M14W073874	Human	UTI	2014	UK	Unpublished
M14W108795	Human	UTI	2014	UK	Unpublished
M14W113876	Human	UTI	2014	UK	Unpublished
M14W125435	Human	UTI	2014	UK	Unpublished
M14W131103-35	Human	UTI	2014	UK	Unpublished
F14W104167-31	Human	UTI	2014	UK	Unpublished
F14W104167-30	Human	UTI	2014	UK	Unpublished
F14W104462-28	Human	UTI	2014	UK	Unpublished
F14W104167-24	Human	UTI	2014	UK	Unpublished
F14W104462-19	Human	UTI	2014	UK	Unpublished
M14W131103-5	Human	UTI	2014	UK	Unpublished
GD45	Environment	River water	2015	UK	Unpublished

F283	Avian	Crow	2012	USA	LRXD00000000
JJ1886	Human	UTI	2012	USA	CP006784.1

Table 5.2. Two hundred and forty-two sequenced ST131 *E. coli* genomes used for comparative phylogenetic analysis in this chapter.

Two hundred and forty-one ST131 *E. coli* genomes were included in this chapter for comparative phylogenetic analysis with the single Nottingham *E. coli* ST131 strain (GD45), isolated from river water in chapter 4. These strains were isolated from various hosts across multiple geographical locations and time periods. The ST131 population includes 125 avian (wild birds), domesticated animal (cats and dogs), livestock (cattle), and human-clinical isolates, sequenced as part of a recent study by McNally *et al*. (2016a); 102 human isolates (clinical and asymptomatic) from a previous phylogenomic study (Petty *et al*., 2014); and 14 human-clinical isolates obtained from the NTU Pathogen Research Group strain collection (unpublished). The genomes were provided by Alan McNally as FASTA files and the accession numbers (or citation) for these genomes are shown in the table.

DA: domesticated animal; SWI: surgical wound infection; UTI: urinary tract infection; RTI: respiratory tract infection; BTI: biliary tract infection; NDL: The Netherlands; UK: United Kingdom; USA: United States of America; CZR: Czech Republic; NZ: New Zealand.

Table 5.3. Eighty-nine sequenced ST648 *E. coli* genomes used for comparative phylogenetic analysis in this chapter.

Strain name	CTX-M carriage
F_12_GNB_311	*bla*$_{CTX-M-15}$
F_30_1_R8	ND
F_GNB_2781	*bla*$_{CTX-M-14}$
F_GNB_2809	ND
F_GNB_2838	*bla*$_{CTX-M-15}$
F_GNB_3697	*bla*$_{CTX-M-15}$
F_N13_1_3	*bla*$_{CTX-M-3}$
F_QUC093	*bla*$_{CTX-M-15}$
GD49	ND
IHIT22921	*bla*$_{CTX-M-15}$
IHIT22927	*bla*$_{CTX-M-15}$
IHIT22988	*bla*$_{CTX-M-15}$
IHIT22990	*bla*$_{CTX-M-15}$
IHIT23010	*bla*$_{CTX-M-15}$
IHIT23044	*bla*$_{CTX-M-15}$
IHIT23167	*bla*$_{CTX-M-15}$
IHIT23176	*bla*$_{CTX-M-15}$
IHIT23177	*bla*$_{CTX-M-15}$
IHIT25637	*bla*$_{CTX-M-15}$
IHIT25686	*bla*$_{CTX-M-15}$
IHIT27893	ND
IMT12298	ND
IMT12560	*bla*$_{CTX-M-1}$
IMT13211	*bla*$_{CTX-M-3}$
IMT16316	*bla*$_{CTX-M-15}$
IMT16343	*bla*$_{CTX-M-15}$
IMT16352	*bla*$_{CTX-M-15}$
IMT17438	*bla*$_{CTX-M-15}$
IMT17486	*bla*$_{CTX-M-14b}$
IMT17507	*bla*$_{CTX-M-15}$
IMT17539	*bla*$_{CTX-M-15}$
IMT17576	*bla*$_{CTX-M-15}$
IMT17887	*bla*$_{CTX-M-15}$
IMT17908	*bla*$_{CTX-M-3}$
IMT18337	*bla*$_{CTX-M-15}$
IMT18340	*bla*$_{CTX-M-15}$
IMT18351	*bla*$_{CTX-M-14}$
IMT18984	*bla*$_{CTX-M-15}$
IMT19322	*bla*$_{CTX-M-15}$
IMT20000	*bla*$_{CTX-M-15}$
IMT20607	*bla*$_{CTX-M-14}$
IMT20610	*bla*$_{CTX-M-14}$
IMT21183	*bla*$_{CTX-M-15}$
IMT21409	*bla*$_{CTX-M-15}$
IMT21500	*bla*$_{CTX-M-15}$

IMT21502	*bla*$_{CTX-M-15}$
IMT21509	*bla*$_{CTX-M-15}$
IMT21529	*bla*$_{CTX-M-14}$
IMT21531	*bla*$_{CTX-M-15}$
IMT22074	*bla*$_{CTX-M-15}$
IMT23463	*bla*$_{CTX-M-14}$
IMT23464	*bla*$_{CTX-M-14}$
IMT23760	*bla*$_{CTX-M-15}$
IMT23775	*bla*$_{CTX-M-15}$
IMT24056	*bla*$_{CTX-M-15}$
IMT24058	*bla*$_{CTX-M-15}$
IMT24081	*bla*$_{CTX-M-15}$
IMT24488	*bla*$_{CTX-M-15}$
IMT24490	*bla*$_{CTX-M-15}$
IMT24495	ND
IMT24616	*bla*$_{CTX-M-15}$
IMT24817	*bla*$_{CTX-M-15}$
IMT24818	*bla*$_{CTX-M-15}$
IMT24834	*bla*$_{CTX-M-15}$
IMT24837	*bla*$_{CTX-M-15}$
IMT24849	*bla*$_{CTX-M-15}$
IMT26356	*bla*$_{CTX-M-15}$
IMT27014	*bla*$_{CTX-M-15}$
IMT33120	ND
IMT33123	*bla*$_{CTX-M-15}$
IMT33127	*bla*$_{CTX-M-15}$
IMT33136	*bla*$_{CTX-M-15}$
IMT33143	*bla*$_{CTX-M-15}$
IMT33148	*bla*$_{CTX-M-15}$
IMT33149	*bla*$_{CTX-M-14}$
IMT33150	*bla*$_{CTX-M-15}$
IMT33151	*bla*$_{CTX-M-15}$
IMT33152	*bla*$_{CTX-M-15}$
IMT33159	*bla*$_{CTX-M-15}$
IMT33167	*bla*$_{CTX-M-15}$
IMT33170	*bla*$_{CTX-M-15}$
IMT33171	*bla*$_{CTX-M-15}$
IMT33601	ND
IMT33602	ND
IMT33608	*bla*$_{CTX-M-32}$
IMT33613	*bla*$_{CTX-M-15}$
IMT34407	ND
IMT34408	ND
NA023	*bla*$_{CTX-M-15}$

Table 5.3. Eighty-nine sequenced ST648 *E. coli* genomes used for comparative phylogenetic analysis in this chapter.

Table 5.3. Eighty-nine sequenced ST648 *E. coli* genomes used for comparative phylogenetic analysis in this chapter.

The genomes of 88 ESBL-producing ST648 *E. coli* strains were included in this chapter, for comparative phylogenetic analysis with the single Nottingham *E. coli* ST648 strain (GD49), isolated from river water in chapter 4. These strains originated from different hosts including humans, companion animals (cats, dogs, and horses), and wild birds, across Western Europe. The genomes were previously sequenced by Sebastian Guenther and colleagues (University of Greifswald) and were provided as FASTA files for this study. The CTX-M types identified for each strain, as determined by *in silico* resistance gene screening using ABRicate, are provided in the table.

ND: not detected.

5.3. Results and Discussion

5.3.1. Comparing the prevalence of STs between the human-clinical and non-human populations of *E. coli*

In silico MLST analysis in chapter 4 revealed that the Nottingham population of *E. coli* isolated from non-human samples in this study is clonally diverse, with a total of 64 different sequence types (STs) identified among 128 strains (section 4.3.4, Table 4.4). The population structure of non-human *E. coli* was comprised largely of a wide variety of different STs, with the ST10 clonal complex representing the most predominant central genotype, along with closely related STs such as ST93, ST746, ST752, and ST1551, which were prevalent in both retail chicken and river water samples. To put the prevalence of sequence types in the non-human *E. coli* population into the context of the human-clinical *E. coli* population, MLST data for a large collection of *E. coli* strains isolated from human-clinical cases in Nottingham were included in the analysis for comparison. Sequence type designations were obtained for 399 human-clinical *E. coli* strains isolated from two previous Nottingham-based studies; 134 of these strains were isolated from cases of urinary tract infection and urosepsis in elderly patients (Croxall *et al*., 2011b); 140 and 125 strains were isolated from bacteraemia patients and urinary samples, respectively, as part of a separate Nottingham-based study (Alhashash *et al*., 2013). Sequence types were determined in these studies by PCR-based MLST and using the typing scheme developed and hosted by Mark Achtman and colleagues (Wirth *et al*., 2006).

Percentage prevalence data for STs with ≥ 3 representative isolates identified among the human-clinical population and ≥ 3 isolates among the non-human population of *E. coli*, are presented in Figure 5.1A and 5.1B, respectively. Considering the larger sample size of isolates analysed for the human-clinical population of *E. coli*, this may explain why a greater number of

unique STs (n = 138) were identified. The most prevalent STs identified in the human-clinical population were ST131 (17.0%), ST73 (14.3%), ST69 (7.0%), and ST95 (5.8%), which are all largely associated with ExPEC infections in humans (Fig. 5.1A). In the non-human population of *E. coli*, on the other hand, the most prevalent STs were ST10 (12.5%), and closely related STs of the ST10 clonal complex, ST752 (5.5%), ST1551 (4.7%), and ST93 (3.9%). The 'other STs' category comprised all STs represented by ≤ 2 isolates in the human-clinical population (Fig. 5.1A) and ≤ 2 isolates in the non-human population (Fig. 5.1B). This category constituted 30.8% of human-clinical isolates and 44.5% of non-human isolates. Although more unique STs were detected in the human-clinical population, the level of genotypic heterogeneity would appear to be similar to that of the non-human population, which revealed a high proportion of isolates belonging to the 'other STs' category within a comparatively small sample size. This was confirmed by calculating the Shannon diversity indices for the abundance of STs in both the human-clinical (3.81) and non-human (3.79) populations of *E. coli*. The Hutcheson t-test was then used to compare the diversity of the two populations. The calculated t-value of 0.11 does not exceed the critical value of 1.96, indicating that the difference between the calculated diversities of both populations is not statistically significant ($p < 0.05$; 95% confidence interval). In comparison, the two populations would appear to differ in terms of population structure, as illustrated by the prevalence of STs in Figure. 5.1. The human-clinical population of *E. coli* is dominated by the four main ExPEC-associated STs, ST69, ST73, ST95, and ST131, which together represent around 44% of this population. It is well documented in the literature that these four sequence types are collectively responsible for a large proportion of *E. coli* urinary tract and bloodstream infections (Kallonen *et al*., 2017; Doumith *et al*., 2015; Alhashash *et al*., 2013; Croxall *et al*., 2011b; Lau *et al*., 2008). The ST95 complex, which is a prominent ExPEC sequence type complex associated with extraintestinal infections in humans and poultry (Vincent *et al*., 2010), was underrepresented in this strain set (5.8% of isolates) in comparison to the other major ExPEC STs reported. The relatively low proportion of ST95 strains compared to the overrepresentation of ST131 in the Nottingham human-clinical population of *E. coli* highlights the recent emergence of a new dominant clone (Johnson *et al*., 2017; Kallonen *et al*., 2017; Clark *et al*., 2012). The non-human population of *E. coli* isolated from this study is not highly represented by the four major STs identified in the human-clinical population, and instead the analysis reveals that the ST10 clonal complex is the most predominant genotype.

Comparison of the prevalence of the most predominant STs of both populations (ST131, ST73, ST69, ST95, and ST10) further demonstrates the disparity in the population structures of human-clinical and non-human *E. coli* (Fig. 5.2). While the important multidrug-resistant (MDR) ExPEC lineage ST131 was the most commonly encountered ST among human-clinical isolates analysed

in this study, only one instance of the same genotype was observed in the non-human population of *E. coli*, indicating a significantly low prevalence in comparison ($p < 0.0001$, two-tailed Fisher's test). Additionally, the well-known human ExPEC sequence types ST73 and ST95 were prevalent in the human-clinical population, but could not be detected among non-human isolates of *E. coli*. ST69, which is a highly virulent strain associated with human extraintestinal infections (Kallonen *et al*., 2017; Alhashash *et al*., 2013; Croxall *et al*., 2011b; Lau *et al*., 2008), as well as some animal models (Cristovao *et al*., 2017; Tartof *et al*., 2005), was identified among human-clinical *E. coli* isolates in addition to non-human (retail chicken) isolates (Fig. 5.2). This ST represented a higher proportion of human-clinical isolates (7.0%) than non-human isolates (3.1%), however, this difference may not be considered to be significant ($p = 0.137$, two-tailed Fisher's test). ST10 is a common human ExPEC ST that is also associated with food animals and retail poultry meat (Aslam *et al*., 2014; Bergeron *et al*., 2012; Vincent *et al*., 2010). It was revealed in this study that ST10 dominates in the non-human population of *E. coli* (prevalent among 12.5% of isolates), but comprised a significantly lower proportion (1.5%) of the human-clinical population of *E. coli* ($p < 0.0001$, two-tailed Fisher's test). This suggests a more widespread prevalence of this ST in the environment and food chain than in human-clinical cases of extraintestinal infection. The paucity of human ExPEC STs 131, 73, 69, and 95 in the non-human population of *E. coli* would suggest that the non-human reservoir of human ExPEC is negligible, at least in so far as it has been sampled in this study, and it is unlikely to be responsible for the majority of human ExPEC infections in Nottingham.

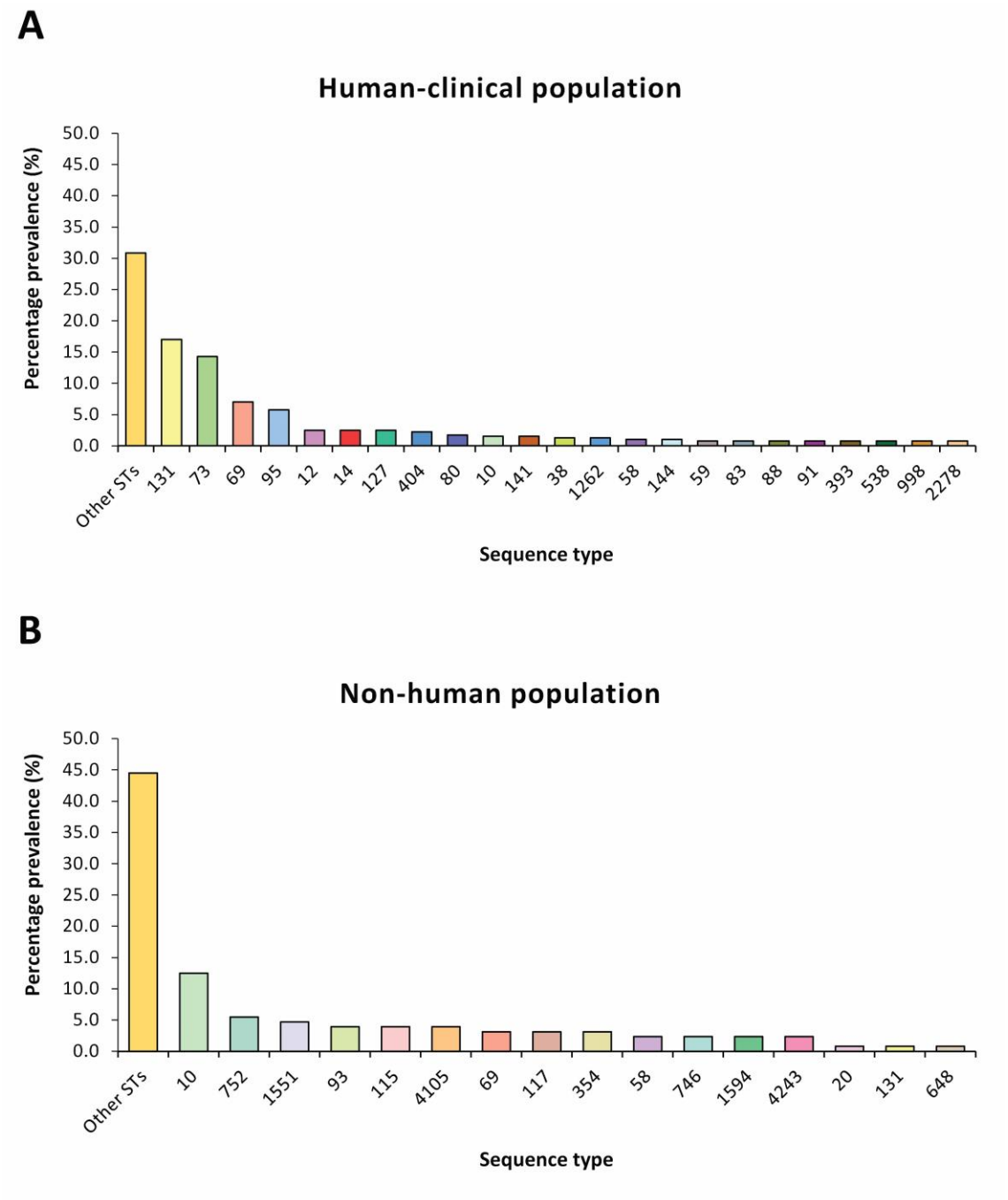


Figure 5.1. Percentage prevalence of *E. coli* sequence types (STs) among *E. coli* strains isolated from (A) human-clinical and (B) non-human samples collected in Nottingham.

The human-clinical population of *E. coli* (n = 399) comprise 259 isolates from cases of UTI and urosepsis in elderly patients, and 140 from bacteraemia patients. STs were determined by PCR-based MLST as part of previous studies (Alhashash *et al.*, 2013; Croxall *et al.*, 2011b). The non-human population comprises 128 sequenced strains from retail chicken and river water samples. ST designations were made by *in silico* MLST analysis of WGS data. Percentage prevalence is presented for all STs with ≥ 3 representative isolates, whilst the 'other STs' category constitutes all STs represented by ≤ 2 isolates, except for STs 20, 131, and 648, which are presented independently in chart B.

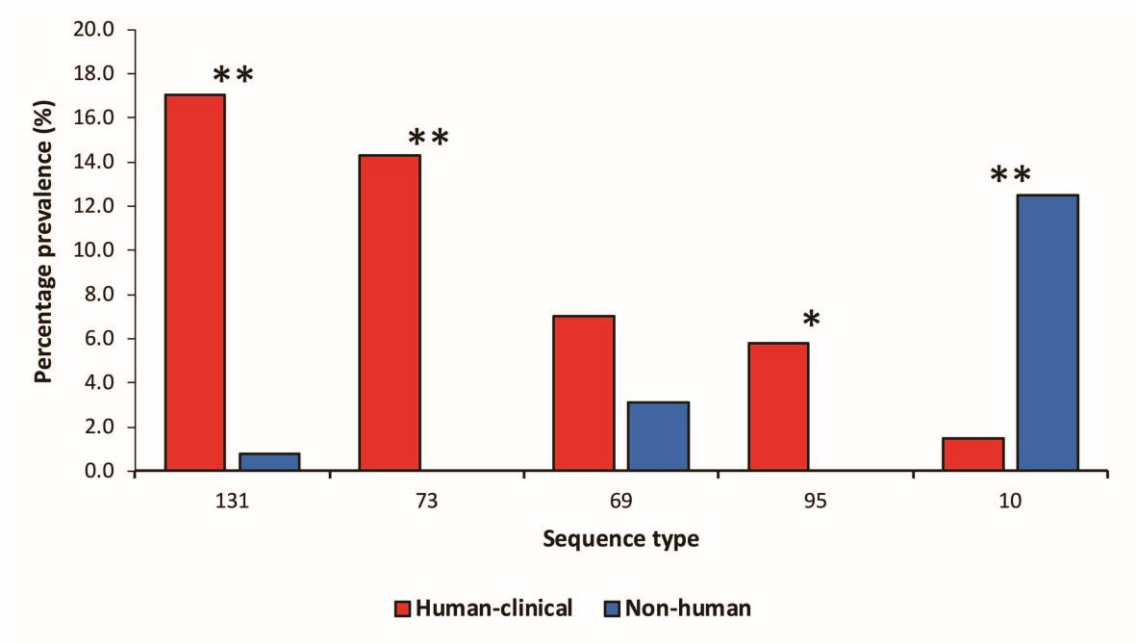


Figure 5.2. Percentage prevalence of the most predominant *E. coli* sequence types (STs) associated with human ExPEC infection in the human-clinical and non-human populations of *E. coli* in Nottingham.

Percentage prevalence comparison is presented for STs 131, 73, 69, 95, and 10, detected in the human-clinical population (n = 399) and non-human population (n = 128) of *E. coli*. ST131 was the dominant clone of the human-clinical population (17.0%), but only one ST131 strain (0.78%) was detected in the non-human population of *E. coli*, indicating a statistically significant difference between the two populations (**human-clinical ST131 vs non-human ST131; $p < 0.0001$, two-tailed Fisher's test). The well-known human ExPEC sequence types ST73 and ST95, which were also prevalent in the human-clinical population, were not identified among non-human isolates of *E. coli* (**human-clinical ST73 vs non-human ST73, $p < 0.0001$, two-tailed Fisher's test; *human-clinical ST95 vs non-human ST95, $p < 0.001$, two-tailed Fisher's test). ST69, on the other hand, was prevalent in both populations and whilst this genotype represented a higher proportion of human-clinical isolates (7.0%) than non-human isolates (3.1%), this may not be considered a statistically significant difference (human-clinical ST69 vs non-human ST69; $p = 0.137$, two-tailed Fisher's test). The prevalence of ST10 was significantly higher in the non-human population (12.5%) than that of the human-clinical (1.5%) population (**human-clinical ST10 vs non-human ST10; $p < 0.0001$, two-tailed Fisher's test). The asterisks represent comparisons made between human-clinical and non-human isolates.

The MLST data for both populations were combined and divided into groups of closely related isolates and clonal complexes, using the PHYLOViZ platform to construct a complete minimum spanning tree (MST) of all the STs in the analysed *E. coli* population (Fig. 5.3). When comparing *E. coli* isolates from human-clinical and non-human samples, the analysis revealed that the majority of these isolates clustered into different sequence types or clonal groups, based on their source of isolation. The MST illustrates a noticeable separation in the population structures of human-clinical and non-human *E. coli*. This is demonstrated by a large proportion of roughly 90% (124/138) of the STs identified among human-clinical *E. coli* isolates being unique to this population (i.e. not found in the non-human population). Similarly, a relatively large proportion of around 78% (50/64) of the STs identified among non-human *E. coli* isolates were unique to this population. This meant that out of the total 186 *E. coli* STs identified among all isolates, only 14 STs (7.5%) were prevalent in both the human-clinical and non-human populations. This would indicate that considerable genotypic differences exist between *E. coli* strains isolated from the two populations, and it suggests that the population of *E. coli* that exists in the environment and food chain is largely distinct from the human-clinical population of *E. coli* associated with human ExPEC infections. Of the 14 STs prevalent in both the human-clinical and non-human populations, clinically important ExPEC strains were identified among ST69, ST131, ST141, ST58, ST93, and ST10 in the non-human population of *E. coli* analysed. ST648 is a highly multidrug-resistant clone observed in human patients globally, and more incidentally in companion animals and wild birds (Guenther, Ewers and Wieler, 2011). This ST was also prevalent in the human-clinical and non-human populations presented in this study. Other STs present in both populations include ST48, ST394, ST409, ST644, ST929, ST1011, and ST2459. The presence of sequence types such as ST131, ST141, ST58, ST93, and ST648 in river water samples would indicate a relatively small level of contamination of surface waters by *E. coli* strains belonging to clinically important clonal groups, due to direct discharge from wastewater treatment plants and agricultural runoff. Human ExPEC strains belonging to ST69 and ST10 were isolated from retail chicken meat in this study, which may suggest a link between contaminated food products and *E. coli* strains that cause extraintestinal infection. Overall, the MST (Fig. 5.3) illustrates two populations of *E. coli* in Nottingham which appear to be distinct in terms of population structure and do not frequently come into contact with each other. Further genomic analyses would be able to determine whether strains of clinically important clonal groups isolated from the non-human population are the same strains implicated in community-acquired and hospital-acquired extraintestinal infections.

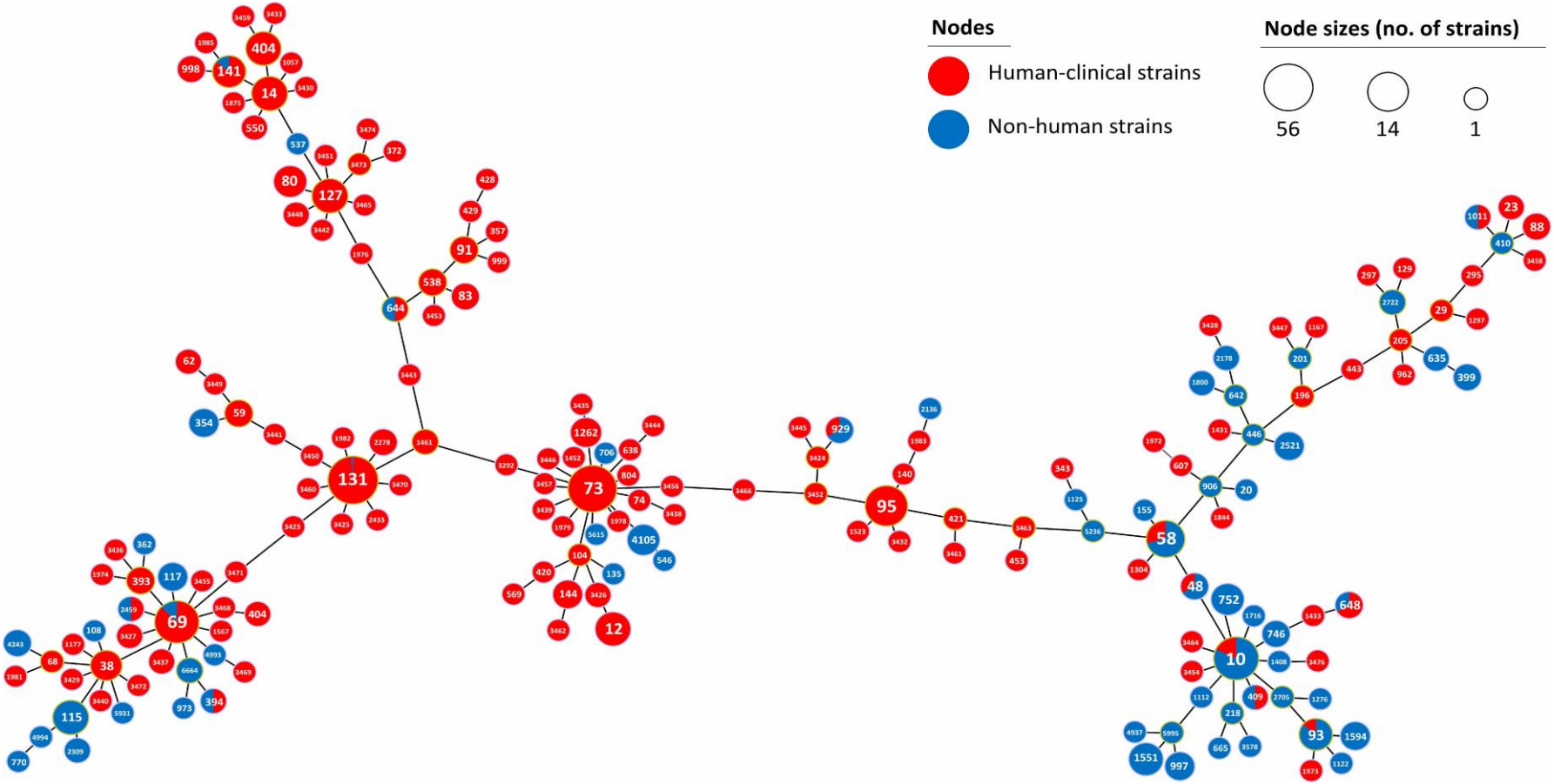


Figure 5.3. Minimum spanning tree (MST) illustrating the *E. coli* STs isolated from human-clinical and non-human samples.

The MST was produced using Phyloviz v3. The size of the nodes reflects the number of strains belonging to each ST. Nodes outlined by a yellow-green ring represent ST complexes. The sample types (human-clinical, red; non-human, blue) are overlaid onto the diagram, which confirms the predominance of ST131, ST73, ST65, and ST95 among human-clinical isolates and the ST10 clonal group among non-human isolates, with a limited number of STs shared between the two populations.

5.3.2. Comparative phylogenomic analysis of the human-clinical and non-human populations of *E. coli*

The collection of sequenced isolates from Nottingham-based studies includes the 128 *E. coli* genomes isolated from river water and retail chicken samples in chapter 4, representing the Nottingham non-human population of *E. coli* in this study. Also included are 136 *E. coli* genomes previously isolated from human-clinical samples, collected from the Queen's Medical Centre (QMC) hospital in Nottingham. The human-clinical population encompasses 47 isolates from cases of bacteraemia, 79 isolates from cases of UTIs, and 10 isolates from cases of neonatal sepsis, obtained from the NTU Pathogen Research Group strain collection (metadata for strains are provided in Table 5.1). It should be noted that these sequenced strains do not represent the true population structure of human-clinical *E. coli* in Nottingham, due to a large proportion of these genomes being selectively sequenced based on their ST designation and association with extraintestinal infections. This would explain the overrepresentation of ST131 and ST73 strains in the human-clinical population of *E. coli*.

A maximum-likelihood SNP-based phylogenetic tree was constructed using Parsnp (Fig. 5.4), which was derived from a core genome alignment, obtained from localised co-linear blocks (length = 62,825 bp, total SNPs = 55,677), of all non-human and human-clinical *E. coli* genomes (n = 264). The population included cryptic clade *E. coli* isolates, which served as the outgroup for the tree. With source of isolation annotated on the tree as coloured bars and phylogenetic clades defined by coloured branches according to phylogroup, the analysis revealed an observable phylogenetic divide between the human-clinical and non-human populations of *E. coli*. There is a clear clustering of isolates according to ST, and due to the overrepresentation of ST131, ST73, ST95, and ST12 strains, the human-clinical population is dominated by phylogroup B2, with 80.1% (109/136) of the population belonging to this phylogenetic group. In contrast, only 7% of the non-human population (9/128) were classified as phylogroup B2, demonstrating the lack of classically pathogenic strains in this population. The difference in population structures between the two populations is therefore quite evident. Other phylogroups present in the human-clinical population of *E. coli* would include phylogroups F (n = 3), D (n = 10), E (n = 2), B2 (n = 4), A (n = 7), and additionally, one strain belonging to cryptic clade C-V was also identified. Non-human *E. coli* isolates outnumber the human-clinical isolates present in these phylogroups, further illustrating the difference in the two population structures.

Although there is a bias towards ST131 and ST73 among the sequenced human-clinical *E. coli* genomes included in this analysis, the distribution of STs and clonal complexes demonstrated in Figure 5.3 indicates a prevalence of STs which are closely related to the larger clonal groups of

ST131, ST73, ST95, and ST69. When this is taken into consideration with the clustering of these ST complexes on the phylogenetic tree (Fig. 5.4), it strongly suggests that the wider human-clinical population of *E. coli* is largely represented by phylogroups B2 and D. This is consistent with previous studies which have indicated that Phylogroup B2 strains are generally more virulent than strains belonging to the other groups (Picard *et al*., 1999; Boyd and Hartl, 1998). Additionally, phylogroup B2 and, to a lesser extent, phylogroup D are predominantly associated with strains that cause extraintestinal infections (Johnson and Stell, 2000; Picard *et al*., 1999). Conversely, most of the non-human population belong to phylogroups A and B1, suggesting a high proportion of largely commensal strains of *E. coli*, as described by previous studies (Duriez *et al*., 2001; Picard *et al*., 1999).

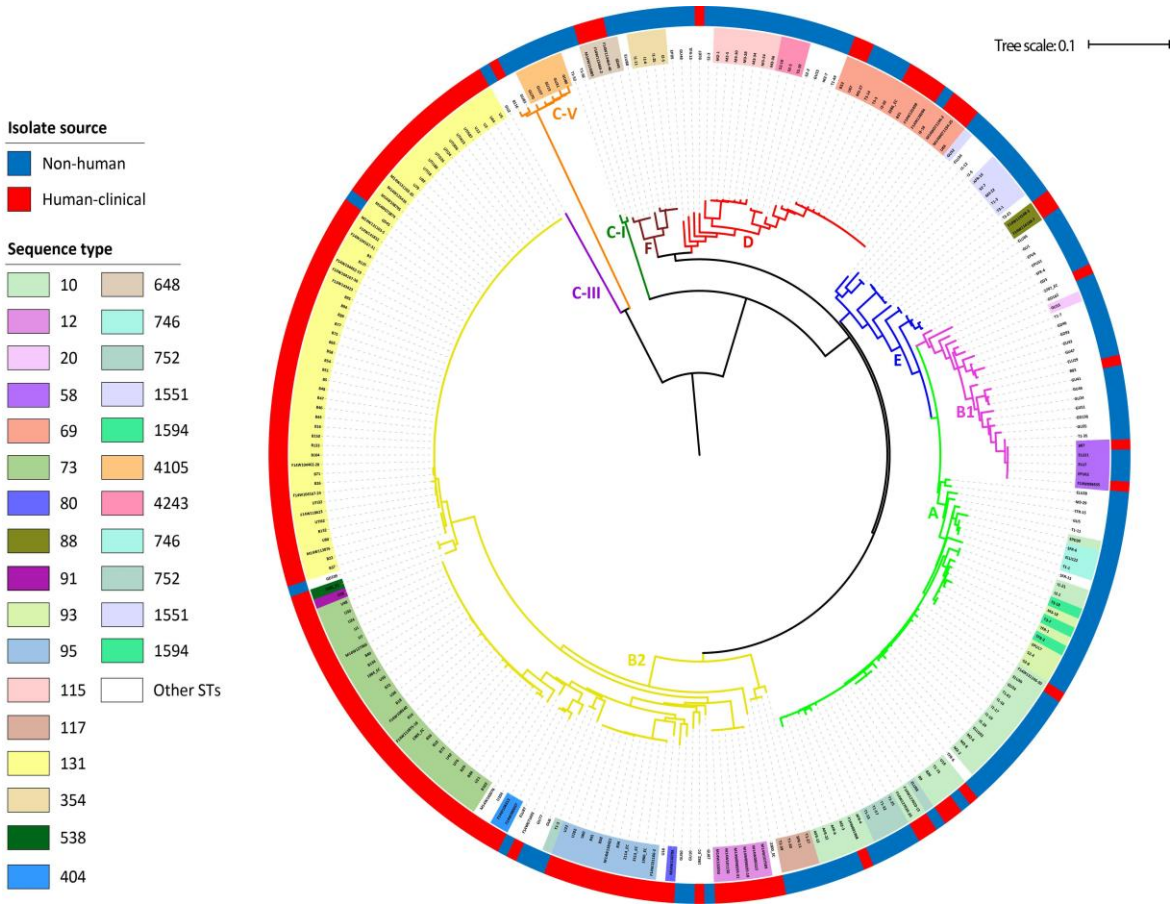


Figure 5.4. Maximum-likelihood SNP-based phylogenetic tree of 264 *E. coli* strains isolated from human-clinical and non-human samples in Nottingham.

The 128 *E. coli* genomes isolated from river water and retail chicken samples in chapter 4, representing the Nottingham non-human population of *E. coli*, are presented on the tree in addition to 136 *E. coli* genomes previously isolated from human-clinical samples. The human-clinical population encompasses 47 isolates from cases of bacteraemia, 79 isolates from cases of UTIs, and 10 isolates from cases of neonatal sepsis (details of strains are provided in section 5.2, Table 5.1). The phylogeny was inferred from a core genome alignment of the combined populations (62,825 bp, 55,677 SNPs, 264 genomes) constructed using Parsnp. The phylogenetic tree was visualised and edited using iTOL (Letunic and Bork, 2016). Source of isolation is annotated on the tree as coloured bars and STs/ST complexes are indicated on the tree as coloured segments behind the strain names. The phylogenetic clades are defined by branch colouring according to each phylogroup/cryptic clade.

5.3.3. Determining the prevalence of ExPEC strains in the human-clinical population of *E. coli*

The high prevalence of phylogroup B2 strains in the human-clinical population of *E. coli* would be suggestive of a high prevalence of virulence-associated genes (VAGs) (Picard *et al*., 1999; Boyd and Hartl 1998). The population is therefore likely to consist of a high proportion of ExPEC strains, due to this phylogroup being predominantly associated with strains that cause extraintestinal infections (Johnson and Stell 2000; Picard *et al*., 1999). In chapter 4, analysis of ExPEC VAG profiles indicated that the prevalence of ExPEC strains in the non-human population of *E. coli* is very low, which is in contrast to previous studies that have reported a consistent observation of specific human ExPEC lineages in poultry (Johnson *et al*., 2017; Jakobsen *et al*., 2010) and surface waters (Gomi *et al*., 2017b; Coleman *et al*., 2013). Only 11 of the 128 non-human *E. coli* isolates (8.6%) were classified as ExPEC, based on the possession of two or more of the following VAGs: *papA* and/or *papC*; *afa/dra*; *kpsMT II*; *iutA*; and *sfa/foc*.

In silico ExPEC VAG profiling was also carried out for the 136 sequenced *E. coli* isolates of the human-clinical population (Fig. 5.5). The genomes were screened for the presence of the above-mentioned VAGs by running the bioinformatics pipeline ABRicate, which scans the Virulence Factors Database (VFDB) to generate VAG profiles for each isolate. This allowed the prevalence of ExPEC strains in the human-clinical population to be determined and compared with that of the non-human population of *E. coli*. It was found that the *pap* operons *papA* and *papC*, as well as the iron acquisition gene *iutA*, are widespread throughout the human-clinical population, whilst the *afa/dra* operons and S fimbrial adhesin and F1C fimbriae cluster (*sfa/foc*) were detected in a number of phylogroup B2, F, and D strains (Fig. 5.5). Similar to the non-human population, the type II capsule marker *kpsMT II* was not detected in human-clinical isolates. Roughly 67% of the human-clinical population (91/136) were defined as ExPEC, representing the majority of the population. ExPEC strains were identified predominantly in phylogroup B2, however 9 ExPEC strains were also detected within phylogroups D, E, and F. ExPEC strains were prevalent among several different sequence types, including STs 131, 73, 95, 12, 404, 648, 88, 69, and 38, indicating the clonal distribution of ExPEC within the human-clinical population of *E. coli*. The difference between the prevalence of ExPEC in the Nottingham human-clinical and non-human populations of *E. coli* revealed in this study is statistically significant (66.9% prevalence in human-clinical vs 8.6% prevalence in non-human; $p < 0.0001$, two tailed Fisher's test). This highlights the possibility that an obvious non-human reservoir of human ExPEC does not exist and therefore, the non-human population of *E. coli* is unlikely to contribute significantly to the weight of human ExPEC infections, in comparison to the human-clinical population *E. coli*.

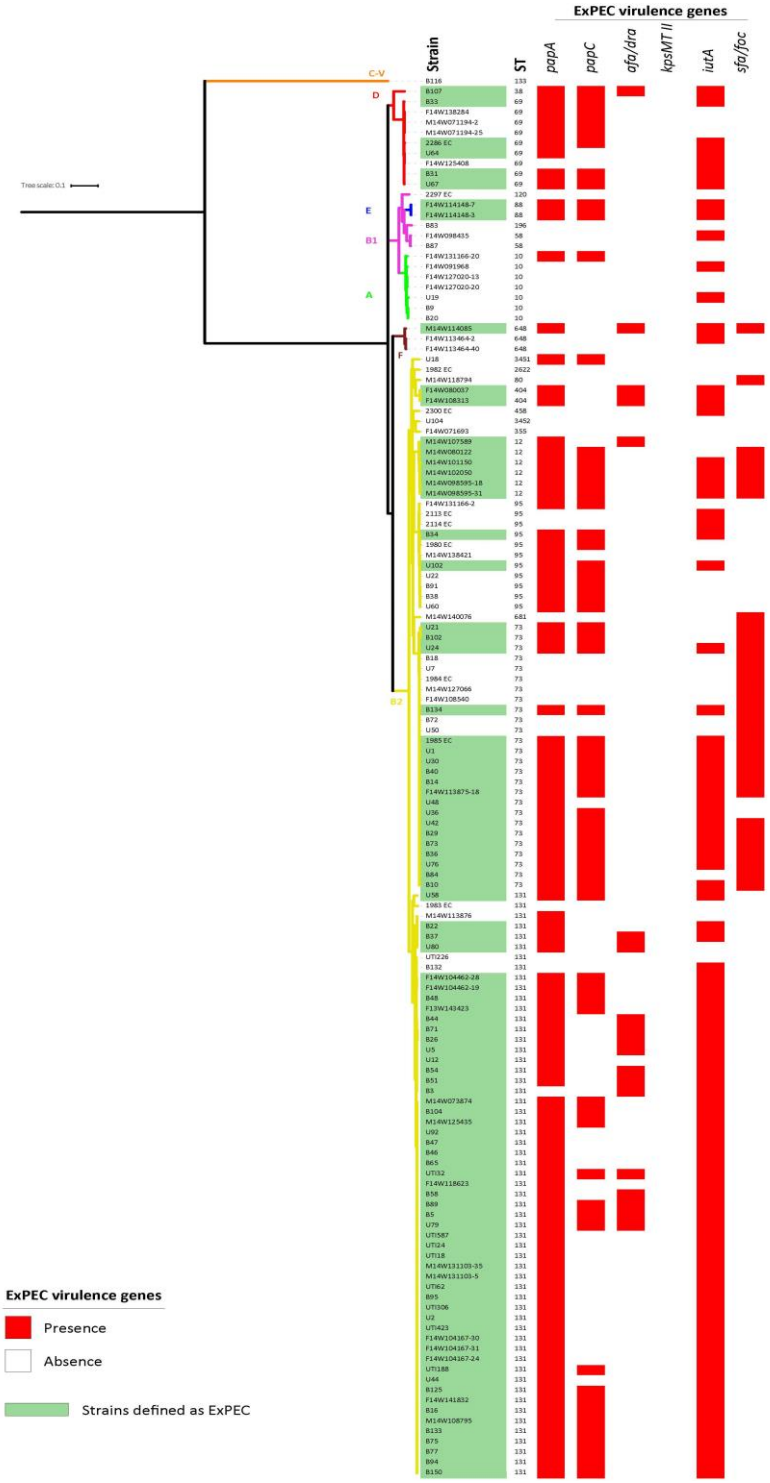


Figure 5.5. Distribution of ExPEC virulence-associated genes among the population of 136 *E. coli* strains isolated from human-clinical samples in Nottingham. Human-clinical *E. coli* genomes (Table 5.1) were mass screened for VAG carriage, by running the ABRicate bioinformatics tool to scan the Virulence Factors Database. The presence of two or more of the following VAGs were used to define the ExPEC pathotype: *papA* and/or *papC*; *afa/dra*; *kpsMT II*; *iutA*; and *sfa/foc*. VAGs are mapped onto the maximum-likelihood phylogenetic tree as red-coloured bars and strains identified as ExPEC are shown in green. ST designations, as determined by *in silico* MLST, are also indicated on the tree. The difference in ExPEC prevalence between human-clinical and non-human *E. coli* is statistically significant (66.9%, human-clinical vs 8.6%, non-human; $p < 0.0001$, two-tailed Fisher's test).

5.3.4. Distribution of antimicrobial resistance genes among human-clinical *E. coli*

In chapter 4, *in silico* antimicrobial resistance gene profiling revealed a diverse range of resistance determinants detected in the Nottingham non-human population of *E. coli*, which corresponded to 8 different antibiotic classes (Fig. 4.6; section 4.3.6). However, contrary to previous reports of MDR *E. coli* consistently being reported from the environment and food sources, a very low prevalence of ESBL genes was observed in the non-human population of *E. coli*. This suggests that multidrug resistance to the extended spectrum of β-lactam antibiotics likely occurs at a very low frequency. To compare the prevalence of antimicrobial resistance genes in the human-clinical population with that of the non-human population, all 136 human-clinical *E. coli* genomes were also screened for the presence of acquired antimicrobial resistance genes. This was achieved *in silico*, by running the bioinformatics pipeline ABRicate, which scans the ResFinder database to generate antibiotic resistance gene profiles for each isolate (Fig. 5.6).

With regards to the human-clinical population of *E. coli*, a total of 51 different resistance determinants were identified, compared to 46 genes identified in the non-human population. These genes corresponded to 11 different antibiotic classes, 8 of which were identified in the non-human population, and 3 additional classes: polymyxins, quinolones, and glycopeptides (vancomycin). Noticeable similarities between the two populations would include the high prevalence of aminoglycoside resistance genes. The streptomycin resistance genes *strA*, *strB*, *aad*A1, and *aad*A5 were detected at high frequencies throughout the human-clinical population. Additionally, a high prevalence of β-lactam ($bla_{\text{TEM-1B}}$), sulphonamide (*sul1* and *sul2*), and tetracycline (*tet*(A), *tet*(B), and *tet*(34)) resistance genes were reported among human-clinical isolates, similar to the non-human population of *E. coli*. On the other hand, some observable differences were also noted between the acquired resistance gene profiles of both populations. The human-clinical population revealed a higher prevalence of chloramphenicol resistance genes, particularly *cat*A1 and *cat*B3, which were identified in 26.5% of all human-clinical isolates, compared to the lower prevalence (3.9%) among non-human *E. coli* isolates. Furthermore, there is a clear contrast between the frequencies of trimethoprim resistance genes observed for both populations. The prevalence of *dfr* alleles, which confer resistance to trimethoprim, was generally low for the non-human population, whilst a high prevalence of *dfr* gene types was identified among human-clinical isolates, in particular *dfr*A17 (33% prevalence), which has frequently been reported in clinical *E. coli* isolates (Seputiene *et al*., 2010). This correlates well with the high level of phenotypic trimethoprim resistance observed for human-clinical isolates of *E. coli* in a previous NTU-based study (Croxall *et al*., 2011b). Additionally, the macrolide resistance gene *mph*(A), which was detected in only two isolates of the non-human population, was considerably more prevalent in the human-clinical population (29.4% of isolates). However,

additional macrolide resistance genes (*mph*(B), *Inu*(F), *Inu*(B), and *Isa*(A)) present in the non-human population were not detected in the human-clinical population. This may suggest that although *E. coli* are intrinsically resistant to macrolide antibiotics (Gomi *et al*., 2017b), there is a disparity in the type of macrolide resistance genes that are acquired and conserved in both populations, perhaps due to the selective pressures that they encounter. A noticeable pattern that can be observed from the antimicrobial resistance gene profiles of the human-clinical population is the concentration of genes in the ST131 clade of phylogroup B2 (Fig. 5.6). These isolates demonstrate a high prevalence of all the major resistance determinants associated with this population, in particular *aad*A1 and *aad*A5, $bla_{\text{CTX-M-15}}$, $bla_{\text{OXA-1}}$, *mph*(A), *sul*1, and *dfrA17*. This may suggest that human-clinical ST131 in this population is likely to be comprised of mainly ST131 clade C isolates, which is the most dominant lineage (currently up to 80% of global ST131 belong to clade C) (Nicolas-Chanoine, Bertrand and Madec, 2014) and is often associated with $bla_{\text{CTX-M-15}}$ carriage.

One of the significant observations made between the two populations is the appreciable difference in prevalence of ESBL genes. Results from this study have revealed a very low prevalence of ESBL genes in the non-human population of *E. coli*, through molecular detection by multiplex PCR assays of 230 isolates (in section 4.3.2 of chapter 4), and confirmed by *in silico* antimicrobial resistance gene profiling of the 128 sequenced non-human *E. coli* genomes (in section 4.3.6 of chapter 4). While the emergence and dissemination of the CTX-M family of ESBLs among *E. coli* within the community is of particular concern, the $bla_{\text{CTX-M}}$ family of genes could not be detected in the non-human population of *E. coli*. By contrast, however, $bla_{\text{CTX-M}}$ type ESBL genes were identified in 21.3% of all human-clinical *E. coli* isolates (Fig. 5.6), with $bla_{\text{CTX-M-15}}$ being the most prevalent ESBL gene (present in 20.6% of all isolates). Furthermore, bla_{OXA} genes, namely $bla_{\text{OXA-1}}$ and $bla_{\text{OXA-10}}$, were identified in only 1.6% of non-human *E. coli* isolates analysed, whereas $bla_{\text{OXA-1}}$ predominated in 21.3% of all human-clinical isolates analysed, indicating a significant difference between the two populations (1.6% vs 21.3%; $p < 0.0001$, two tailed Fisher's test). SHV-type ESBLs, which were not detected in the non-human population, were less prevalent than the other ESBLs and were identified in only 2.9% of all human-clinical isolates. However, the high prevalence of $bla_{\text{CTX-M}}$ and bla_{OXA} genes in the Nottingham human-clinical population of *E. coli* is consistent with previous reports of ESBLs being commonly detected in human-clinical isolates (Kallonen *et al*., 2017; Alhashash *et al*., 2013; Croxall *et al*., 2011b; Lau *et al*., 2008), suggesting that these isolates may act as reservoirs for ESBLs. The absence of $bla_{\text{CTX-M}}$-carrying plasmids in the non-human population of *E. coli* analysed would indicate that horizontal transfer of such genetic elements between isolates from non-human sources occurs at low frequencies. This would suggest that non-human isolates of the Nottingham *E. coli*

population do not readily encounter ESBL-producing strains of the human-clinical population, and horizontal gene transfer and recombination between the two populations are likely to be limited.

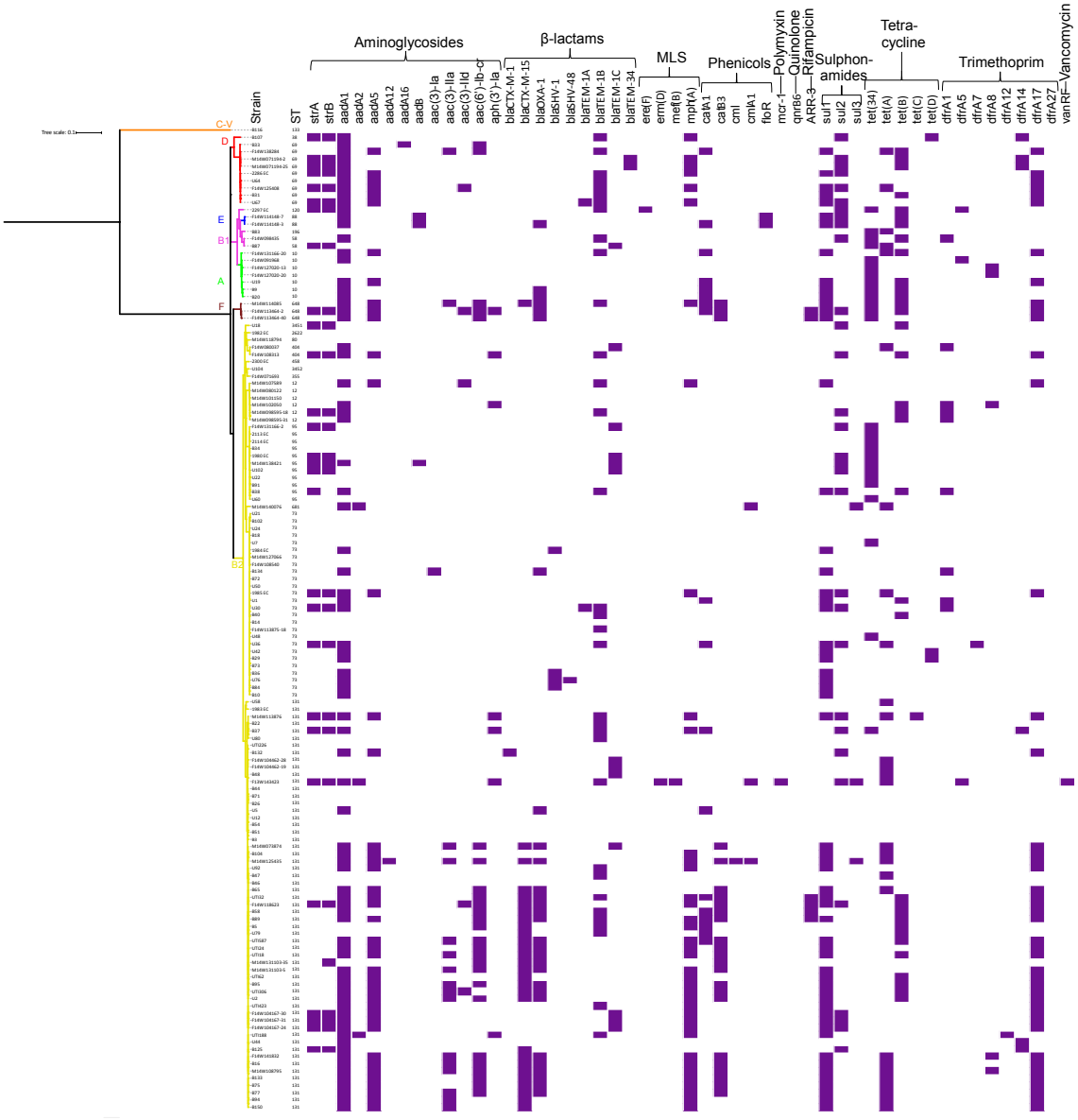


Figure 5.6. Distribution of antibiotic resistance gene profiles across the population of 136 *E. coli* strains isolated from human-clinical samples in Nottingham.

The genomes of the 136 human-clinical *E. coli* strains included in the study population were mass screened for antibiotic resistance gene carriage, by running the ABRicate bioinformatics tool using the ResFinder database. Presence of resistance determinants are shown on the phylogenetic tree as purple-coloured bars and are grouped by antibiotic class. ST designations, as determined by *in silico* MLST, are also indicated next to the taxa on the tree. The human-clinical population showed higher prevalence of chloramphenicol and trimethoprim resistance genes, in addition to ESBLs (*bla*$_{CTX-M}$ and *bla*$_{OXA}$), when compared to the non-human population (section 4.3.6, Fig. 4.5).

5.3.5. Phylogenetic analyses of the *E. coli* lineages ST131 and ST648

5.3.5.1. *E. coli* ST131

Of the limited number of clinically significant sequence types identified in the non-human population of *E. coli*, only a single strain belonging to the ST131 clonal group was isolated from river water samples. The ST131 lineage of extraintestinal pathogenic *E. coli* has been rapidly globally disseminated to become the dominant MDR strain of *E. coli* from urinary tract and bloodstream infections, across the globe (Banerjee and Johnson, 2014). The phylogenetic structure of the ST131 lineage consists of three distinct clades (Petty *et al*., 2014). These are defined as clades A, B, and C, of which clade C, also known as H30Rx, is associated with the rapid expansion and global dissemination of MDR isolates carrying the $bla_{CTX\text{-}M\text{-}15}$ ESBL gene. Core genome phylogenetic analysis of the human-clinical and non-human populations of *E. coli* (Fig. 5.4) did not provide significant genomic distinction between strains belonging to the same clonal group. To gain further resolution between closely related strains found in both human-clinical and non-human sources, these strains must be analysed in a wider context of strains within the same clonal group.

To determine the position of the single non-human ST131 isolate from this study in a wider population of the ST131 lineage, it was included in a global collection of 242 *E. coli* ST131 genome sequences, from multiple ecosystems (Table 5.2). In addition to the river water isolate, 125 were avian (wild birds), domesticated animal (cats and dogs), livestock (cattle), and human-clinical isolates, sequenced as part of a recent study by McNally *et al*. (2016a); 102 were human-clinical isolates from a previous phylogenomic study (Petty *et al*., 2014); and the remaining 14 were human-clinical isolates obtained from the NTU Pathogen Research Group strain collection (unpublished). A core genome alignment and maximum likelihood phylogeny was obtained from localised co-linear blocks (length = 3,727,932 bp, total SNPs = 287,012) for all 242 genomes, which confirmed the 3-clade structure of *E. coli* ST131, as previously described (Fig. 5.7). This analysis revealed that the river water strain (GD45) had clustered with clade B, which appears to be the most genetically diverse clade, as indicated by longer branch lengths and all isolate sources are represented in this clade (Fig. 5.7). It is therefore important to note that GD45 is not a clade C strain, which is the most globally dominant lineage. Global longitudinal studies revealed that clade B was predominant among ST131 before the 1990s, however, since the 2000s clade C has become the most dominant lineage with up to 80% of global ST131 belonging to clade C (Nicolas-Chanoine, Bertrand and Madec, 2014). Clade C is mostly fluoroquinolone-resistant and is often associated with carrying the ESBL gene $bla_{CTX\text{-}M\text{-}15}$, whilst clade B is most often fluoroquinolone-susceptible and rarely carries plasmids with $bla_{CTX\text{-}M\text{-}15}$. This is consistent

with the GD45 strain isolated in this study, which is ESBL-negative. Given the position of GD45 within clade B, it can be deduced that this isolate is not associated with the lineage of ST131 that is currently responsible for the majority of extraintestinal infections in humans and is therefore unlikely to be of significant concern to human health.

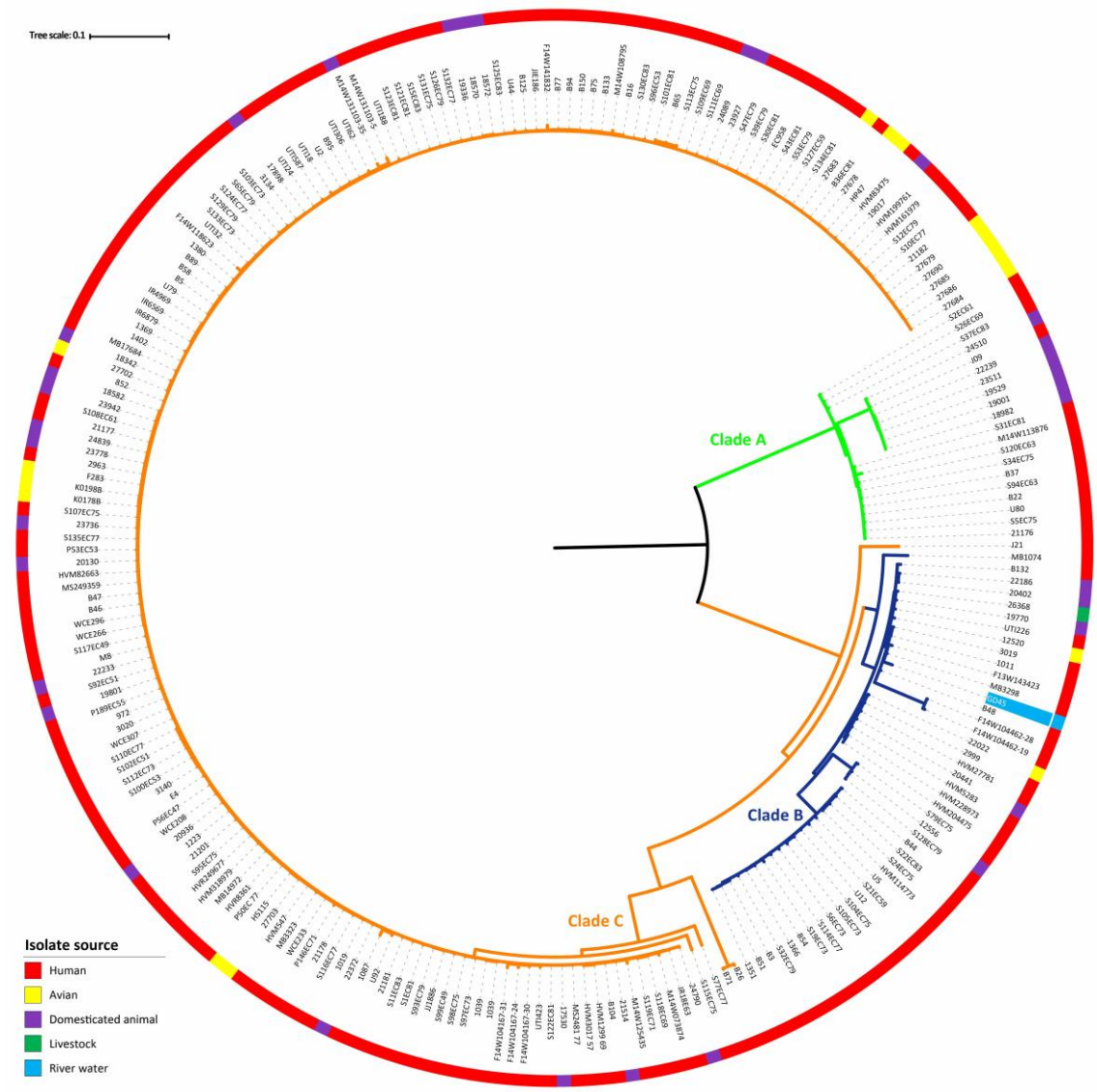


Figure 5.7. Maximum likelihood SNP-based phylogenetic tree of 242 *E. coli* ST131 isolates.

A global collection of ST131 strains isolated from humans, wild birds (avian), dogs and cats (domesticated animals), cattle (livestock), and river water are indicated by coloured bars at the tips of the tree. The phylogeny was inferred from a core genome alignment of all strains (length = 3,727,932 bp, total SNPs = 287,012, 242 genomes) constructed using Parsnp. The phylogenetic tree was visualised and edited using iTOL. The names of the taxa correspond to those listed in Table 5.2. The phylogenetic clades A, B, and C are indicated by colour coding of the branches and the ST131 strain isolated from river water in this study (GD45) is highlighted in blue within clade B.

5.3.5.2. *E. coli* ST648

ST648 is another ExPEC-associated sequence type, in addition to ST131, which is associated with worldwide dissemination of *E. coli* producing the CTX-M β-lactamase (Pitout, 2012). ST648, which belongs to phylogroup F, has been identified previously in livestock and companion animals (Ewers *et al*., 2012). Of particular concern, are ST648 strains with an ESBL phenotype which have been observed globally in human patients, and more incidentally from domesticated and wild animals in Europe and Asia (Guenther, Ewers and Wieler 2011; Guenther *et al*., 2010; Zong and Yu, 2010; Cortes *et al*., 2010). The frequent occurrence of CTX-M-15-type ESBLs in ST648 isolates, among samples from predominantly companion animals and horses, highlights the widespread dissemination of this ESBL-producing genotype, which has been detected in livestock, wildlife, and humans. A previous study by Ewers and co-authors (2014) suggested that the prevalence of CTX-M-15 – a predominantly human-linked β-lactamase type – among companion animal ST648 isolates indicates that a mutual exchange of such strains between clinical, community, and environmental surroundings is likely.

In the present study, only a single isolate of ST648 was identified among the non-human population of *E. coli* analysed. To determine the phylogenetic relation of this strain within the context of a wider population of the ST648 lineage, it was included in a European collection of 89 *E. coli* ST648 genome sequences, from multiple host sources (Table 5.3). In addition to the river water isolate, 88 strains were isolated from different hosts including humans, companion animals (cats, dogs, and horses), and wild birds, across Western Europe. A core genome alignment and maximum likelihood phylogeny was obtained from localised co-linear blocks (length = 4,002,826 bp, total SNPs = 208,172) for all 89 genomes (Fig. 5.8). The phylogeny reveals three predominant clades that are observable in the population, with the Nottingham river-water isolate (GD49) positioned within the population of European isolates (not forming a separate branch). Annotation of identified CTX-M types onto the tree exposes a predominance of the $bla_{\text{CTX-M-15}}$ ESBL type in the population. An overwhelming majority of approximately 88% (78/89) of the population were found to harbour a CTX-M ESBL gene, with $bla_{\text{CTX-M-15}}$ being detected in roughly 72% (64/89) of the population. The 14 other CTX-M-positive isolates were found to carry $bla_{\text{CTX-M-1}}$, $bla_{\text{CTX-M-3}}$, $bla_{\text{CTX-M-14}}$, and $bla_{\text{CTX-M-32}}$. There were only 11 isolates in the population in which CTX-M type ESBLs could not be detected, including GD49 isolated from the current study. Interestingly, GD49 clusters with the clade of primarily CTX-M-negative isolates, and isolates carrying $bla_{\text{CTX-M-3}}$, $bla_{\text{CTX-M-14}}$, and $bla_{\text{CTX-M-32}}$ were also present, whilst only two carriers of $bla_{\text{CTX-M-15}}$ were identified within this clade. The majority of CTX-M-15-producing isolates therefore constitute the other two clades within the population. This suggests that a level of genetic diversity exists within this population of ST648, where strains have lost their CTX-

M-associated plasmid or acquired different CTX-M-associated plasmids, perhaps due to ecological separation and encountering a change in environment.

It is also interesting to note that despite the geographical separation between the ESBL-negative GD49 strain (isolated from Nottingham) and the rest of the population, consisting of primarily MDR isolates from animals (isolated from mainland Western Europe), GD49 still falls within the tree indicating its phylogenetic relation to the European population of ST648. This would support a widespread dispersion of this sequence type, as has been reported globally (Guenther *et al*., 2012; Cortes *et al*., 2010; Sidjabat *et al*., 2009), and it may also imply that the ST648 strain isolated from the environment could originate from humans, companion animals, or livestock. Based on the association of the $bla_{\text{CTX-M-15}}$ ESBL type in ST648 isolates with human-clinical cases observed in previous studies (Zong and Yu 2010; Cortes *et al*., 2010; Nicolas-Chanoine *et al*., 2008), it would seem unlikely that ST648 isolates that do not possess a CTX-M-carrying plasmid, in particular $bla_{\text{CTX-M-15}}$, can be attributed to the global expansion of ST648 in humans. Having identified an ESBL-negative ST648 strain in the non-human (river water) population of *E. coli* in this study, it may suggest that ST648 strains in this population pose little clinical significance to humans.

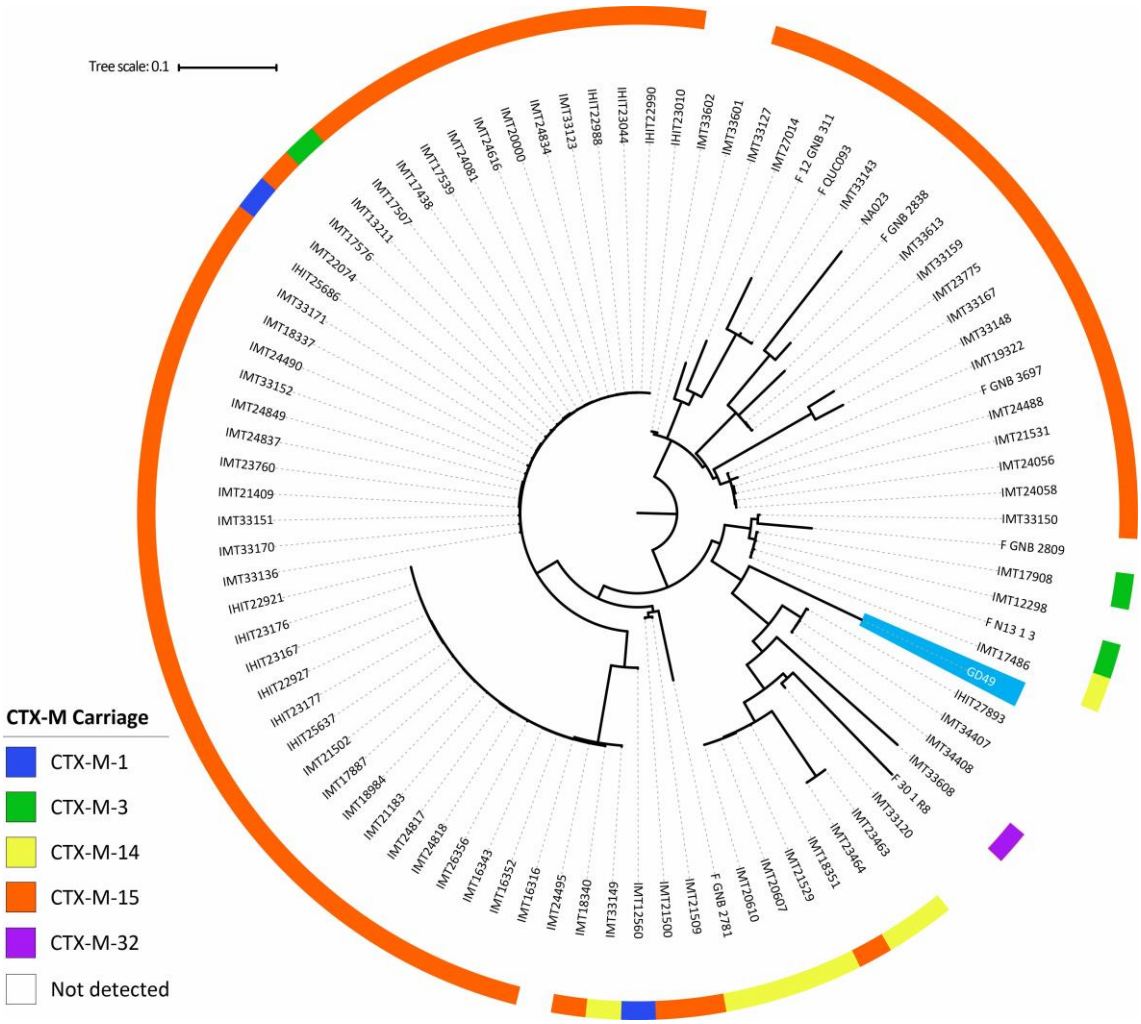


Figure 5.8. Maximum likelihood SNP-based phylogenetic tree of 89 *E. coli* ST648 isolates.

A European collection of ST648 strains isolated from humans, companion animals (cats, dogs, and horses), wild birds, and river water are included in the phylogenetic tree. The phylogeny was inferred from a core genome alignment of all strains (length = 4,002,826 bp, 89 genomes) constructed using Parsnp. The phylogenetic tree was visualised and edited using iTOL. *bla*$_{\text{CTX-M}}$ carriage, as determined by *in silico* resistance gene screening using ABRicate, is annotated on the tips of the tree as coloured bars. The names of the taxa correspond to the strains listed in Table 5.3. The single ST648 strain isolated from river water in this study (GD49) is highlighted in blue on the tree. Five different CTX-M types were identified in the population: *bla*$_{\text{CTX-M-1}}$, *bla*$_{\text{CTX-M-3}}$, *bla*$_{\text{CTX-M-14}}$, *bla*$_{\text{CTX-M-15}}$, and *bla*$_{\text{CTX-M-32}}$. *bla*$_{\text{CTX-M-15}}$ was found to be the most frequently detected CTX-M type, whilst GD49 was one of the few strains that did not possess a CTX-M-carrying plasmid.

5.3.6. Comparative genomic analyses

5.3.6.1. Pan-genome approach to compare human-clinical and non-human *E. coli*

Comparative analyses between the non-human and human-clinical populations of *E. coli* in this study has thus far suggested two distinct populations of *E. coli* in Nottingham. If this is the case, then significant levels of genetic exchange would not be expected between the two populations. To be able to accurately determine and compare the entire gene contents of multiple genomes, a pan-genome approach was applied when comparing the human-clinical and non-human populations of *E. coli* in this study. Pan-genome analysis offers a higher resolution than core genome phylogeny and MLST typing methods, as it analyses the entire bacterial genome (Hall, Ehrlich and Hu, 2010). Since there are only a limited number of clinically relevant STs present in the non-human population, all strains were included in the pan-genome to look for any patterns of gene flow between the two populations. Determination of the core and pan-genomes of all 264 human-clinical and non-human *E. coli* strains of the Nottingham study population was achieved using the Roary pan-genome bioinformatics pipeline (Page *et al*., 2015).

Overall, the pan-genome of the Nottingham *E. coli* population was composed of 69,645 genes, of which, only 596 genes were represented at least once in ≥ 95% of strains (Fig. 5.9A). Despite the presence of 10 cryptic clade *Escherichia* strains in the population, this would indicate a species with a large dispensable/accessory genome, reflecting the diversity of *E. coli,* with only a small proportion of core genes conserved for basic biological and phenotypic functions. This was confirmed by re-running the pan-genome analysis on all *E. coli sensu stricto* strains (n = 254) by excluding the 10 cryptic *Escherichia* strains (Fig. 5.9B). The analysis revealed that the number of core genes represented at least once in ≥ 95% of strains remained unchanged (596 core genes), suggesting a very small core genome for the species. A large number of strain-specific genes are therefore present in the *E. coli* pan-genome, as well as genes encoding species diversity and providing selective advantages such as niche adaptation, antibiotic resistance, and virulence factors. Of the total number of genes which make up the pan-genome of the study population, 29,139 genes were found to be unique strain-specific genes (a gene possessed by exactly one isolate), and the frequency of gene occurrence plot for the pan-genome of all 264 strains (Fig. 5.9A) indicates that a minimal number of core genes are shared by all strains of the study population. Determining the level of horizontal gene transfer of dispensable accessory genes between the two populations would provide an insight into the frequency of strain movement, and thus, whether the human-clinical and non-human populations of *E. coli* overlap. To investigate this, analysis of the accessory genomes of all strains is required.

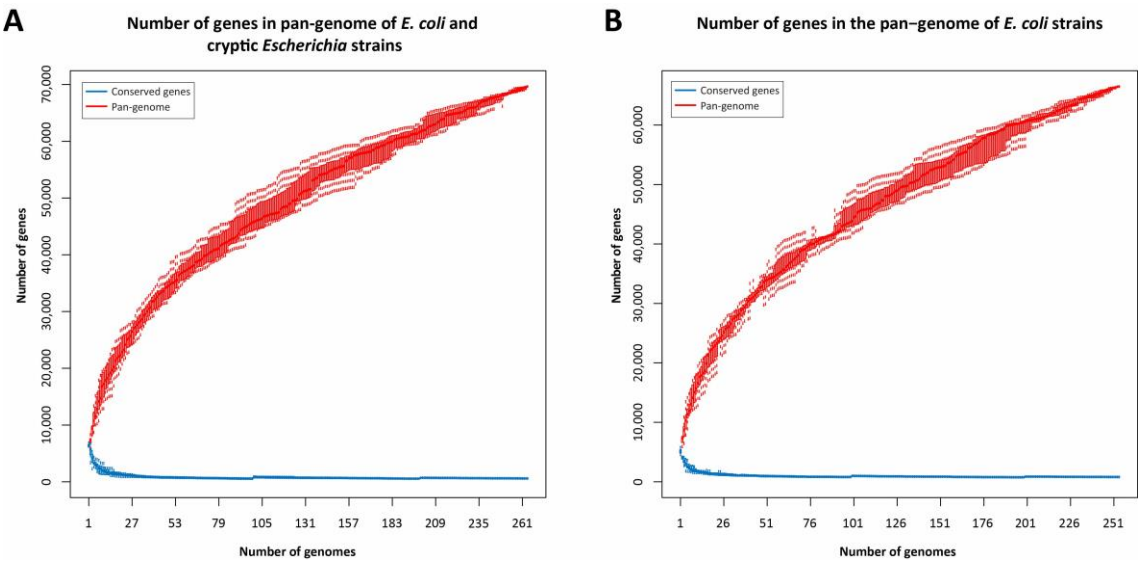


Figure 5.9. (A) Core and pan-genome frequency plots of all 264 *E. coli* and cryptic clade *Escherichia* strains isolated from human-clinical and non-human samples in Nottingham. (B) Core and pan-genome frequency plots of all 254 *sensu stricto E. coli* strains isolated from human-clinical and non-human samples in Nottingham.

The box and whisker plots display the number of genes on the y-axis against the number of genomes on the x-axis. The pan-genome was determined using the Roary pan-genome bioinformatics pipeline and the plots were generated using the R plots (ggplot2) flag. As the number of genomes increases, the total number of genes in the pan-genome increases (shown in red). The pan-genome analysis was first run on all 264 *E. coli* genomes, inclusive of 10 cryptic clade *Escherichia* strains (A). A total of 69,645 genes were identified in the pan-genome of all 264 strains, of which, only 596 genes were represented at least once in ≥ 95% of strains (shown in blue as 'conserved genes'). The pan-genome analysis was re-run on the *E. coli* population, excluding the 10 cryptic clade *Escherichia* strains (B). A total of 66,477 genes were identified in the pan-genome of 254 strains, while the number of core genes represented at least once in ≥ 95% of strains remained unchanged (596 core genes).

5.3.6.2. Comparisons of the accessory genomes of human-clinical and non-human *E. coli*

Comparisons between the accessory genomes of all 264 human-clinical and non-human *E. coli* strains would allow for a determination of the proportion of accessory gene clusters or loci unique to either population, as well as the proportion that is shared between both populations. Comparing the accessory genomes of human-clinical and non-human *E. coli* isolates would provide an insight into whether gene sharing occurs frequently or is restricted between the two populations. Additionally, accessory gene sharing between strains of the two populations may provide some indication of whether strain movement occurs readily between the two populations. For this purpose, genes considered to be core for all strains were excluded from this analysis to allow for accessory genome comparisons. Genes present in at least 85% of strains conforms to the definition of "soft core" genes (Gordienko, Kazanov and Gelfand, 2013), and it was therefore decided to use this as exclusion criteria for this particular analysis. The accessory genomes of all strains were obtained by excluding genes present in ≥ 85% of strains from the pan-genome matrix, which would result in the removal of primarily core genes from the pan-genome. To confirm that the excluded genes represent predominantly core/soft core genes of the *E. coli* genome, the gene product functions of these genes were evaluated and their proportions are presented in Figure 5.10. It was revealed that the majority of these genes encode for basic biological and phenotypic functions, such as DNA replication, lipid biosynthesis, cell wall and plasma membrane maintenance, motility, respiration, cell division, hydrolase activity, and stress response. The most common gene functions encoded by these genes included transport, protein biosynthesis and general metabolism of the cell. Gene functions for cell adhesion (1.32%), antibiotic resistance (0.74%), enzyme inhibitors (0.37%), and antibiotic biosynthesis (0.16%) were also reported, however these are negligible in comparison to the major gene functions identified. This analysis confirmed that the excluded genes predominantly encode for the basic biochemical and phenotypic functions associated with the core genome of *E. coli*, justifying the exclusion of these genes from further analysis.

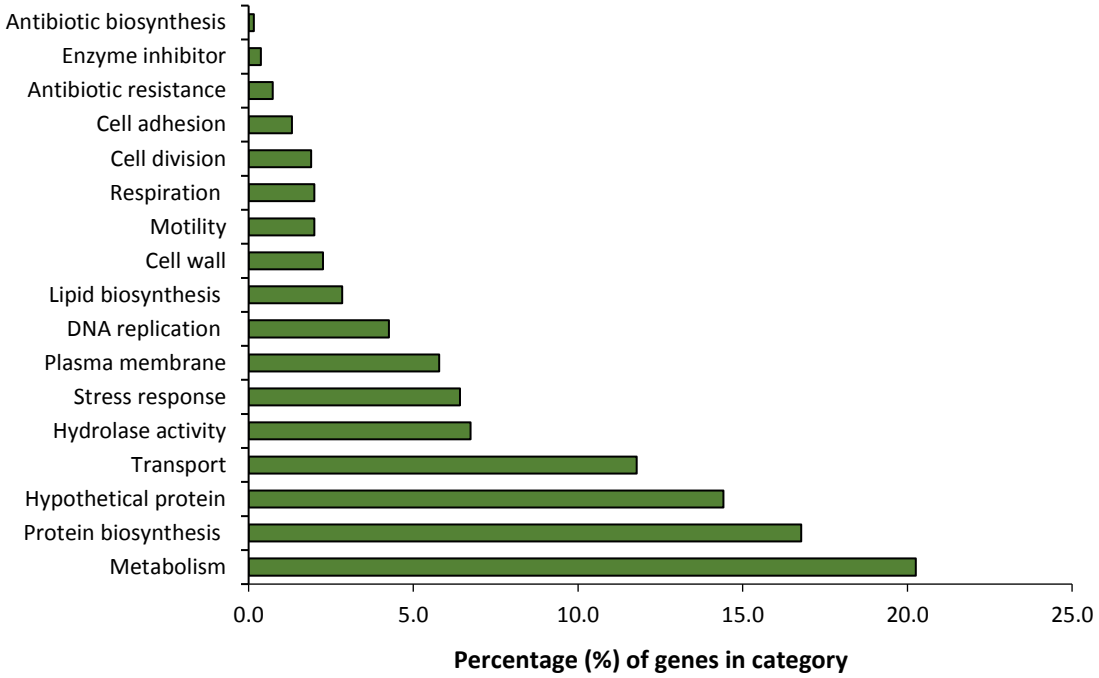


Figure 5.10. Functional categories of genes present in ≥ 85% of all 264 *E. coli* strains isolated from human-clinical and non-human samples in Nottingham.

The graph shows the distribution of functional categories for genes present in ≥ 85% of the 264 genomes included in the pan-genome analysis (section 5.3.6.1), with percentage prevalence of genes in each category shown on the x-axis. Genes were identified by creating a pan-genome of all genomes, using the Roary bioinformatics pipeline. The graph confirms that these genes predominantly encode for the basic biochemical and phenotypic functions associated with core/soft core genes of *E. coli*, thus justifying the exclusion of these genes for accessory genome analyses.

The resulting accessory gene presence/absence matrix was analysed using the Roary query_pan_genome script, to identify the genetic loci uniquely associated with the human-clinical population or the non-human population, as well as genes that are associated with both populations of strains. The total number of genes unique to either population and the total number of shared genes between the two populations are presented in the Venn diagram (Fig. 5.11). The analysis revealed that removal of "soft" core genes (present in ≥ 85% of strains) resulted in an accessory genome comprising 64,835 genes, representing a large proportion of the total pan-genome. Of the total number of accessory genes detected, 12,311 genes were unique to human-clinical strains and 31,468 genes were identified among non-human *E. coli* strains exclusively, whilst a total 21,056 accessory genes were present in both populations. This corresponded to ~40% of genes identified in non-human *E. coli* and ~63% of genes identified in human-clinical *E. coli* being shared between the two populations. Scanning the gene functions of the 21,056 accessory genes shared between the two populations revealed that 150 antimicrobial resistance genes and 785 phage-associated elements were identified. In the context of shared accessory genes, these numbers are relatively high indicating that there is likely a high number of mobile genetic elements present in both populations. This suggests that gene exchange between the two populations does occur and the extent of this should be investigated further. Although numerous accessory genes were identified as being unique to either population, the frequency of gene occurrence plots for genes unique to the human-clinical (Fig. 5.12A) and non-human (Fig. 5.12B) populations indicate a large proportion of strain-specific genes for each population. Of the 12,311 genes unique to the human-clinical population, 8,879 genes were strain-specific genes, whilst 20,039 of the 31,468 genes unique to the non-human population were also strain-specific genes. This equates to 72% and 63% of unique genes being strain-specific genes in the human-clinical and non-human populations, respectively. The frequency of gene occurrence plot for the 21,056 genes present in both populations (Fig. 5.12C) reveals a more even distribution of accessory genes, with the number of shared genes decreasing as the number of genomes included increases. This was expected given that these genes are, by definition, not strain-specific. The majority of the shared genes would appear to be ST- or clade-specific genes, while only 17 common genes are prevalent in a maximum of 209 out of 264 genomes across the two populations. Due to the high genomic diversity of strains included in the study population, the number of unique genes detected in each population is skewed towards strain-specific genes, which does not provide a good indicator of how genetically distinct human-clinical and non-human *E. coli* are as populations. On the other hand, these observations would suggest that examining the gene set of shared accessory genes between the two populations in further detail would provide a higher resolution for genomic comparison of strains isolated from human-clinical and non-human sources.

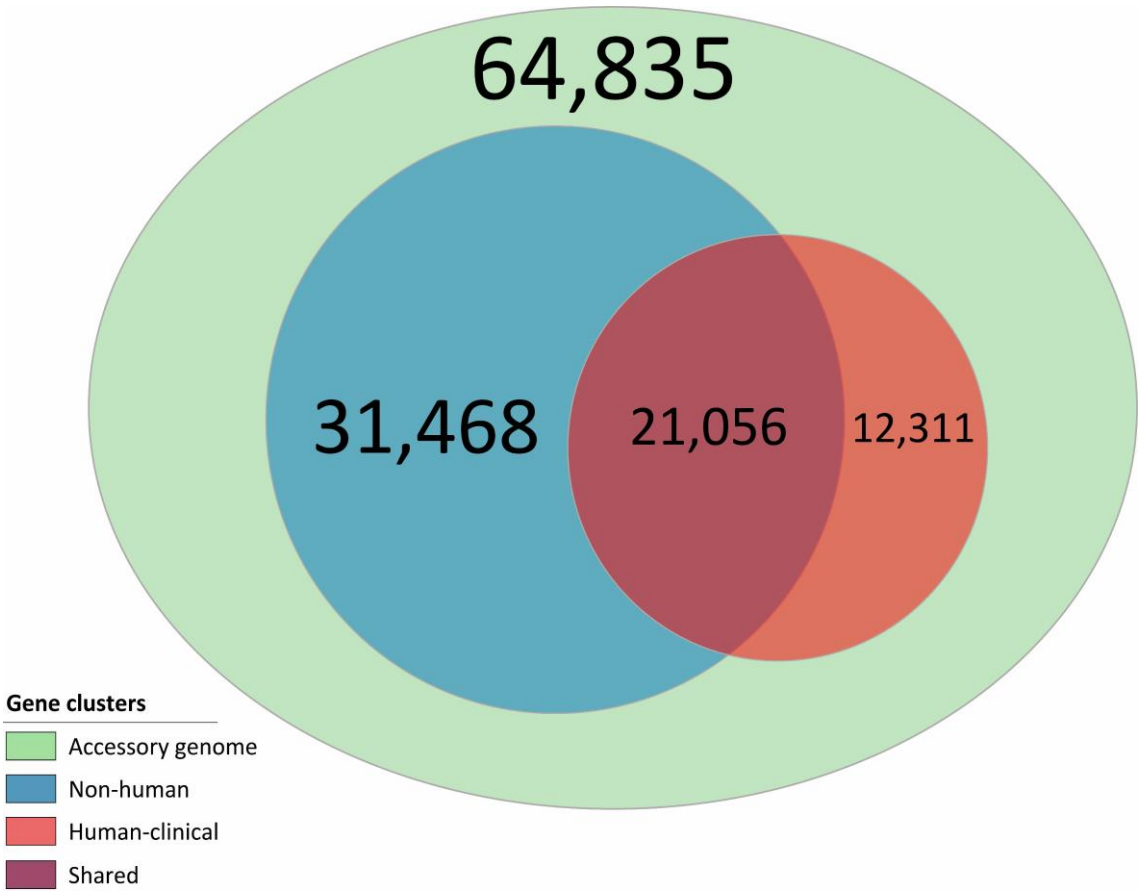


Figure 5.11 Comparison of population-unique accessory genes and shared accessory genes between the human-clinical and non-human populations of *E. coli* in Nottingham.

The Venn diagram displays the number of genes unique to the non-human population of *E. coli* (left, blue), unique to the human-clinical population of *E. coli* (right, red), and shared between the two populations (centre, mauve). The total number of accessory genes comprising the two populations is shown in green. The accessory genome was determined by generating a pan-genome matrix using the Roary pipeline and excluding genes present in ≥ 85% of all 264 genomes included in the analysis. Unique and shared genes were identified by running the Roary query_pan_genome script on the accessory gene presence/absence matrix. Of the total 64,835 accessory genes detected, 12,311 genes were unique to human-clinical strains and 31,468 genes were identified among non-human *E. coli* strains exclusively, whilst 21,056 genes were identified in both populations. This corresponded to ~40% of genes identified in non-human *E. coli* and ~63% of genes identified in human-clinical *E. coli* being shared between the two populations.

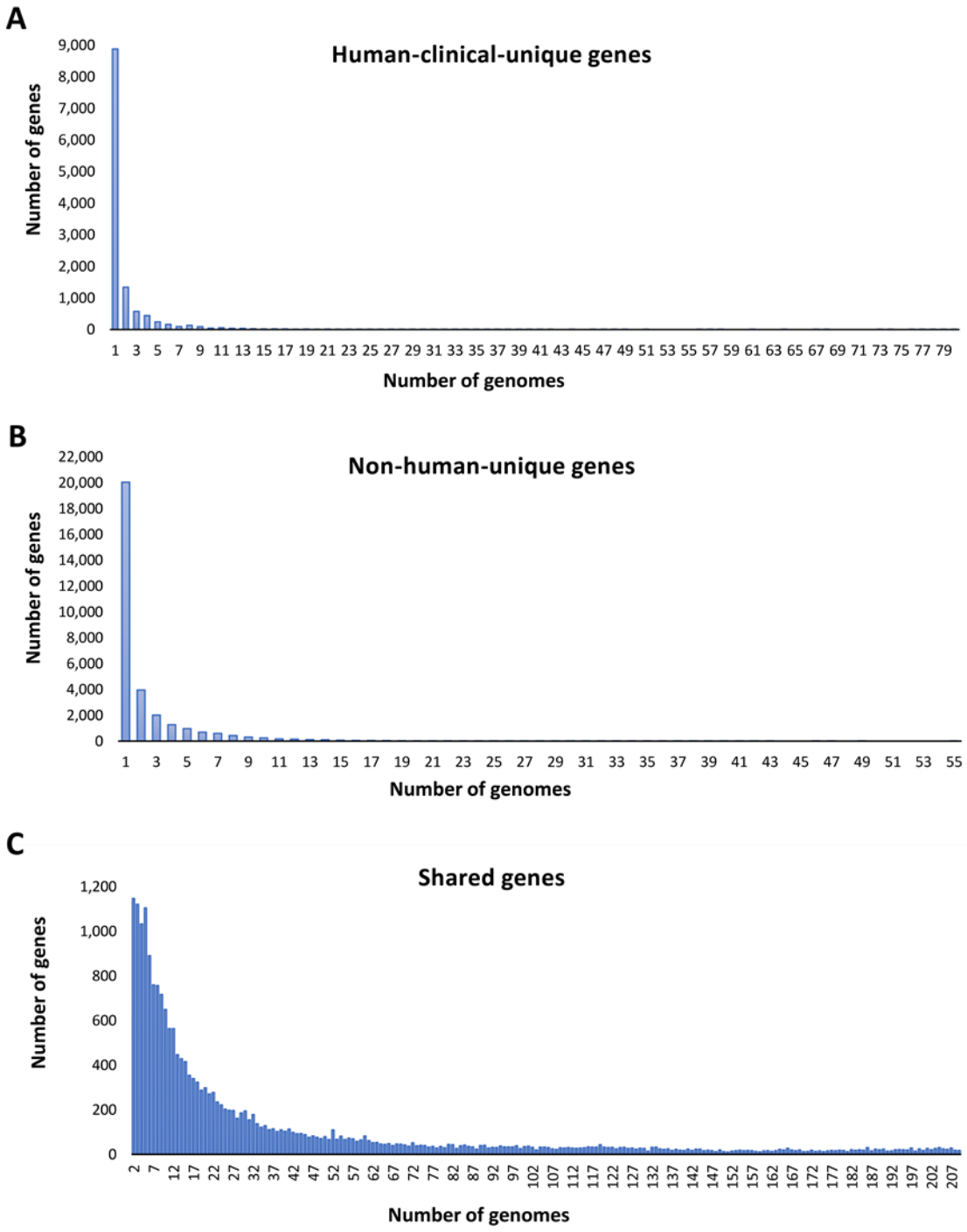


Figure 5.12. Frequency of gene occurrence plots for population-unique and shared accessory genes for the human-clinical and non-human populations of *E. coli* in Nottingham.

The graphs plot the number of genes on the y-axis against the number of genomes on the x-axis for accessory genes unique to the human-clinical population of *E. coli* (A), unique to the non-human population of *E. coli* (B), and shared between the two populations (C). The graphs indicate a large proportion of unique genes being specific to one or a few genomes in each population. The majority of shared genes may represent ST or clade-specific genes, while only 17 genes are prevalent in a maximum of 209 of the 264 genomes across the two populations.

The number of accessory genes identified as present in both populations (21,056) may be considered high relative to the numbers of unique genes reported (Fig. 5.11), particularly for the human-clinical population, so statistical analysis was performed to investigate this further. The probability of a gene being identified from the non-human population is the proportion of shared and non-human-unique genes out of the total, which is 81% ([21,056 + 31,468 / 64,835] × 100). On the other hand, the probability of a gene being identified from the human-clinical population is 51.5% ([21,056 + 12,311 / 64,835] × 100), which corresponds to the proportion of shared and human-clinical unique genes out of the total accessory genome. Assuming that these two probabilities are independent of one another, the expected probability of encountering a gene identified in both populations can be derived by multiplying the two proportions ([81% × 51.5%] / 100), which gives an expected probability of 41.7% of identifying a gene present in both populations. However, the observed frequency of genes shared between the human-clinical and non-human populations reported here is 32.5% ([21,056 / 64,835] × 100). Therefore, it can be deduced that the observed frequency of shared genes equates to approximately 78% of the expected frequency, and the probability of encountering a gene that is present in both populations is 9.2% lower than the expected probability.

The chance of a gene being identified in multiple groups is not independent of the number of genomes the gene is in. Therefore, to provide a statistically robust comparison between the observed and expected proportions of genes shared between the human-clinical and non-human populations, a permutation test was carried out with pseudo-random re-sampling of the population without replacement of genomes. This approach took into account gene frequencies, and all genes that were only present once in each population (i.e. strain-specific genes) were excluded from the analysis. Permutations were carried out 1,000 times iterating for each gene category, and each permutation involved picking the same number of genomes as there are genes in that category. One-tailed empirical p-values were calculated to compare the expected frequencies of gene sharing to those of the observed data set. The simulated proportions are plotted as histograms on the graphs of Appendix 6, with the observed proportions mapped on the graphs for comparison. It is clear from the graphs that the observed data shows far less gene sharing than is expected by chance. The majority of gene categories for genes present in up to 80 genomes show significantly less gene sharing than expected ($p = 0$). Only when the number of genomes is quite high (> 80) does this situation change and the observed proportions correlate with those of the simulated data.

Statistical analysis of the observed data suggests that there is perhaps a restriction in the amount of gene sharing that takes place between the human-clinical and non-human populations, presumably because of the ecological barriers that exist between the two

populations. Environmental selection pressures are likely to play a part in the differences in gene content observed between the two populations. Infectious human ExPEC strains that have infiltrated the urinary tract and bloodstream are generally believed to originate from the faecal flora (Smith, Fratamico and Gunther, 2007). The environmental conditions of the human intestines, urinary tract, and bloodstream are considerably different from those encountered by *E. coli* that inhabit the wider environment, such as river and surface waters. These conditions would impose different selection pressures on the bacteria and therefore the evolution of *E. coli* in non-human environments is likely to be driven by a need to survive and proliferate under extreme environmental stresses, such as desiccation, temperature variation, and osmotic pressure, in addition to surviving transit in food and water (Boor, 2006). *E. coli* strains of the non-human population may have acquired many new advantageous genes, via horizontal gene transfer, from the large genepool of Enterobacteriaceae in the wider environment, which may explain the genetic differences observed between the non-human and human-clinical populations of *E. coli*. Considering that the accessory genome also encodes for the strain diversity within a population of bacteria, the observation of non-human *E. coli* genomes possessing more accessory genes than human-clinical strains is consistent with the high level of genomic variation of these strains, as revealed through phylogenetic analyses in this study.

To explore in more detail the distribution of accessory genes detected in both populations, the accessory gene presence/absence matrix of shared genes was used to annotate the core genome phylogenetic tree, which was visualised with associated metadata using Phandango (Hadfield *et al*., 2018), as shown in Figure 5.13. It is noticeable that there are patterns within the accessory gene profiles which appear to correlate with the ST clusters on the phylogenetic tree. The patterns are not entirely clear-cut for all STs in the population but are most apparent for STs with multiple representative isolates, such as STs 131, 95, 73, 69, 115, 117, 1551, 12, 354, and 648. On closer inspection of the accessory gene profiles for certain STs, some observable differences between strains isolated from human-clinical and non-human samples were noted. One example would include ST69, represented by both human-clinical and non-human strains, in which three strains isolated from non-human samples (I2-18, M3-27, and I2-20) appear to be distinct from the human-clinical strains in this clonal group, as they share fewer accessory genes with the human-clinical population. This is demonstrated to a greater extent in the ST10 cluster, which not only exhibits the highest diversity of strains among all STs, but also reveals a difference in the accessory gene profiles between human-clinical and non-human strains. The majority of non-human strains in this clonal group (AFR-12, AFR-22, AFR-6, AFR-4, M2-3, M2-2, M2-8, M2-4, and I1-24) share fewer genes with the human-clinical population, which may suggest limited gene flow due to ecological barriers between the human-clinical and non-human population of

E. coli in Nottingham. Lineage-specific pan-genome and recombination analyses of human-clinical and non-human ST69 and ST10 strains is required to assess the extent of gene sharing between the two populations in greater detail. This was undertaken in section 5.3.7.

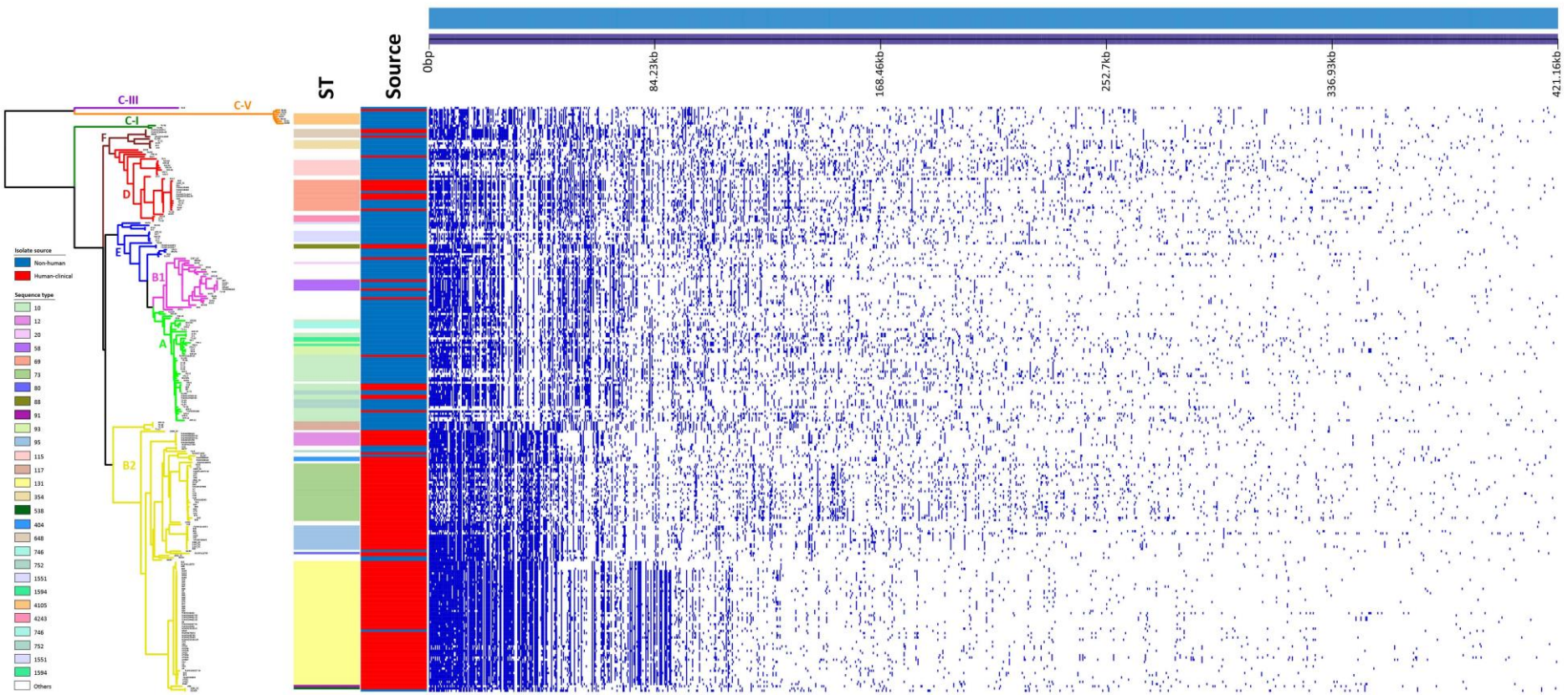


Figure 5.13. Distribution of accessory genes in the human-clinical and non-human populations of *E. coli* in Nottingham.

The heat map shows the presence or absence of accessory loci from the pan-genome of all 264 sequenced *E. coli* strains. The genomes are arranged on the y-axis, according to the maximum-likelihood phylogenetic tree shown on the left, with individual loci on the x-axis. Blue indicates presence of the locus in a strain and white indicates absence. The phylogenetic clades are defined by tree branch colouring and STs and sources of isolation are annotated on the tree as coloured bars.

5.3.7. Pan-genome and recombination analysis of *E. coli* lineages ST69 and ST10

Although there are only a limited number clinically relevant STs that are represented in both the human-clinical and non-human populations in this study, multiple representative strains of ST69 and ST10 have been identified among both populations. *E. coli* ST10 strains are natural colonisers of the human gastrointestinal tract and are widespread in the environment. They are usually associated with low virulence and antibiotic susceptibility (Manges and Johnson, 2012), however, several recent case studies have identified an association of ST10 isolates with human disease, ESBL carriage, and livestock (Peirano *et al*., 2012; Leverstein-van Hall *et al*., 2011; Cortes *et al*., 2010). The ST69 clonal group, on other hand, is well established among human ExPEC infections and ST69 strains have been isolated worldwide from cases of UTI and bacteraemia, from both hospital- and community-acquired infections (Kallonen *et al*., 2017; Dias *et al*., 2009). The presence of these STs in both the human-clinical and non-human populations of *E. coli* analysed in this study warranted further investigation. Pan-genome analysis and detection of core genome recombination events was performed exclusively for ST69 and ST10 strains, to further compare the gene content of human-clinical and non-human strains, and assess the level of genetic exchange occurring between related strains isolated from the two different environments.

5.3.7.1. *E. coli* ST69

The pan-genomes of all 14 ST69 strains, identified among the Nottingham human-clinical and non-human *E. coli* populations analysed in this study, were determined using Roary. The resulting gene presence/absence matrix was used to annotate the core genome phylogenetic tree of ST69, which was visualised with isolate source data (Fig. 5.14). The pan-genome of this clonal group comprised 10,673 genes, 3,299 of which were core genes represented at least once in all strains. The number of genes identified for the ST69 population in this study represents a strikingly large pan-genome for this clonal group. With less than one-third of the pan-genome corresponding to core genes, this would indicate a large accessory or variable genome and thus highly diversity among the clonal group. Running the Roary query_pan_genome script on the pan-genome matrix had identified 2,957 genes unique to non-human strains, 2,194 genes unique to human-clinical strains, and 2,223 accessory (non-core) genes shared between ST69 strains of both populations. The independent probability expectation for shared genes was calculated to test the significance of these numbers. The probability of an accessory gene being identified from the non-human population is 70.2%, whilst the probability of a gene being identified from the human-clinical population is 59.9%. Assuming that these two probabilities are independent of one another, the expected probability of encountering an accessory gene

identified in both populations is 42%. However, the observed frequency of accessory genes shared between the human-clinical and non-human populations of ST69 is 30.1%. It can therefore be inferred that the observed frequency of shared genes equates to approximately 71.7% of the expected frequency, and the probability of encountering a gene that is present in both populations is 11.9% lower than the expected probability. The pan-genome presence/absence matrix (Fig. 5.14) confirms the findings of the accessory genome analysis (section 5.3.6.2) that outside of the core genome, there are noticeable differences in the gene profiles that characterise human-clinical and non-human strains of ST69. Most striking are strains M3-27, I2-18, and I2-20 isolated from non-human sources, which possess more accessory genes than the human-clinical ST69 strains. Interestingly, gene clusters which are circled in green in Figure 5.14 are unique to strains I2-18 and I2-20, and additionally, gene clusters circled in orange are unique to strain M3-27, which further demonstrates the disparity between these strains and the rest of the ST69 population. The fact that these non-human *E. coli* strains possess more accessory genes, as well as many unique genes, would suggest a high frequency of acquisition of advantageous genes, possibly from other environmental/non-human strains of the *Escherichia* genus, or other species of Enterobacteriaceae, which may prove to be beneficial for the bacteria to survive and proliferate in different environmental conditions.

The difference in gene profiles observed between human-clinical and non-human ST69 strains may also suggest low levels of homologous recombination, due to ecological barriers between the two populations. To investigate this notion, regions of genomic recombination were detected by running the Gubbins algorithm (Croucher *et al*., 2015) on a core genome alignment of all ST69 genomes (length = 4,069,033 bp; 14 genomes). Any regions of recombination detected were visualised as coloured blocks against the phylogenetic tree constructed by Gubbins (Fig. 5.15). Blue blocks indicate recombination events that have occurred in a single isolate, while red blocks indicate signatures of recombination shared between multiple isolates. The horizontal position of the blocks represents their position in the alignment. It is evident from this analysis that recombination events unique to individual ST69 isolates are more prevalent than recombination events between multiple isolates across the alignment. Recombination between ST69 isolates of the human-clinical and non-human populations is rare, as indicated by the red blocks (Fig. 5.15). A cluster of human-clinical isolates (U67, B33, B31, and 2286_EC) which exhibit the lowest levels of recombination, shared only a single recombination event with non-human isolates (I2-18, T3-14, and T3-3), suggesting that these isolates do not inhabit the same place at the same time. Another example of limited gene flow between non-human and human-clinical strains would include I2-20 and I2-18, isolated from the same retail chicken sample, which do not recombine readily with other ST69 isolates in the strain set. Given the

suggestion that human-clinical ST69 may have non-human origins, due to reports of ST69 being isolated from pork, chicken, and beef (Vincent *et al*., 2010; Jakobsen *et al*., 2010), it would be reasonable to expect that opportunities for recombination to occur between non-human ST69 and human-clinical ST69 isolates would be more frequent. However, from the evidence of the current study, it can be inferred that only negligible levels of recombination occur between human-clinical and non-human *E. coli* ST69 strains, which would indicate limited gene flow between strains of the two populations. This suggests that chance encounters between ST69 strains of human and non-human origins in the same habitat do not occur frequently, indicating that ecological barriers may play a role in limiting the amount of gene sharing that takes place between ST69 strains of the two populations.

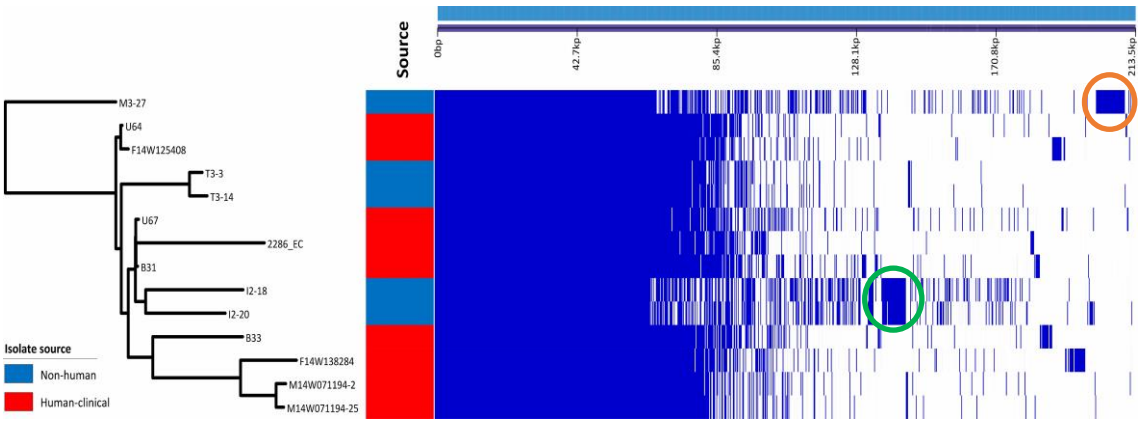


Figure 5.14. Distribution of gene profiles for ST69 *E. coli* strains isolated from human-clinical and non-human samples in Nottingham. The heat map shows the presence or absence of every genetic locus from the *E. coli* ST69 pan-genome of 14 sequenced strains (human-clinical, n = 9; non-human, n = 5). The genomes are arranged on the y-axis according to the maximum-likelihood phylogenetic tree shown on the left, with individual loci on the x-axis. Blue indicates presence of the locus in a strain and white indicates absence. Sources of isolation are annotated on the tree as coloured bars. The pan-genome matrix appears to be totally concordant with the phylogenetic tree, where similar strains have more similar gene profiles. It is noticeable that non-human strains generally share fewer accessory genes with human-clinical strains.

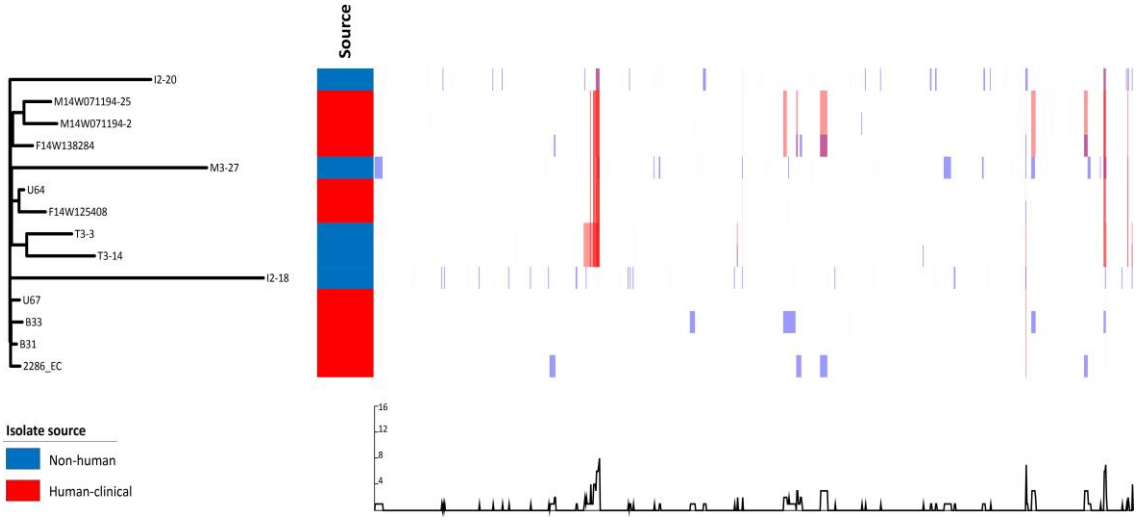


Figure 5.15. Distribution of core genome recombination events for ST69 *E. coli* strains isolated from human-clinical and non-human samples in Nottingham. Recombination events were predicted by running the Gubbins algorithm on a core genome alignment of all 14 human-clinical and non-human ST69 genomes (length = 4,069,033 bp). The phylogeny of ST69, as constructed by Gubbins, is shown on the left with source of isolation annotated on the tree. For each isolate, blocks representing the regions identified as recombination events by Gubbins are indicated by coloured blocks; blue blocks are unique to a single isolate while red blocks are shared by multiple isolates. The horizontal position of the blocks represents their position in the alignment. The number of genomes for which each recombination event is detected is indicated on the graph at the bottom.

5.3.7.2. *E. coli* ST10

The pan-genomes of all 27 ST10 strains, identified among the Nottingham human-clinical and non-human *E. coli* populations analysed in this study, were constructed. The resulting gene presence/absence matrix was annotated on the core genome phylogenetic tree of ST10 with isolate source data (Fig. 5.16). It was found that the ST10 clonal group is composed of 19,642 genes, of which, 882 were core genes represented at least once in all strains. Quite striking is the size of the pan-genome of this clonal group, which is noticeably larger than the ST69 pan-genome (10,673 genes). While only a small proportion (4.5%) of the ST10 pan-genome corresponds to core genes, the vast majority would represent accessory genes, some of which may play a role in adaptation to special growth conditions, such as those involved in the colonisation of new ecological niches. The large gene repertoire of *E. coli* ST10 would suggest that these strains are widespread in nature and have greater access to the global microbial gene pool. Roary query_pan_genome analysis identified 13,092 genes unique to non-human strains, while only 978 genes were unique to human-clinical strains, and 4,690 accessory (non-core) genes were shared between ST10 strains of both populations. The independent probability expectation for shared genes was calculated to test the significance of these numbers. The probability of an accessory gene being identified from the non-human population is very high at 94.8%, whilst the probability of a gene being identified from the human-clinical population is only 30.2%. If these two probabilities are independent of one another, the expected probability of encountering a gene identified in both populations is 28.6%. However, the observed frequency of genes shared between the human-clinical and non-human populations of ST10 is 25%. Therefore, it can be interpreted that the observed frequency of shared genes equates to only 87.4% of the expected frequency, and the probability of encountering a gene that is present in both populations is 3.6% lower than the expected probability. The pan-genome presence/absence matrix (Fig. 5.16) reveals noticeable patterns in the gene profiles for human-clinical and non-human ST10 strains, which corroborate the findings for ST69 (section 5.3.7.1) that genetic differences exist between human-clinical and non-human strains, perhaps due to niche adaptation. This pattern is even more prominent for ST10 where the human-clinical strains F14W131166-20, B20, B9, U19, F14W091968, F14W127020-13, and F14W127020-20 all possess highly diminished accessory genomes in comparison to the non-human strains of the ST10 strain set. This is also demonstrated by non-human ST10 strains possessing more unique genes when compared to the human-clinical strains. This would indicate that *E. coli* ST10 is a genetically diverse clonal group, and the larger gene pool of non-human *E. coli* ST10 would suggest that these strains typically comprise more genes encoding niche-specific fitness factors, whereas

human-clinical ExPEC strains may have undergone several gene loss events as a trade-off for enhanced fitness and survival in the bladder and bloodstream.

Recombination events were also determined to assess the amount of gene flow that occurs between human-clinical and non-human ST10 strains. The Gubbins algorithm was run on a core genome alignment of all ST10 genomes (length = 2,049,758 bp; 23 genomes), and any regions of recombination detected are shown in Figure 5.17. Overall, significantly more recombination events were detected between *E. coli* ST10 strains when compared to ST69, which would be expected considering the genetic diversity of the ST10 clonal group. Most of the recombination events shared between multiple isolates were detected between non-human strains, particularly in two clusters. On the other hand, recombination events unique to the human-clinical strains appears at much lower levels, and the overall level of detected recombination is lower in ST69 and ST10 strains of human-clinical origin. This was confirmed by quantifying the number of bases in the detected recombination events for each strain (Table 5.4). Interestingly, a cluster of non-human strains (AFR-12, AFR-6, AFR-22, M2-3, and AFR-4) do not recombine at all with the rest of the population, and only a limited amount of recombination occurs between these strains, in addition to these strains being more distantly related to the other ST10 strains, which may suggest that they form part of a subclade of the ST10 clonal complex. Furthermore, the number of SNPs falling within predicted recombination events were quantified (Table 5.4), and for the majority of non-human strains, these were generally lower than the number of base substitutions falling outside of predicted recombination events (i.e. those arising by point mutation), which would perhaps suggest lowered recombination levels between these strains and the rest of the population. Considering the high levels of recombination detected within the non-human *E. coli* population of ST10 strains, the amount of recombination observed between non-human and human-clinical strains would appear to be reduced. This is most noticeable for human-clinical strains F14W091968, F14W127020-13, and F14W127020-20, which exhibited much lower levels of shared recombination events with the rest of the strain set (Fig. 5.17; Table 5.4). Given the common reporting of *E. coli* ST10 in both community and hospital settings, as well as in non-human sources such as food animals and the environment, it would be reasonable to expect that human-clinical and non-human strains of this ST may come into close contact by chance occurrence, and therefore opportunities for recombination would be more frequent.

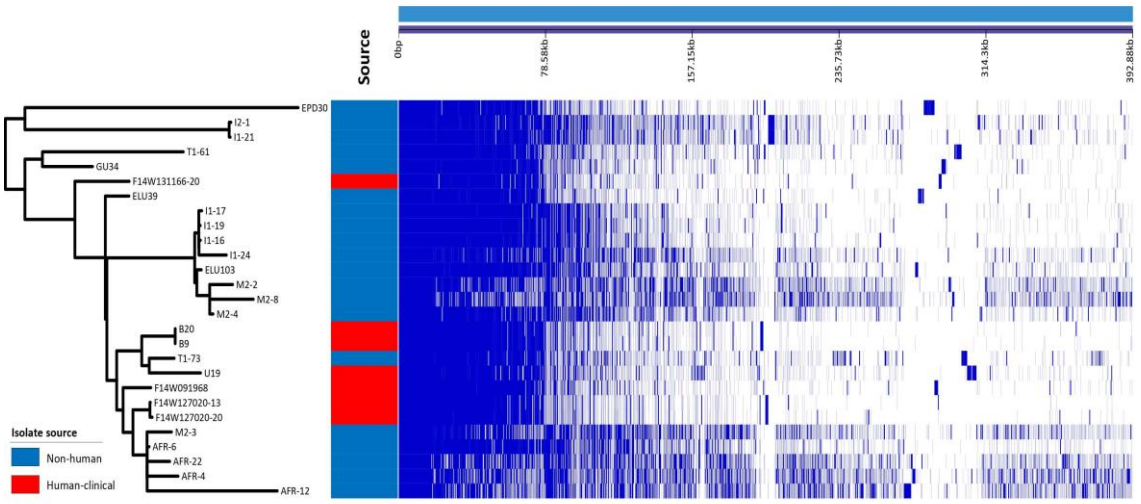


Figure 5.16. Distribution of gene profiles for ST10 *E. coli* strains isolated from human-clinical and non-human samples in Nottingham. The heat map shows the presence (blue) or absence (white) of every genetic locus from the pan-genome of 23 *E. coli* ST10 strains (human-clinical, n = 7; non-human, n = 16). The genomes are arranged on the y-axis according to the phylogenetic tree shown on the left, with individual loci on the x-axis. Sources of isolation are annotated on the tree as coloured bars. Human-clinical strains share noticeably fewer accessory genes with the non-human strains of the ST10 strain set.

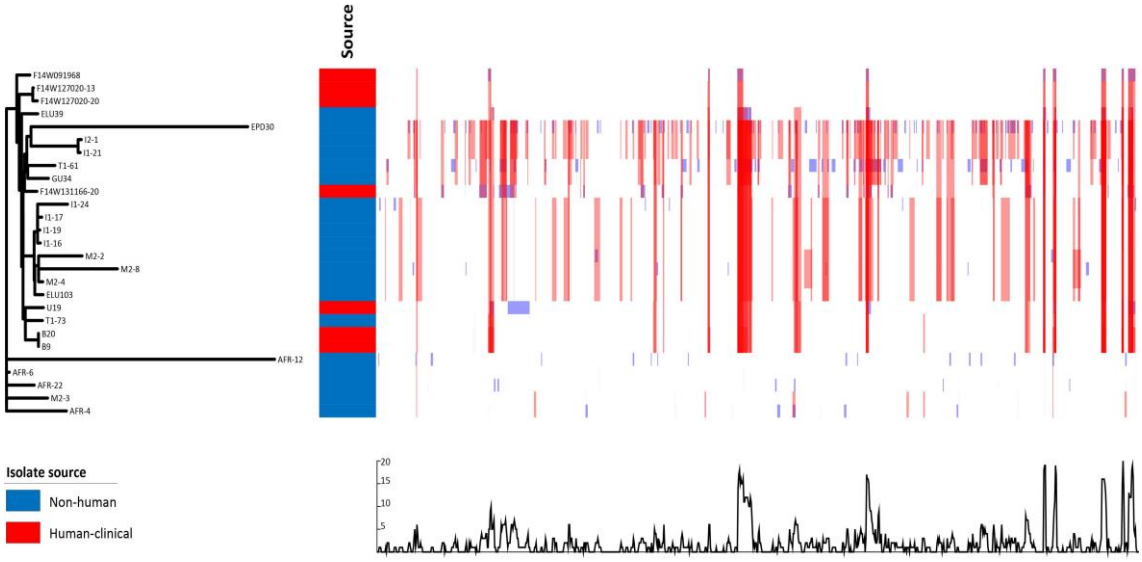


Figure 5.17. Distribution of genomic recombination events for ST10 *E. coli* strains isolated from human-clinical and non-human samples in Nottingham. Recombination events were predicted by running Gubbins on a core genome alignment of all 23 ST10 genomes (length = 2,049,758 bp). The phylogeny constructed by Gubbins is shown on the left, with isolate source annotated on the tree. Regions of recombination are indicated by horizontal coloured blocks for each genome. Blue blocks are unique to a single isolate, while red blocks are shared by multiple isolates. The number of genomes for which recombination events are detected is indicated on the graph at the bottom.

Table 5.4. Length of genome of recombinant origin within ST10 *E. coli* strains isolated from human-clinical and non-human samples in Nottingham.

Node	Bases identified within recombination events (bp)	No. of recombination blocks	No. of SNPs inside recombinations (out of total SNPs)
F14W091968	94,307	9	598 (814)
F14W127020-13	94,795	0	0 (0)
F14W127020-20	94,795	0	0 (75)
ELU39	149,857	9	427 (677)
EPD30	897,655	119	5,942 (10,316)
I2-1	813,606	0	0 (81)
I1-21	813,606	0	0 (0)
T1-61	682,182	46	4,227 (4,806)
GU34	524,200	0	0 (0)
F14W131166-20	403,350	25	1,385 (1,637)
I1-24	441,277	14	355 (857)
I1-17	422,806	0	0 (0)
I1-19	422,806	1	4 (46)
I1-16	422,816	1	7 (46)
M2-2	454,752	3	73 (713)
M2-8	477,053	9	204 (1,384)
M2-4	451,121	0	0 (0)
ELU103	422,797	0	0 (0)
U19	241,201	9	1,136 (1,387)
T1-73	180,500	0	0 (0)
B20	182,187	0	0 (0)
B9	182,187	0	0 (0)
AFR-12	45,881	16	514 (3,517)
AFR-6	6,264	0	0 (0)
AFR-22	24,198	11	283 (712)
M2-3	37,391	0	0 (0)
AFR-4	59,365	12	266 (1,646)

Recombination events were predicted by running Gubbins on a core genome alignment of all 23 ST10 genomes (length = 2,049,758 bp). The number of bases in recombination blocks is shown for each node, which represents the total length of all recombination events, as quantified by the Gubbins algorithm (Fig. 5.17). The total number of recombination blocks reconstructed onto each branch was also quantified, as well as the number of base substitutions reconstructed onto the branch that fall within a predicted recombination. For the majority of non-human strains, the number of base substitutions identified within recombination events is generally lower than those falling outside of predicted recombination events (i.e. those arising by point mutation), which may suggest lowered recombination levels between these strains and the rest of the population.

5.4. Conclusions

What we currently know about the ecology of *E. coli* is based on data from studies of representative isolates. A large amount of data has been collected for antimicrobial-resistant *E. coli* isolated from human-clinical cases. In chapter 4, genomic analyses of a large, unbiased population of *E. coli* isolated from non-human sources revealed a population of significant genomic diversity, with a low prevalence of specific human ExPEC and multidrug-resistant lineages. The population structure of *E. coli* responsible for human extraintestinal infection is well-described in the current literature. Human-clinical ExPEC strains have often been characterised at the sequence type level by MLST, and a small number of STs, namely ST69, ST73, ST95, and ST131, were found to predominate among cases of UTIs and bloodstream infections (Kallonen *et al*., 2017; Riley, 2014; Alhashash *et al*., 2013; Croxall *et al*., 2011b). Previous studies that have attempted to attribute transmission of ExPEC in human infections to poultry or environmental sources, have largely focussed on MDR *E. coli* and ESBL-producing strains, using traditional typing methods. These studies are not likely to provide sufficient resolution to assess the relatedness of strains isolated from non-human and human-clinical sources (Jang *et al*., 2013; Platell *et al*., 2011b; Dolejska *et al*., 2011b; Dolejska *et al*., 2011a). To address this, comparative genomic analyses of WGS data for geographically-constrained non-human and human-clinical populations of *E. coli* were carried out, to provide a higher level of resolution to determine the extent of genetic overlap that may between these two populations.

The ST-designations of a collection of 399 human-clinical strains, previously isolated from cases of bacteraemia and UTIs in Nottingham (Alhashash *et al*., 2013; Croxall *et al*., 2011b), confirmed a clonally diverse population dominated by the four major ST complexes: ST131, ST73, ST69, and ST95, and an observable presence of the ST10 complex. Comparison of the prevalence of these clinically-dominant STs between the human-clinical population and the non-human population of *E. coli*, characterised in chapter 4, further demonstrates a clear disparity in the population structures of human-clinical and non-human *E. coli*. Whilst the important MDR ExPEC lineage, ST131, was the most commonly encountered ST among human-clinical isolates analysed in this study, only one instance of this genotype was observed in the non-human population of *E. coli*. Additionally, STs 73 and 95, which were prevalent in the human-clinical population analysed in this study, could not be detected among non-human isolates of *E. coli*. ST69 was identified in both the human-clinical and non-human populations, and although this ST represented a higher proportion of human-clinical isolates (7.0%) than non-human isolates (3.1%), this difference was not considered to be statistically significant (p = 0.137, two-tailed Fisher's test). ST10 is a common human ExPEC ST that is also associated with food animals and retail poultry meat

(Manges 2016; Aslam *et al*., 2014; Bergeron *et al*., 2012; Vincent *et al*., 2010). It was revealed in this study that ST10 was the dominant lineage in the non-human population of *E. coli* (12.5%) and was significantly more prevalent than human-clinical ST10 isolates (1.5%; $p < 0.0001$, two-tailed Fisher's test). This suggests a more widespread prevalence of this ST in the environment and the food chain than in human-clinical cases of ExPEC. The paucity of human ExPEC STs 131, 73, 69, and 95 in the non-human population of *E. coli* would indicate that the non-human reservoir of human ExPEC is negligible and is unlikely to be responsible for the majority of human ExPEC infections. The distinct nature of the population structures that define human-clinical and non-human *E. coli* in Nottingham, in addition to the lack of shared STs, would suggest that the two populations do not frequently encounter each other despite existing within a geographically-constrained ecosystem.

A core genome phylogenetic tree, encompassing 136 whole-genome sequenced human-clinical strains (isolated from cases of bacteraemia, UTI, and neonatal sepsis in Nottingham) and the 128 non-human *E. coli* genomes from chapter 4, provided an insight into the phylogroup composition of the two populations. Due to the overrepresentation of ST131 and ST73 strains, the human-clinical population was dominated by phylogroup B2 (80.1%), while in contrast, only 7% of the non-human population were classified as phylogroup B2. This would demonstrate a lack of classically pathogenic strains in the non-human population, according to previous observations that phylogroup B2 strains are generally more virulent than strains belonging to the other phylogroups (Picard *et al*., 1999; Boyd and Hartl, 1998). On the contrary, the non-human population of *E. coli* is largely composed of phylogroup A and B1, indicating a high proportion of largely commensal strains of *E. coli*, as reported by previous studies (Duriez *et al*., 2001; Picard *et al*., 1999). Considering that the human-clinical *E. coli* study population consisted of isolates primarily from elderly patients and neonates with suspected urinary and bloodstream infections, the lack of human-clinical isolates from classical commensal phylogroups suggests that pathogenic *E. coli* may be more likely to opportunistically colonise patients with compromised immune systems. This is supported by the significantly higher prevalence of ExPEC strains identified in the human-clinical population when compared to the non-human population of *E. coli*, as determined by *in silico* VAG profiling (human-clinical, 66.9% vs non-human 8.6%; $p < 0.0001$, two tailed Fisher's test). Another disparity noted between the two populations is the prevalence of potentially MDR strains of *E. coli*. *In silico* antimicrobial resistance gene profiling of all strains revealed a higher prevalence of ciprofloxacin, trimethoprim, and macrolide resistance genes among human-clinical isolates when compared to non-human isolates of *E. coli*, corroborating the levels of phenotypic resistance observed for these antibiotic classes in human-clinical *E. coli* previously (Kallonen *et al*., 2017; Alhashash *et*

al., 2013; Croxall *et al*., 2011b). More importantly, however, was the appreciable difference in prevalence of ESBL genes observed between the two populations. The $bla_{CTX\text{-}M}$ family of ESBL genes, responsible for the majority of MDR infections worldwide (Nicolas-Chanoine *et al*., 2008; Lau *et al*., 2008), were identified in 21.3% of all human-clinical *E. coli* isolates, while such genes could not be detected in the non-human population. Furthermore, a significantly higher prevalence of bla_{OXA} genes were detected in human-clinical isolates (21.3%), when compared to non-human *E. coli* isolates (1.6%; $p < 0.0001$, two tailed Fisher's test). When taken together, the low prevalence of human ExPEC and MDR strains in the non-human population of *E. coli* may suggest that strains of the non-human population of *E. coli* in Nottingham do not readily encounter strains of the human-clinical population, thereby lowering the frequency of horizontal transfer of ESBL and virulence-associated genes between the two populations.

Of the limited number of clinically important sequence types identified in the non-human population of *E. coli* analysed in this study, only a single ST131 strain (GD45) and a single ST648 strain (GD49) were isolated from river water samples. GD45 was included in a core genome phylogenetic tree of a global collection of 242 *E. coli* ST131 genomes, obtained from multiple ecosystems, to determine the relatedness of the non-human ST131 strain to a wider population of the ST131 lineage. Of the three known *E. coli* ST131 subclades, GD45 was revealed to be a clade B strain, and is therefore not related to the clade C lineage of ST131, which is often associated with carrying the ESBL gene $bla_{CTX\text{-}M\text{-}15}$ and is currently responsible for the majority of extraintestinal infections in humans (Nicolas-Chanoine, Bertrand and Madec, 2014). Given the position of GD45 in clade B and its lack of a CTX-M-carrying plasmid, it would suggest that this particular non-human ST131 strain is unlikely to be implicated in human ExPEC infections, and perhaps indicates limited overlap between non-human and human-clinical populations of *E. coli*. However, to determine whether a low prevalence of clade C ST131 strains is true for the majority of non-human populations of *E. coli,* deeper genomic sampling from additional non-human and environmental sources would be required in future work. Similar phylogenetic analysis was performed for the non-human ST648 strain (GD49), which was included in a core genome phylogenetic tree of 89 European *E. coli* ST648 genomes, from multiple host sources. Despite the geographical and ecological separation between GD49 (isolated from river water in Nottingham) and the rest of the population (isolated largely from companion animals in mainland Western Europe), GD49 still falls within the phylogenetic tree, indicating its relation to the European population of ST648. This would support a widespread dispersion of this ST as previously reported globally (Guenther *et al*., 2012; Cortes *et al*., 2010), and it may also suggest that environmental ST648 strains originate from humans, companion animals, or livestock. However, of importance is the fact that GD49 lacks a CTX-M-carrying plasmid, which is contrary

to the overwhelming majority of the ST648 population (88%), which were found to harbour a CTX-M gene. Based on the association of *bla*$_{CTX\text{-}M\text{-}15}$ in ST648 isolates with human-clinical cases observed in previous studies (Zong and Yu, 2010; Cortes *et al*., 2010; Nicolas-Chanoine *et al*., 2008), it would seem unlikely that ST648 isolates lacking this ESBL type can be attributed to the global expansion of ST648 and ExPEC infections in humans. A broader sample size of ST648 isolates from environmental and foodborne sources should be studied in future work to determine whether these isolates follow a general trend of lacking the *bla*$_{CTX\text{-}M\text{-}15}$ ESBL. The Nottingham *E. coli* study population includes other instances of STs prevalent in both populations, such as ST10 (human-clinical, n = 7; non-human, n = 16) and ST69 (human-clinical, n = 9; non-human, n = 5). However, it was decided that phylogenetic comparative analysis would not be carried out for the non-human *E. coli* ST10 strains within the wider population of ST10, because aside from a small clade that are largely CTX-M-1 producers (Leverstein-van Hall *et al*., 2011), ST10 is a very diverse clonal complex associated with low virulence and antibiotic susceptibility (Manges and Johnson, 2012). Pan-genome analysis would therefore provide the appropriate method to distinguish between non-human and human-clinical strains. Regarding ST69, there was simply not a large enough genome collection from multiple sources to carry out a substantial phylogenetic reconstruction of the wider ST69 population.

A pan-genome approach was taken in this study, as it offers a much higher resolution than core genome phylogenetic and MLST analyses, because it takes into consideration the entire bacterial genome (Hall, Ehrlich and Hu, 2010). The pan-genome analysis of all 264 human-clinical and non-human *E. coli* strains of the Nottingham study population revealed a highly heterogeneous genomic data set, where only 596 of the total 69,645 genes were represented at least once in ≥ 95% of strains. The analysis also indicated a vast number of strain-specific genes (29,139) present in the pan-genome, suggesting a large accessory genome which encodes for the strain diversity and selective advantages for virulence, antibiotic resistance and niche adaptation (Tettelin *et al*., 2008). This prompted further analysis of the accessory genomes to determine proportions of accessory gene clusters or loci that are unique to either the human-clinical or non-human population, as well as the proportion of genes shared between both populations. It was found that 12,311 genes and 31,468 genes were unique to the human-clinical and non-human populations, respectively, however the majority of these genes corresponded to strain-specific genes. It was therefore the number of genes shared between the two populations (n = 21,056) which was of interest for comparing the two populations. Statistical analysis revealed that the probability of encountering a gene that is present in both the human-clinical and non-human populations of Nottingham is 9.2% lower than expected. This would perhaps be indicative of a restriction in the amount of gene sharing that takes place between the human-

clinical and non-human populations of *E. coli*, presumably because of the ecological barriers and difference in environmental selection pressures that exist between the two populations.

Pan-genome analysis and detection of core genome recombination events were also carried out specifically for strains of ST69 and ST10, as multiple representative strains were identified for these STs in both the human-clinical and non-human populations. ST10 is a diverse lineage and several recent studies have reported an association of ST10 isolates with human disease, ESBL carriage, and livestock (Bergeron *et al*., 2012; Peirano *et al*., 2012; Leverstein-van Hall *et al*., 2011; Cortes *et al*., 2010). ST69 is one of the more recognised clonal groups implicated in human ExPEC infection worldwide (Kallonen *et al*., 2017; Dias *et al*., 2009; Johnson *et al*., 2009). Identification of these clinically relevant STs in both the human-clinical and non-human populations of *E. coli* therefore warranted further genomic comparisons to assess the level of genetic flow between closely related strains of the two populations. Noticeable differences in the gene profiles of human-clinical strains compared to non-human strains were observed for both the ST69 and ST10 strain sets, though these were more prominent for ST10. The non-human strains shared few accessory genes with the human-clinical strains, and the non-human strains possessed more accessory genes, perhaps due to these strains requiring more genes encoding niche-specific fitness factors for adaptation and survival in the environment. Lowered levels of recombination events were detected between human-clinical and non-human strains of ST69, which is concordant with the divergent gene profiles of human-clinical and non-human strains. This observation is consistent with previous studies reporting reduced genetic flow between environmental and enteric *E. coli* (Luo *et al*., 2011), and additionally between poultry and human-clinical isolates of *E. coli* (de Been *et al*., 2014), suggesting that human-clinical and non-human *E. coli* form two distinct populations. To further investigate the level of genetic overlap between human-clinical and non-human strains, additional sequence type-specific pan-genome and recombination analyses should be carried out in future work.

Overall, the results of this investigation revealed that there are clear differences in the population structures of *E. coli* isolated from human-clinical and non-human sources, with regards to the major STs that comprise these populations. At the phylogenetic level, a small number of clinically important, clonally-related strains such as ST131, ST648, ST69, and ST10 were identified in both populations. However, comparative analyses at the whole genome level revealed a low prevalence of potential MDR and ExPEC strains in the non-human population compared to the human-clinical population. Furthermore, gene sharing between *E. coli* ST69 strains of the human-clinical and non-human populations is limited, suggesting that ecological barriers may contribute to reduced levels of recombination between the two populations. Further lineage-specific specific analyses, encompassing additional non-human sources of *E. coli*

would be required to confirm this. Collectively, these results lead to the conclusion that an obvious non-human reservoir of human MDR *E. coli* and ExPEC strains does not exist, and therefore, it is unlikely that the non-human population of *E. coli* contributes significantly to the burden of hospital- and community-acquired extraintestinal infections.

CHAPTER 6

Conclusions and future directions

6. Conclusions and future directions

Accessing information on the full ecology of bacterial pathogens has previously been a deficiency of many microbial ecological studies, and consequently, has hindered our understanding of the evolution and dissemination of these organisms, which cause many debilitating and life-threatening diseases in humans. Before the advent of whole-genome sequence-based analysis, the study of microbial ecology was largely restricted to low-resolution and subjective molecular and biochemical techniques. To effectively probe the structures of microbial populations and fully appreciate species diversity, sequencing of single isolates must be employed to achieve the genomic resolution required (Quince *et al*., 2017). Using this approach, the current study aimed to elucidate the unknown ecology of two important pathogens to humans, *Y. pseudotuberculosis* and *E. coli*. Previous genome-scale analyses revealed hidden ecological inferences for the enteric pathogen *Y. enterocolitica* (Reuter *et al*., 2015; Reuter *et al*., 2014), which warranted further investigation to determine whether similar hidden patterns may exist in *Y. pseudotuberculosis*.

Phylogenetic analysis of globally and temporally distributed *Y. pseudotuberculosis* genomes from multiple ecosystems identified a clear phylogeographic split in the population, which was previously undetected for pathogenic *Yersinia* species. This was characterised by an ancestral clade of strains of primarily Asian origin and a second low diversity clade of mainly European strains. This phenomenon was once postulated for *Y. enterocolitica* with Old World and New World strains (Wang *et al*., 2011), however, recent population genomics studies by our group have shown this is not the case (Reuter *et al*., 2015; Reuter *et al*., 2014). The Asian ancestry of *Y. pseudotuberculosis* is consistent with the estimated ancestry of *Y. pestis* (Morelli *et al*., 2010; Achtman *et al*., 1999). Phylogenetic dating provided an estimated TMRCA for the entire *Y. pseudotuberculosis* data set, which is in the same range (10,000–40,000 years before present) as that estimated for the emergence of *Y. pestis* (Achtman *et al*., 1999). Based on this evidence, we could argue that the emergence of *Y. pestis* coincided with a larger population dispersal event across *Y. pseudotuberculosis*. Additionally, the dating analysis indicated a recent geographical divergence of the European and Asian clades occurring ~12,500 years before present. It would be tempting to suggest that a bottleneck event occurring in the recent past may have resulted in the establishment of a small number of successful clones in new ecosystems, leading to subsequent dissemination of these clonal lineages into Europe and the rest of the world. Mapping serotype designations onto the *Y. pseudotuberculosis* phylogeny indicated that successfully established clones in Europe belong almost exclusively to serotypes

O:1a and O:1b, consistent with previous studies reporting a predominance of these serotypes in Europe, whilst greater serotype diversity exists in Asia (Fukushima *et al*., 2001).

One of the limitations of the phylogenetic dating analysis was the availability of only a relatively small number of isolation dates (n = 46) representing the full diversity of the phylogeny. This made it difficult to accurately date the phylogeny of the entire data set. A greater accuracy of dating would have been achieved had there been more isolation dates available. Furthermore, it would be suggested that a more thorough and dense genomic sampling should be considered for further analysis of *Y. pseudotuberculosis*. This would reveal whether the levels of diversity observed in geographic clades in this study are maintained in the broader population of *Y. pseudotuberculosis*. Significantly more isolates from Africa and the Americas should be included in further analyses, as these isolates were noticeably underrepresented in the population analysed in this study. This should provide stronger evidence to support a separate migration of the species into Africa and the Americas, as evidenced by the small transitional cluster of strains between the Asian and European clades.

Further genomic analysis of the *Y. pseudotuberculosis* population identified clearly distinct phylogenetic subgroups within each clade. These phylogroups have unique combinations of accessory genes, with little variation in their accessory genomes, and they share very similar patterns of core genome homologous recombination. Similar phylogroup signatures were identified from analysis of accessory gene sharing between serotype-specific clades of *Y. enterocolitica*, in a previous study by our group (Reuter *et al*., 2015). It was concluded that the restricted gene sharing between clades was due to ecological separation. In *Y. pseudotuberculosis*, however, the observation of phylogenetic subgroups, which are intimately connected to the CRISPR spacer designations, presents a strong case for the role of the CRISPR system in the formation of these phylogroups. The inference of CRISPR playing a role in mediating this genetic restriction is supported by the observation that different CRISPR cluster-type strains coexist in the same geographical habitat, and several clusters appear in samples separated by at least a decade. Moreover, it would be expected that the CRISPR cluster signature would be gradually eroded over time (Dearlove *et al*., 2016), especially considering that *Y. pseudotuberculosis* is widespread in nature, and no active barrier would prevent recombination between strains coexisting in the same ecological niche. As a result, clear phylogroup signatures should not be observed within the population (Sheppard *et al*., 2008), given the high levels of core genome recombination detected for *Y. pseudotuberculosis*. However, it can be concluded from the data that the CRISPR system is strongly associated with restriction of both accessory and core gene exchange between tightly maintained *Y. pseudotuberculosis* phylogroups. To our knowledge, this study provides the first evidence of a possible causative link between CRISPRs

and the evolution of distinct phylogroups in a bacterial species. The data suggest that large population perturbations led to the emergence of geographically isolated clones. These clones encounter geographically localised exogenous DNA, creating a CRISPR array of immunity which controls the range of genetic material that can be exchanged with other clones of *Y. pseudotuberculosis*. Clones comprising the same CRISPR array can freely exchange genes without any restriction; however, horizontal gene transfer between different clones is not sufficient to erode the clonal phylogenetic structure, and thus distinct phylogroups of *Y. pseudotuberculosis* persist within the population.

Whilst a strong correlation between CRISPRs and restriction of genetic flow exists in the *Y. pseudotuberculosis* population, this study did not reveal any phylogenetic signature associated with the ecology of the species. Isolates from different hosts/ecosystems were distributed throughout the phylogeny and did not cluster separately from each other. This is indicative of the ability of *Y. pseudotuberculosis* to colonise many different non-human hosts and environments, suggesting a broad and widespread ecology for the species. Although several previous studies have reported a role for ecological barriers, in shaping distinct ecotypes within populations of important bacterial pathogens (Reuter *et al*., 2015; Sheppard *et al*., 2014; McNally *et al*., 2013; Luo *et al*., 2011), the present study has demonstrated that *Y. pseudotuberculosis* is a host generalist species with an ability to occupy multiple ecological niches, from where pathogenic lineages can easily be transferred to humans. Our findings indicate that evolution of *Y. pseudotuberculosis* is driven by factors other than those that prevent physical contact. The study therefore creates a new window of research for microbial ecology and evolution where CRISPR can be used to investigate how distinct ecotypes may emerge for important human pathogens.

Comprehensive population genomic analyses were also performed on *E. coli*, another model organism used to investigate microbial ecology in this study. Multiple reports exist of *E. coli*, isolated from environmental waters and retail chicken meat, that resemble strains responsible for human extraintestinal infections. This has led to the suggestion that there may be a non-human reservoir for human multidrug-resistant ExPEC. Many of these environmental microbiological studies tend to bias towards resistant isolates in their sampling procedure, thus leading to an overrepresentation of MDR – specifically ESBL-producing – *E. coli* strains in the literature (Hussain *et al*., 2017; Manges, 2016; Lazarus *et al*., 2015; de Been *et al*., 2014). Given the shortcoming of past studies, it was imperative to determine the population structure of non-human *E. coli* from river water and retail chicken samples in this study. This was achieved using an unbiased culture-based approach, to sample all isolates in order to achieve a more accurate snapshot of the prevalence of MDR and pathogenic *E. coli*. An *E. coli* population, consisting of

128 sequenced genomes, was obtained from river water and retail chicken samples in Nottingham and subjected to whole-genome analysis. To put this population in the context of *E. coli* causing human extraintestinal infection, comparative population genomics was performed with previously isolated human-clinical *E. coli* strains from urinary and bloodstream infection cases in Nottingham, allowing for high-resolution examination of any genetic overlap between the two populations.

In silico MLST analysis revealed that the non-human *E. coli* population is not dominated by the STs that are frequently associated with urinary tract and bloodstream infections; ST131, ST73, ST69, and ST95. Rather, the non-human population of *E. coli* is comprised of a wide variety of different STs, with most strains not grouping into larger clonal complexes, demonstrating the genetic diversity within the population. The ST10 clonal complex emerged as the most frequently encountered clonal group, representing approximately 15% of the non-human *E. coli* population analysed in this study. These findings represent a clear disparity in the population structures of non-human and human-clinical *E. coli*. The lack of STs associated with human extraintestinal infection in the non-human population would suggest that the environmental and foodborne risk for human extraintestinal infection, specifically through contaminated water and retail chicken meat, is very low. The limited number of clinically associated STs identified may also be indicative of a limitation in the non-human sampling. Isolates were only obtained from river water and retail chicken samples in this study. To access the full diversity of *E. coli* in the environment and food chain, future work should consider a much deeper sampling of non-human sources, such as slurry, plants, wild and domesticated animals, other retail meats, and dairy products.

Phylogenetic analysis of the non-human and human-clinical populations of *E. coli* provided an insight into the phylogroup composition of the two populations. An overrepresentation of human-clinical *E. coli* genomes belonging to ST131 and ST73 meant that this population was dominated by phylogroup B2 strains (~80%), while by contrast, only 7% of the non-human population were classified as phylogroup B2. This demonstrated a lack of classically pathogenic strains in the non-human population, according to previous reports that phylogroup B2 strains are generally more virulent than strains belonging to other phylogroups (Picard *et al*., 1999; Boyd and Hartl, 1998). On the other hand, the non-human population of *E. coli* was largely composed of phylogroup A and B1 strains, indicating a predominance of largely commensal and non-pathogenic strains of *E. coli* (Duriez *et al*., 2001; Picard *et al*., 1999). To confirm the distribution of pathogenic strains between the human-clinical and non-human populations of *E. coli*, *in silico* VAG profiling was performed for all strains to identify the virulence factors associated with ExPEC. A significantly higher prevalence of ExPEC strains in the human-clinical

population (66.9%) was revealed when compared to the non-human population of *E. coli* (8.6%). Given that the human-clinical *E. coli* population analysed in this study was largely derived from elderly patients and neonates diagnosed with urinary and bloodstream infections, the high prevalence of ExPEC would indicate that pathogenic *E. coli* strains are more likely to cause opportunistic infection in immunocompromised hosts. Therefore, the non-human population, with a very low prevalence of ExPEC, presents little or no significance to cases of human extraintestinal infection in Nottingham. It must be noted that the definition of ExPEC used in this study is based on the presence of five virulence markers, but that the actual pathogenicity of strains cannot be determined just by the presence/absence of these five markers. Furthermore, the human-clinical genomic data set analysed in this study does not reflect an accurate representation of the prevalence of pathogenic strains in the population. The genomic data set consists of sequenced *E. coli* genomes from human-clinical samples, from Nottingham, that were available at the time of this study; however, the population is largely overrepresented by ST131 and ST73 strains, which were selectively sequenced as part of previous projects by our group (Alhashash *et al*., 2016; Clark *et al*., 2012). To accurately compare the phylogroup structure and prevalence of ExPEC in non-human and human-clinical populations of *E. coli* in future work, a similar approach should be taken with human-clinical samples, where isolates are sequenced to represent the full spectrum of diversity in the population. Another important difference noted between the two populations would be the prevalence of potential MDR strains of *E. coli*. The $bla_{\text{CTX-M}}$ family of ESBL genes, which is associated with the majority of MDR infections worldwide (Nicolas-Chanoine *et al*., 2008; Lau *et al*., 2008), was identified in 21.3% of all human-clinical *E. coli* isolates, while these genes were not present among non-human *E. coli* strains. The low incidence of ESBL gene carriers and ExPEC strains, in the non-human population of *E. coli*, reflects a genuine representation of the prevalence of potential MDR and/or pathogenic strains, in the natural environment and in food products. This contrasts with previous studies that have selectively enriched for antimicrobial-resistant isolates and have reported an increasing occurrence of ESBL producers (Hussain *et al*., 2017; Manges, 2016; Lazarus *et al*., 2015; de Been *et al*., 2014). The lack of phenotypic characterisation of the non-human *E. coli* isolates in this study is noted. Future work may consider performing antimicrobial susceptibility testing on these isolates, as well as assays for physiological virulence factors such as adhesins, capsule production, flagella, and toxins, to assess the virulence potential of these isolates *in vitro*.

Given the paucity of *E. coli* sequence types associated with human extraintestinal infection in the non-human population (namely ST131, ST73, ST95, and ST648), producing pan-genomes specific to these STs was not feasible. Therefore, a pan-genome of all 264 human-clinical and non-human *E. coli* strains of the Nottingham study population was constructed, which revealed

a highly heterogeneous genomic data set sharing only a small proportion of core genes; a result of the inclusion of distantly related strains representing the full spectrum of phylogenetic diversity across the species (Rasko *et al*., 2008). The analysis also revealed a vast number of strain-specific genes present in the pan-genome, indicating the high level of genome plasticity in the *E. coli* accessory genome, which encodes for the strain diversity and selective advantages for virulence, antibiotic resistance, and niche adaptation (Tettelin *et al*., 2008). This led to subsequent investigation of the accessory genome, to determine the number of accessory genes that are unique to either the human-clinical or non-human population, as well as the number of genes shared between both populations. This uncovered a noticeable difference in the number of accessory genes unique to either population, though these numbers largely represented strain-specific genes. Closer scrutiny of the number of genes shared between the two populations revealed that the probability of encountering a gene that is present in both the human-clinical and non-human populations is 9.2% lower than expected. This indicates that there could be ecological barriers limiting the amount of gene sharing that takes place between the two populations. A restriction in genetic flow may also indicate limited movement of MDR plasmids between strains of the two populations. Given the very low prevalence of ESBL-carrying plasmids in the non-human population, it was decided that comparative analyses of plasmid DNA with human-clinical strains would not be carried out in this study, and thus, the pan-genome approach was employed to provide an indication of gene flow. The limitation of constructing a pan-genome of both combined populations is that the reported number of genes unique to either population may be misleading, due to the overrepresentation of strain-specific genes. Constructing pan-genomes specific to each of the dominant *E. coli* STs responsible for human extraintestinal infection would perhaps give a clearer indication of movement of mobile genetic elements between the two populations. Genome sequencing of additional non-human *E. coli* ST131, ST73, ST95, and ST648 strains, from the same geographical region, would therefore be required for further genomic comparison with human-clinical strains of the same STs.

Only single representative isolates of the clinically important ST131 (GU45) and ST648 (GD49) strains were detected in the non-human population of *E. coli*. SNP-based core genome phylogenetic trees were constructed to situate the ST131 and ST648 strains within wider populations of their respective STs. The ST131 isolate obtained from this study was situated within clade B of the ST131 phylogeny, comprising globally dispersed ST131 genomes from multiple ecosystems. The clade C lineage of ST131 is most often associated with carriage of the ESBL gene $bla_{\text{CTX-M-15}}$ and is currently recognised as the major cause of human extraintestinal infections (Nicolas-Chanoine, Bertrand and Madec, 2014). The absence of clade C, CTX-M-15-producing *E. coli* ST131 strains in the non-human population analysed in this study would

suggest that ST131 in the environment poses very little concern to human health. Subclones of ST131 clade C have recently been identified among ESBL-producing *E. coli*, in wastewater from treatment plants and hospitals in Japan (Gomi *et al*., 2017a), suggesting dissemination of clinically important lineages into the environment. However, to determine the prevalence of clade C ST131 clones within the context of the non-human *E. coli* population of Nottingham, a much more thorough and dense genomic sampling of isolates from additional non-human sources in this region, would be required in future genomic analyses. The single non-human *E. coli* ST648 strain recovered from this study was situated phylogenetically within a European population of ST648, isolated mainly from companion animals. Despite the phylogenetic relatedness of GD49 to the rest of the population, this strain did not harbour the $bla_{\text{CTX-M-15}}$ ESBL gene, whilst the vast majority (88%) of the ST648 population was associated with $bla_{\text{CTX-M}}$ carriage. Given the worldwide prevalence of ST648 strains possessing the $bla_{\text{CTX-M-15}}$ ESBL gene, and the implication of these strains in human-clinical cases (Zong and Yu 2010; Cortes *et al*., 2010; Nicolas-Chanoine *et al*., 2008), it would be reasonable to suggest that ST648 strains lacking a CTX-M-carrying plasmid, from non-human sources, are unlikely to contribute to the global expansion of ST648 in human extraintestinal infections. A broader collection of non-human ST648 isolates would be required in future genomic studies, to determine whether ST648 from the wider non-human/environmental population in Nottingham generally lack the $bla_{\text{CTX-M-15}}$ ESBL gene.

Pan-genome analysis was also performed specifically for ST69 and ST10, owing to multiple representative strains of these STs in both the human-clinical and non-human populations of *E. coli*. The non-human strains possessed more unique genes than those shared with human-clinical strains, suggesting a greater repertoire of genetic material acquired from the gene pool, such as those encoding niche-specific fitness factors for adaptation and survival in the environment. Recombinant regions within the core genomes of ST69 and ST10 strains were detected to assess the extent of genetic exchange between non-human and human-clinical *E. coli*. Negligible levels of recombination were detected between human-clinical and non-human ST69 strains, suggesting limited opportunity for genetic exchange, likely due to these strains not coexisting in the same space at the same time. Significantly more recombination events were shared between human-clinical and non-human ST10 strains, however, indicating genetic exchange between these strains. Given the widespread, diverse nature and frequent isolation of *E. coli* ST10 in community, hospital, and environmental locales, it would be expected that opportunities for recombination would be more frequent due to chance encounters. While strains of the same ST from human-clinical and non-human sources are phylogenetically related, there is a clear difference in gene content between non-human and human-clinical strains. This

would suggest that these strains once shared a recent common source, but have since diversified and become more adapted to the surrounding environment they presently inhabit. Further sequence type-specific investigations into the pan-genomes and core genome recombination events between non-human and human-clinical strains are required, to confirm the level of genetic overlap between clinically associated lineages of the two populations.

In summary, the work detailed in this thesis was achieved through whole-genome sequence analyses, providing a high-resolution view of the population structures, ecology, and evolution of *Y. pseudotuberculosis* and *E. coli*. Through in-depth phylogenetic analysis, this study yielded definitive evidence of a phylogeographic split of a globally dispersed population of *Y. pseudotuberculosis* with an Asian ancestry, and subsequent dissemination of successful clonal lineages into Europe and the rest of the world. This study provides the first evidence of a possible causative link between CRISPR cassettes and the evolution of these distinct lineages in *Y. pseudotuberculosis*. Bayesian analysis of core genome recombination revealed that restriction of genetic exchange maintains the discrete lineage structure in the population, despite coexistence of lineages for thousands of years. The study did not identify a role for ecological barriers in defining the distinct lineage structure of this species, suggesting that *Y. pseudotuberculosis* is a host generalist pathogen with an ability to persist in multiple niches. The use of CRISPR in providing evolutionary insights into the emergence of bacterial lineages warrants further investigation in other important human pathogens. Comparative population genomics of *E. coli* uncovered clear differences in the population structures of *E. coli* isolated from human-clinical and non-human sources. A snapshot revealing a low prevalence of clinically-associated sequence types, multidrug resistance, and potential ExPEC strains among non-human *E. coli* when compared to human-clinical *E. coli* would suggest two distinct populations within the region. This was supported by a clear difference in accessory genome content between the two populations and only minimal levels of genetic exchange between closely related strains, such as ST69, suggesting ecological barriers to recombination may play a role in driving evolution within *E. coli*. The evidence gathered from this study led to the conclusion that the non-human population of *E. coli* is unlikely to contribute significantly to the burden of hospital- and community-acquired extraintestinal infections in humans.

References

Achard, A., Villers, C., Pichereau, V. and Leclercq, R., 2005. New lnu(C) Gene Conferring Resistance to Lincomycin by Nucleotidylation in *Streptococcus agalactiae* UCN36. *Antimicrobial Agents and Chemotherapy,* 49 (7), 2716-2719.

Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A. and Carniel, E., 1999. *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proceedings of the National Academy of Sciences of the USA,* 96 (24), 14043-8.

Achtman, M., 2017. Multiple time scales for dispersals of bacterial disease over human history. *In:* M. Petraglia, N. Boivin and R. Crassard, eds., *Human Dispersal and Species Movement: From Prehistory to the Present.* Cambridge: Cambridge University Press, 2017, pp. 454-476.

Adams-Sapper, S., Diep, B.A., Perdreau-Remington, F. and Riley, L.W., 2013. Clonal composition and community clustering of drug-susceptible and -resistant *Escherichia coli* isolates from bloodstream infections. *Antimicrobial Agents and Chemotherapy,* 57 (1), 490-497.

Ajiboye, R.M., Solberg, O.D., Lee, B.M., Raphael, E., Debroy, C. and Riley, L.W., 2009. Global spread of mobile antimicrobial drug resistance determinants in human and animal *Escherichia coli* and *Salmonella* strains causing community-acquired infections. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America,* 49 (3), 365-371.

Al-Hasan, M.N., Eckel-Passow, J.E. and Baddour, L.M., 2010. Bacteremia complicating Gram-negative urinary tract infections: a population-based study. *The Journal of Infection,* 60 (4), 278-285.

Alhashash, F., Wang, X., Paszkiewicz, K., Diggle, M., Zong, Z. and McNally, A., 2016. Increase in bacteraemia cases in the East Midlands region of the UK due to MDR *Escherichia coli* ST73: high levels of genomic and plasmid diversity in causative isolates. *The Journal of Antimicrobial Chemotherapy,* 71 (2), 339-343.

Alhashash, F., Weston, V., Diggle, M. and McNally, A., 2013. Multidrug-Resistant *Escherichia coli* Bacteremia. *Emerging Infectious Diseases,* 19 (10), 1699-1701.

Amann, R.I., Binder, B.J., Olson, R.J., Chisholm, S.W., Devereux, R. and Stahl, D.A., 1990. Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Applied and Environmental Microbiology,* 56 (6), 1919-1925.

Aslam, M., Toufeer, M., Narvaez Bravo, C., Lai, V., Rempel, H., Manges, A. and Diarra, M.S., 2014. Characterization of Extraintestinal Pathogenic *Escherichia coli* isolated from retail poultry meats from Alberta, Canada. *International Journal of Food Microbiology,* 177, 49-56.

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

Avery, S.M., Moore, A. and Hutchison, M.L., 2004. Fate of *Escherichia coli* originating from livestock faeces deposited directly onto pasture. *Letters in Applied Microbiology,* 38 (5), 355-359.

Bajpai, T., Pandey, M., Varma, M. and Bhatambare, G.S., 2017. Prevalence of TEM, SHV, and CTX-M Beta-Lactamase genes in the urinary isolates of a tertiary care hospital. *Avicenna Journal of Medicine,* 7 (1), 12-16.

Banerjee, R. and Johnson, J.R., 2014. A new clone sweeps clean: the enigmatic emergence of *Escherichia coli* sequence type 131. *Antimicrobial Agents and Chemotherapy,* 58 (9), 4997-5004.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A. and Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology,* 19 (5), 455-477.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A. and Horvath, P., 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science (New York, N.Y.),* 315 (5819), 1709-1712.

Bauchart, P., Germon, P., Bree, A., Oswald, E., Hacker, J. and Dobrindt, U., 2010. Pathogenomic comparison of human extraintestinal and avian pathogenic *Escherichia coli* – search for factors involved in host specificity or zoonotic potential. *Microbial Pathogenesis,* 49 (3), 105-115.

Belanger, L., Garenaux, A., Harel, J., Boulianne, M., Nadeau, E. and Dozois, C.M., 2011. *Escherichia coli* from animal reservoirs as a potential source of human extraintestinal pathogenic E. coli. *FEMS Immunology and Medical Microbiology,* 62 (1), 1-10.

Bergeron, C.R., Prussing, C., Boerlin, P., Daignault, D., Dutil, L., Reid-Smith, R., Zhanel, G.G. and Manges, A.R., 2012. Chicken as Reservoir for Extraintestinal Pathogenic *Escherichia coli* in Humans, Canada. *Emerging Infectious Diseases,* 18 (3), 415-421.

Bergsbaken, T. and Cookson, B.T., 2009. Innate immune response during *Yersinia* infection: critical modulation of cell death mechanisms through phagocyte activation. *Journal of Leukocyte Biology,* 86 (5), 1153-1158.

Bergthorsson, U. and Ochman, H., 1998. Distribution of chromosome length variation in natural isolates of *Escherichia coli*. *Molecular Biology and Evolution,* 15 (1), 6-16.

Bert, F., Johnson, J.R., Ouattara, B., Leflon-Guibout, V., Johnston, B., Marcon, E., Valla, D., Moreau, R. and Nicolas-Chanoine, M.H., 2010. Genetic diversity and virulence profiles of *Escherichia coli* isolates causing spontaneous bacterial peritonitis and bacteremia in patients with cirrhosis. *Journal of Clinical Microbiology,* 48 (8), 2709-2714.

Berthe, T., Ratajczak, M., Clermont, O., Denamur, E. and Petit, F., 2013. Evidence for coexistence of distinct *Escherichia coli* populations in various aquatic environments and their survival in estuary water. *Applied and Environmental Microbiology,* 79 (15), 4684-4693.

Blyton, M.D., Pi, H., Vangchhia, B., Abraham, S., Trott, D.J., Johnson, J.R. and Gordon, D.M., 2015. Genetic Structure and Antimicrobial Resistance of *Escherichia coli* and Cryptic Clades in Birds with Diverse Human Associations. *Applied and Environmental Microbiology,* 81 (15), 5123-5133.

Bok, E., Mazurek, J., Stosik, M., Wojciech, M. and Baldy-Chudzik, K., 2014. Prevalence of Virulence Determinants and Antimicrobial Resistance among Commensal *Escherichia coli* Derived from Dairy and Beef Cattle. *International Journal of Environmental Research and Public Health,* 12 (1), 970-985.

Bonardi, S., Bruini, I., D'Incau, M., Van Damme, I., Carniel, E., Bremont, S., Cavallini, P., Tagliabue, S. and Brindani, F., 2016. Detection, seroprevalence and antimicrobial resistance of *Yersinia enterocolitica* and *Yersinia pseudotuberculosis* in pig tonsils in Northern Italy. *International Journal of Food Microbiology,* 235, 125-132.

Boor, K.J., 2006. Bacterial Stress Responses: What Doesn't Kill Them Can Make Them Stronger. *PLoS Biology,* 4 (1), e23.

Bottone, E.J., 1999. *Yersinia enterocolitica*: overview and epidemiologic correlates. *Microbes and Infection,* 1 (4), 323-333.

Bottone, E.J., 1997. *Yersinia enterocolitica*: the charisma continues. *Clinical Microbiology Reviews,* 10 (2), 257-276.

Bouckaert, R., Heled, J., Kuhnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A. and Drummond, A.J., 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology,* 10 (4), e1003537.

Boyd, E.F., and Hartl, D.L., 1998. Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a phylogenetically clustered distribution. *Journal of Bacteriology,* 180 (5), 1159-1165.

Bradford, P.A., 2001. Extended-spectrum beta-lactamases in the 21st century: Characterization, epidemiology, and detection of this important resistance threat. *Clinical Microbiology Reviews,* 14 (4), 933-951; 933.

Brennan, F.P., Abram, F., Chinalia, F.A., Richards, K.G. and O'Flaherty, V., 2010. Characterization of environmentally persistent *Escherichia coli* isolates leached from an Irish soil. *Applied and Environmental Microbiology,* 76 (7), 2175-2180.

Brouns, S.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J., Snijders, A.P., Dickman, M.J., Makarova, K.S., Koonin, E.V. and van der Oost, J., 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science (New York, N.Y.),* 321 (5891), 960-964.

Brubaker, R.R., 1991. Factors promoting acute and chronic diseases caused by *Yersiniae*. *Clinical Microbiology Reviews,* 4 (3), 309-324.

Carniel, E., 2003. Evolution of pathogenic *Yersinia*, some lights in the dark. *Advances in Experimental Medicine and Biology,* 529, 3-12.

Carniel, E., 1999. The *Yersinia* high-pathogenicity island. *International Microbiology: The Official Journal of the Spanish Society for Microbiology,* 2 (3), 161-167.

Chen, L., Zheng, D., Liu, B., Yang, J. and Jin, Q., 2015. VFDB 2016: hierarchical and refined dataset for big data analysis – 10 years on. *Nucleic Acids Research,* 44, D694-D697.

Chen, P., Hung, C., Huang, P., Chen, J., Huang, I., Chen, W., Chiou, Y., Hung, W., Wang, J. and Cheng, M., 2016. Characteristics of CTX-M Extended-Spectrum β-Lactamase-Producing *Escherichia coli* Strains Isolated from Multiple Rivers in Southern Taiwan. *Applied and Environmental Microbiology,* 82 (6), 1889-1897.

Chiles, M.C., Madhusudhan, K.T., Greenson, J.K., Scott, M.A., Bronner, M.P., Havens, J.M., Dean, P.J. and Lamps, L.W., 2002. Pathogenic *Yersinia pseudotuberculosis* and *Yersinia enterocolitica* DNA is detected in bowel and mesenteric nodes from Crohn's disease patients. *Modern Pathology,* 15 (1), 518.

Clark, G., Paszkiewicz, K., Hale, J., Weston, V., Constantinidou, C., Penn, C.W., Achtman, M. and McNally, A., 2012. Genomic and molecular epidemiology analysis of clinical *Escherichia coli* ST131 isolates suggests circulation of a genetically monomorphic but phenotypically heterogeneous ExPEC clone. *Journal of Antimicrobial Chemotherapy,* 67, 868-77.

Clarridge, J.E. 3rd, 2004. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews*, 17 (4), 840-862.

Clements, A., Young, J.C., Constantinou, N. and Frankel, G., 2012. Infection strategies of enteric pathogenic *Escherichia coli*. *Gut Microbes,* 3 (2), 71-87.

Clermont, O., Christenson, J.K., Denamur, E. and Gordon, D.M., 2013. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environmental Microbiology Reports,* 5 (1), 58-65.

Clermont, O., Bonacorsi, S. and Bingen, E., 2000. Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Applied and Environmental Microbiology,* 66 (10), 4555-4558.

Cohan, F., and Kopac, S., 2011. Microbial Genomics: *E. coli* Relatives Out of Doors and Out of Body. *Current Biology,* 21 (15), R587-R589.

Cohen Stuart, J., van den Munckhof, T., Voets, G., Scharringa, J., Fluit, A. and Hall, M.L., 2012. Comparison of ESBL contamination in organic and conventional retail chicken meat. *International Journal of Food Microbiology,* 154 (3), 212-214.

Coleman, B.L., Louie, M., Salvadori, M.I., McEwen, S.A., Neumann, N., Sibley, K., Irwin, R.J., Jamieson, F.B., Daignault, D., Majury, A., Braithwaite, S., Crago, B. and McGeer, A.J., 2013. Contamination of Canadian private drinking water sources with antimicrobial resistant *Escherichia coli*. *Water Research,* 47 (9), 3026-3036.

Collyn, F., Billault, A., Mullet, C., Simonet, M. and Marceau, M., 2004. YAPI, a New *Yersinia pseudotuberculosis* Pathogenicity Island. *Infection and Immunity,* 72 (8), 4784-4790.

Conway, T., and Cohen, P.S., 2015. Commensal and Pathogenic *Escherichia coli* Metabolism in the Gut. *Microbiology Spectrum,* 3 (3), 10.1128/microbiolspec.MBP-0006-2014.

Cookson, S.T., and Nataro, J.P., 1996. Characterization of HEp-2 cell projection formation induced by diffusely adherent *Escherichia coli*. *Microbial Pathogenesis,* 21 (6), 421-434.

Cornelis, G.R. and Wolf-Watz, H., 1997. The *Yersinia* Yop virulon: a bacterial system for subverting eukaryotic cells. *Molecular Microbiology,* 23 (5), 861-867.

Cortes, P., Blanc, V., Mora, A., Dahbi, G., Blanco, J.E., Blanco, M., Lopez, C., Andreu, A., Navarro, F., Alonso, M.P., Bou, G., Blanco, J. and Llagostera, M., 2010. Isolation and characterization of potentially pathogenic antimicrobial-resistant *Escherichia coli* strains from chicken and pig farms in Spain. *Applied and Environmental Microbiology,* 76 (9), 2799-2805.

Cristovao, F., Alonso, C.A., Igrejas, G., Sousa, M., Silva, V., Pereira, J.E., Lozano, C., Cortes-Cortes, G., Torres, C. and Poeta, P., 2017. Clonal diversity of extended-spectrum beta-lactamase producing *Escherichia coli* isolates in fecal samples of wild animals. *FEMS Microbiology Letters,* 364 (5), 10.1093/femsle/fnx039.

Croucher, N.J., Page, A.J., Connor, T.R., Delaney, A.J., Keane, J.A., Bentley, S.D., Parkhill, J. and Harris, S.R., 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research,* 43 (3), e15.

Croxall, G., Weston, V., Joseph, S., Manning, G., Cheetham, P. and McNally, A., 2011a. Increased Human Pathogenic Potential of *Escherichia coli* from Polymicrobial Urinary Tract Infections in Comparison to Isolates from Monomicrobial Culture Samples. *Journal of Medical Microbiology,* 60, 102-102-109.

Croxall, G., Hale, J., Weston, V., Manning, G., Cheetham, P., Achtman, M. and McNally, A., 2011b. Molecular epidemiology of extraintestinal pathogenic *Escherichia coli* isolates from a regional cohort of elderly patients highlights the prevalence of ST131 strains with increased antimicrobial resistance in both community and hospital care settings. *Journal of Antimicrobial Chemotherapy,* 66 (11), 2501-2508.

Croxen, M. and Finlay, B., 2010. Molecular mechanisms of *Escherichia coli* pathogenicity. *Nature Reviews Microbiology,* 8, 26-26-38.

Cuzon, G., Naas, T., Lesenne, A., Benhamou, M. and Nordmann, P., 2010. Plasmid-mediated carbapenem-hydrolysing OXA-48 β-lactamase in *Klebsiella pneumoniae* from Tunisia. *International Journal of Antimicrobial Agents,* 36 (1), 91-93.

Danzeisen, J.L., Wannemuehler, Y., Nolan, L.K. and Johnson, T.J., 2013. Comparison of multilocus sequence analysis and virulence genotyping of *Escherichia coli* from live birds, retail poultry meat, and human extraintestinal infection. *Avian Diseases,* 57 (1), 104-108.

Darfeuille-Michaud, A., 2002. Adherent-invasive *Escherichia coli*: a putative new *E. coli* pathotype associated with Crohn's disease. *International Journal of Medical Microbiology,* 292 (3), 185-193.

de Been, M., Lanza, V.F., de Toro, M., Scharringa, J., Dohmen, W., Du, Y., Hu, J., Lei, Y., Li, N., Tooming-Klunderud, A., Heederik, D.J.J., Fluit, A.C., Bonten, M.J.M., Willems, R.J.L., de, l.C. and van Schaik, W., 2014. Dissemination of Cephalosporin Resistance Genes between *Escherichia coli* Strains from Farm Animals and Humans by Specific Plasmid Lineages. *PLoS Genetics,* 10 (12), e1004776.

Dearlove, B.L., Cody, A.J., Pascoe, B., Meric, G., Wilson, D.J. and Sheppard, S.K., 2016. Rapid host switching in generalist *Campylobacter* strains erodes the signal for tracing human infections. *The ISME Journal,* 10 (3), 721-729.

Denamur, E., 2011. The 2011 Shiga toxin-producing *Escherichia coli* O104:H4 German outbreak: a lesson in genomic plasticity. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases,* 17 (8), 1124-1125.

Dias, R.C.S., Marangoni, D.V., Smith, S.P., Alves, E.M., Pellegrino, F.L.P.C., Riley, L.W. and Moreira, B.M., 2009. Clonal composition of *Escherichia coli* causing community-acquired urinary tract infections in the state of Rio de Janeiro, Brazil. *Microbial Drug Resistance,* 15 (4), 303-308.

Dolejska, M., Matulova, M., Kohoutova, L., Literak, I., Bardon, J. and Cizek, A., 2011a. Extended-spectrum beta-lactamase-producing *Escherichia coli* in turkey meat production farms in the Czech Republic: national survey reveals widespread isolates with bla(SHV-12) genes on IncFII plasmids. *Letters in Applied Microbiology,* 53 (3), 271-277.

Dolejska, M., Frolkova, P., Florek, M., Jamborova, I., Purgertova, M., Kutilova, I., Cizek, A., Guenther, S. and Literak, I., 2011b. CTX-M-15-producing *Escherichia coli* clone B2-O25b-ST131 and *Klebsiella* spp. isolates in municipal wastewater treatment plant effluents. *Journal of Antimicrobial Chemotherapy,* 66 (12), 2784-2790.

Dominguez, E., Zarazaga, M., Saenz, Y., Brinas, L. and Torres, C., 2002. Mechanisms of antibiotic resistance in *Escherichia coli* isolates obtained from healthy children in Spain. *Microbial Drug Resistance (Larchmont, N.Y.),* 8 (4), 321-327.

Doumith, M., Day, M., Ciesielczuk, H., Hope, R., Underwood, A., Reynolds, R., Wain, J., Livermore, D.M. and Woodford, N., 2015. Rapid Identification of Major *Escherichia coli* Sequence Types Causing Urinary Tract and Bloodstream Infections. *Journal of Clinical Microbiology,* 53 (1), 160-166.

Dublan Mde, L., Ortiz-Marquez, J.C., Lett, L. and Curatti, L., 2014. Plant-adapted *Escherichia coli* show increased lettuce colonizing ability, resistance to oxidative stress and chemotactic response. *PloS One,* 9 (10), e110416.

Dumbrell, A.J., Nelson, M., Helgason, T., Dytham, C., and Fitter, A.H., 2010. Relative roles of niche and neutral processes in structuring a soil microbial community. *The ISME Journal*, 4 (3), 337-345.

Duriez, P., Clermont, O., Bonacorsi, S., Bingen, E., Chaventre, A., Elion, J., Picard, B. and Denamur, E., 2001. Commensal *Escherichia coli* isolates are phylogenetically distributed among geographically distinct human populations. *Microbiology (Reading, England),* 147 (Pt 6), 1671-1676.

Egervarn, M., Borjesson, S., Byfors, S., Finn, M., Kaipe, C., Englund, S. and Lindblad, M., 2014. *Escherichia coli* with extended-spectrum beta-lactamases or transferable AmpC beta-lactamases and *Salmonella* on meat imported into Sweden. *International Journal of Food Microbiology,* 171, 8-14.

Eppinger, M., Rosovitz, M.J., Fricke, W.F., Rasko, D.A., Kokorina, G., Fayolle, C., Lindler, L.E., Carniel, E. and Ravel, J., 2007. The complete genome sequence of *Yersinia pseudotuberculosis* IP31758, the causative agent of Far East scarlet-like fever. *PLoS Genet,* 3 (8), e142.

Escobar-Paramo, P., Sabbagh, A., Darlu, P., Pradillon, O., Vaury, C., Denamur, E. and Lecointre, G., 2004. Decreasing the effects of horizontal gene transfer on bacterial phylogeny: the *Escherichia coli* case study. *Molecular Phylogenetics and Evolution,* 30 (1), 243-250.

Ewers, C., Bethe, A., Semmler, T., Guenther, S. and Wieler, L.H., 2012. Extended-spectrum beta-lactamase-producing and AmpC-producing *Escherichia coli* from livestock and companion animals, and their putative impact on public health: a global perspective. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases,* 18 (7), 646-655.

Ewers, C., Bethe, A., Stamm, I., Grobbel, M., Kopp, P.A., Guerra, B., Stubbe, M., Doi, Y., Zong, Z., Kola, A., Schaufler, K., Semmler, T., Fruth, A., Wieler, L.H. and Guenther, S., 2014. CTX-M-15-D-ST648 *Escherichia coli* from companion animals and horses: another pandemic clone combining multiresistance and extraintestinal virulence? *The Journal of Antimicrobial Chemotherapy,* 69 (5), 1224-1230.

Ewers, C., Grobbel, M., Stamm, I., Kopp, P.A., Diehl, I., Semmler, T., Fruth, A., Beutlich, J., Guerra, B., Wieler, L.H. and Guenther, S., 2010. Emergence of human pandemic O25:H4-ST131 CTX-M-

15 extended-spectrum-β-lactamase-producing *Escherichia coli* among companion animals. *Journal of Antimicrobial Chemotherapy,* 65 (4), 651-660.

Fang, H., Ataker, F., Hedin, G. and Dornbusch, K., 2008. Molecular epidemiology of Extended-Spectrum {beta}-Lactamases among *Escherichia coli* isolates collected in a Swedish hospital and its associated health care facilities from 2001 to 2006. *Journal of Clinical Microbiology,* 46 (2), 707-712.

Feil, E.J., Li, B.C., Aanensen, D.M., Hanage, W.P. and Spratt, B.G., 2004. eBURST: Inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *Journal of Bacteriology,* 186 (5), 1518-1530.

Flores-Mireles, A., Walker, J.N., Caparon, M. and Hultgren, S.J., 2015. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nature Reviews Microbiology,* 13 (5), 269-284.

Foster, J.W., 2004. *Escherichia coli* acid resistance: tales of an amateur acidophile. *Nature Reviews Microbiology,* 2 (11), 898-907.

Fox, M.G. and Sorhannus, U.M., 2003. *RpoA*: A Useful Gene for Phylogenetic Analysis in Diatoms. *Journal of Eukaryotic Microbiology*, 50 (6), 471-475.

Foxman, B., 2014. Urinary tract infection syndromes: occurrence, recurrence, bacteriology, risk factors, and disease burden. *Infectious Disease Clinics of North America,* 28 (1), 1-13.

Francisco, A.P., Vaz, C., Monteiro, P.T., Melo-Cristino, J., Ramirez, M. and Carrico, J.A., 2012. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics,* 13, 87-2105-13-87.

Frank, C., Werber, D., Cramer, J.P., Askar, M., Faber, M., an der Heiden, M., Bernard, H., Fruth, A., Prager, R., Spode, A., Wadl, M., Zoufaly, A., Jordan, S., Kemper, M.J., Follin, P., Muller, L., King, L.A., Rosner, B., Buchholz, U., Stark, K., Krause, G. and HUS Investigation Team, 2011. Epidemic profile of Shiga-toxin-producing *Escherichia coli* O104:H4 outbreak in Germany. *The New England Journal of Medicine,* 365 (19), 1771-1780.

Fukushima, H., Matsuda, Y., Seki, R., Tsubokura, M., Takeda, N., Shubin, F.N., Paik, I.K. and Zheng, X.B., 2001. Geographical heterogeneity between Far Eastern and Western countries in prevalence of the virulence plasmid, the superantigen *Yersinia pseudotuberculosis*-derived mitogen, and the high-pathogenicity island among *Yersinia pseudotuberculosis* strains. *Journal of Clinical Microbiology,* 39 (10), 3541-3547.

G.F. Asensi, EMF, d.R., EM, D.A., Rodrigues, D.d.P., J.T. Silva and V.M.F. Paschoalin, 2009. Detection of *Escherichia coli* and Salmonella in chicken rinse carcasses. *British Food Journal,* 111 (6), 517-527.

Gage, K.L., 2012. 320 - Plague and Other *Yersinia* Infections A2 - Goldman, Lee. *In:* A.I. Schafer, ed., *Goldman's Cecil Medicine (Twenty-Fourth Edition).* Philadelphia: W.B. Saunders, 2012, pp. 1895-1900.

Galindo, C.L., Rosenzweig, J.A., Kirtley, M.L. and Chopra, A.K., 2011. Pathogenesis of *Y. enterocolitica* and *Y. pseudotuberculosis* in Human Yersiniosis. *Journal of Pathogens,* 2011, 182051.

Gaulin, C., Levac, E., Ramsay, D., Dion, R., Ismail, J., Gingras, S. and Lacroix, C., 2012. *Escherichia coli* O157:H7 outbreak linked to raw milk cheese in Quebec, Canada: use of exact probability calculation and casecase study approaches to foodborne outbreak investigation. *Journal of Food Protection,* 75 (5), 812-818.

Gest, H., 2004. The discovery of microorganisms by Robert Hooke and Antoni Van Leeuwenhoek, fellows of the Royal Society. *Notes and Records of the Royal Society of London,* 58 (2), 187-201.

Gibreel, T.M., Dodgson, A.R., Cheesbrough, J., Fox, A.J., Bolton, F.J. and Upton, M., 2012. Population structure, virulence potential and antibiotic susceptibility of uropathogenic *Escherichia coli* from Northwest England. *The Journal of Antimicrobial Chemotherapy,* 67 (2), 346-356.

Gomi, R., Matsuda, T., Matsumura, Y., Yamamoto, M., Tanaka, M., Ichiyama, S. and Yoneda, M., 2017a. Occurrence of Clinically Important Lineages, Including the Sequence Type 131 C1-M27 Subclone, among Extended-Spectrum-beta-Lactamase-Producing *Escherichia coli* in Wastewater. *Antimicrobial Agents and Chemotherapy,* 61 (9), 10.1128/AAC.00564-17.

Gomi, R., Matsuda, T., Matsumura, Y., Yamamoto, M., Tanaka, M., Ichiyama, S. and Yoneda, M., 2017b. Whole-Genome Analysis of Antimicrobial-Resistant and Extraintestinal Pathogenic *Escherichia coli* in River Water. *Applied and Environmental Microbiology,* 83 (5), 10.1128/AEM.02703-16.

Goncalves, L.F., de Oliveira Martins-Junior, P., de Melo, A.B.F., da Silva, R.C.R.M., de Paulo Martins, V., Pitondo-Silva, A. and de Campos, T.A., 2016. Multidrug resistance dissemination by extended-spectrum beta-lactamase-producing *Escherichia coli* causing community-acquired urinary tract infection in the Central-Western Region, Brazil. *Journal of Global Antimicrobial Resistance,* 6, 1-4.

Gordienko, E.N., Kazanov, M.D. and Gelfand, M.S., 2013. Evolution of pan-genomes of *Escherichia coli*, *Shigella* spp., and *Salmonella enterica*. *Journal of Bacteriology,* 195 (12), 2786-2792.

Guenther, S., Aschenbrenner, K., Stamm, I., Bethe, A., Semmler, T., Stubbe, A., Stubbe, M., Batsajkhan, N., Glupczynski, Y., Wieler, L.H. and Ewers, C., 2012. Comparable high rates of extended-spectrum-beta-lactamase-producing *Escherichia coli* in birds of prey from Germany and Mongolia. *PloS One,* 7 (12), e53039.

Guenther, S., Ewers, C. and Wieler, L.H., 2011. Extended-Spectrum Beta-Lactamases Producing *E. coli* in Wildlife, yet Another Form of Environmental Pollution? *Frontiers in Microbiology,* 2, 246.

Guenther, S., Grobbel, M., Beutlich, J., Bethe, A., Friedrich, N.D., Goedecke, A., Lubke-Becker, A., Guerra, B., Wieler, L.H. and Ewers, C., 2010. CTX-M-15-type extended-spectrum beta-lactamases-producing *Escherichia coli* from wild birds in Germany. *Environmental Microbiology Reports,* 2 (5), 641-645.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England),* 29 (8), 1072-1075.

Hacker, J., and Blum-Oehler, G., 2007. In appreciation of Theodor Escherich. *Nature Reviews Microbiology,* 5, 902-902.

Hadfield, J., Croucher, N.J., Goater, R.J., Abudahab, K., Aanensen, D.M. and Harris, S.R., 2018. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics,* 34 (2), 292-293.

Hall, B.G., Ehrlich, G.D. and Hu, F.Z., 2010. Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. *Microbiology,* 156 (4), 1060-1068.

Heath, P.T., Nik Yusoff, N.K. and Baker, C.J., 2003. Neonatal meningitis. *Archives of Disease in Childhood.Fetal and Neonatal Edition,* 88 (3), F173-8.

Heroven, A.K. and Dersch, P., 2014. Coregulation of host-adapted metabolism and virulence by pathogenic *Yersinia*e. *Frontiers in Cellular and Infection Microbiology,* 4, 146.

Ho, N., Kondakova, A.N., Knirel, Y.A. and Creuzenet, C., 2008. The biosynthesis and biological role of 6-deoxyheptose in the lipopolysaccharide O-antigen of *Yersinia pseudotuberculosis*. *Molecular Microbiology,* 68 (2), 424-447.

Howie, R.L., Folster, J.P., Bowen, A., Barzilay, E.J. and Whichard, J.M., 2010. Reduced azithromycin susceptibility in *Shigella sonnei*, United States. *Microbial Drug Resistance (Larchmont, N.Y.),* 16 (4), 245-248.

Hu, J., Shi, J., Chang, H., Li, D., Yang, M. and Kamagata, Y., 2008. Phenotyping and genotyping of antibiotic-resistant *Escherichia coli* isolated from a natural river basin. *Environmental Science & Technology,* 42 (9), 3415-3420.

Hurst, M.R., Becher, S.A., Young, S.D., Nelson, T.L. and Glare, T.R., 2011. *Yersinia entomophaga* sp. nov., isolated from the New Zealand grass grub *Costelytra zealandica*. *International Journal of Systematic and Evolutionary Microbiology,* 61 (Pt 4), 844-849.

Hussain, A., Shaik, S., Ranjan, A., Nandanwar, N., Tiwari, S.K., Majid, M., Baddam, R., Qureshi, I.A., Semmler, T., Wieler, L.H., Islam, M.A., Chakravortty, D. and Ahmed, N., 2017. Risk of Transmission of Antimicrobial Resistant *Escherichia coli* from Commercial Broiler and Free-Range Retail Chicken in India. *Frontiers in Microbiology,* 8, 2120.

Hubbell, S.P., 2001. The unified neutral theory of biodiversity and biogeography (MPB-32). New Jersey: Princeton University Press.

Illumina, 2017. An Introduction to Next-Generation Sequencing Technology. https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Ingle, D.J., Clermont, O., Skurnik, D., Denamur, E., Walk, S.T. and Gordon, D.M., 2011. Biofilm Formation by and Thermal Niche and Virulence Characteristics of *Escherichia* spp. *Applied and Environmental Microbiology,* 77 (8), 2695-2700.

Inoue, M., Nakashima, H., Ueba, O., Ishida, T., Date, H., Kobashi, S., Takagi, K., Nishu, T. and Tsubokura, M., 1984. Community outbreak of *Yersinia pseudotuberculosis*. *Microbiology and Immunology,* 28 (8), 883-891.

Jaakkola, K., Somervuo, P. and Korkeala, H., 2015. Comparative Genomic Hybridization Analysis of *Yersinia enterocolitica* and *Yersinia pseudotuberculosis* Identifies Genetic Traits to Elucidate Their Different Ecologies. *BioMed Research International,* 2015, 760494.

Jakobsen, L., Garneau, P., Bruant, G., Harel, J., Olsen, S.S., Porsbo, L.J., Hammerum, A.M. and Frimodt-Moller, N., 2012. Is *Escherichia coli* urinary tract infection a zoonosis? Proof of direct link with production animals and meat. *European Journal of Clinical Microbiology & Infectious Diseases: Official Publication of the European Society of Clinical Microbiology,* 31 (6), 1121-1129.

Jakobsen, L., Spangholm, D.J., Pedersen, K., Jensen, L.B., Emborg, H.D., Agerso, Y., Aarestrup, F.M., Hammerum, A.M. and Frimodt-Moller, N., 2010. Broiler chickens, broiler chicken meat, pigs and pork as sources of ExPEC related virulence genes and resistance in *Escherichia coli* isolates from community-dwelling humans and UTI patients. *International Journal of Food Microbiology,* 142 (1-2), 264-272.

Jalava, K., Hakkinen, M., Valkonen, M., Nakari, U.M., Palo, T., Hallanvuo, S., Ollgren, J., Siitonen, A. and Nuorti, J.P., 2006. An outbreak of gastrointestinal illness and erythema nodosum from grated carrots contaminated with *Yersinia pseudotuberculosis*. *The Journal of Infectious Diseases,* 194 (9), 1209-1216.

Jang, J., Suh, Y.S., Di, D.Y., Unno, T., Sadowsky, M.J. and Hur, H.G., 2013. Pathogenic *Escherichia coli* strains producing extended-spectrum beta-lactamases in the Yeongsan River basin of South Korea. *Environmental Science & Technology,* 47 (2), 1128-1136.

Janke, B., Dobrindt, U., Hacker, J. and Blum-Oehler, G., 2001. A subtractive hybridisation analysis of genomic differences between the uropathogenic *E. coli* strain 536 and the *E. coli* K-12 strain MG1655. *FEMS Microbiology Letters,* 199 (1), 61-66.

Jelinek, T. and Kollaritsch, H., 2008. Vaccination with Dukoral against travelers' diarrhea (ETEC) and cholera. *Expert Review of Vaccines,* 7 (5), 561-567.

Jena, J., Sahoo, R.K., Debata, N.K. and Subudhi, E., 2017. Prevalence of TEM, SHV, and CTX-M genes of extended-spectrum beta-lactamase-producing *Escherichia coli* strains isolated from urinary tract infections in adults. *3 Biotech,* 7 (4), 244-017-0879-2.

Jeraldo, P., Sipos, M., Chia, N., Brulc, J.M., Dhillon, A.S., Konkel, M.E., Larson, C.L., Nelson, K.E., Qu, A., Schook, L.B., Yang, F., White, B.A. and Goldenfeld, N., 2012. Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes. *Proceedings of the National Academy of Sciences of the USA*, 109 (25), 9692-9698.

Johnson, J.R., Johnston, B., Clabots, C., Kuskowski, M.A. and Castanheira, M., 2010. *Escherichia coli* Type ST131 as the Major Cause of Serious Multidrug-Resistant *E. coli* Infections in the United States. *Clinical Infectious Diseases,* 51, 286-286-294.

Johnson, J.R., and Clabots, C., 2006. Sharing of virulent *Escherichia coli* clones among household members of a woman with acute cystitis. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America,* 43 (10), e101-8.

Johnson, J.R., Delavari, P., O'Bryan, T.T., Smith, K.E. and Tatini, S., 2005a. Contamination of retail foods, particularly turkey, from community markets (Minnesota, 1999-2000) with antimicrobial-resistant and extraintestinal pathogenic *Escherichia coli*. *Foodborne Pathogens and Disease,* 2 (1), 38-49.

Johnson, J.R., Kuskowski, M.A., Smith, K., O'Bryan, T.T. and Tatini, S., 2005b. Antimicrobial-resistant and extraintestinal pathogenic *Escherichia coli* in retail foods. *The Journal of Infectious Diseases,* 191 (7), 1040-1049.

Johnson, J.R., Porter, S.B., Johnston, B., Thuras, P., Clock, S., Crupain, M. and Rangan, U., 2017. Extraintestinal Pathogenic and Antimicrobial-Resistant *Escherichia coli*, Including Sequence Type 131 (ST131), from Retail Chicken Breasts in the United States in 2013. *Applied and Environmental Microbiology,* 83 (6), 10.1128/AEM.02956-16.

Johnson, J.R., Johnston, B., Clabots, C.R., Kuskowski, M.A., Roberts, E. and DebRoy, C., 2008. Virulence Genotypes and Phylogenetic Background of *Escherichia coli* Serogroup O6 Isolates from Humans, Dogs, and Cats. *Journal of Clinical Microbiology,* 46 (2), 417-422.

Johnson, J.R., Johnston, B., Clabots, C., Kuskowski, M.A., Swaroop, P., DebRoy, C., Nowicki, B. and Rice, J., 2010. *Escherichia coli* sequence type ST131 as an emerging fluoroquinolone-resistant uropathogen among renal transplant recipients. *Antimicrobial Agents and Chemotherapy,* 54 (1), 546-546-550.

Johnson, J.R., Menard, M., Johnston, B., Kuskowski, M.A., Nichol, K. and Zhanel, G.G., 2009. Epidemic clonal groups of *Escherichia coli* as a cause of antimicrobial-resistant urinary tract infections in Canada, 2002 to 2004. *Antimicrobial Agents and Chemotherapy,* 53 (7), 2733-2739.

Johnson, J.R. and Stell, A.L., 2000. Extended virulence genotypes of *Escherichia coli* strains from patients with urosepsis in relation to phylogeny and host compromise. *The Journal of Infectious Diseases,* 181, 261-272.

Johnson, T.J., Wannemuehler, Y., Johnson, S.J., Stell, A.L., Doetkott, C., Johnson, J.R., Kim, K.S., Spanjaard, L. and Nolan, L.K., 2008. Comparison of Extraintestinal Pathogenic *Escherichia coli* Strains from Human and Avian Sources Reveals a Mixed Subset Representing Potential Zoonotic Pathogens. *Applied and Environmental Microbiology,* 74 (22), 7043-7050.

Jorgensen, S.B., Soraas, A.V., Arnesen, L.S., Leegaard, T.M., Sundsfjord, A. and Jenum, P.A., 2017. A comparison of extended spectrum beta-lactamase producing *Escherichia coli* from clinical, recreational water and wastewater samples associated in time and location. *PloS One,* 12 (10), e0186576.

Kaas, R.S., Friis, C., Ussery, D.W. and Aarestrup, F.M., 2012. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics,* 13, 577-2164-13-577.

Kaasch, A.J., Dinter, J., Goeser, T., Plum, G. and Seifert, H., 2012. *Yersinia pseudotuberculosis* bloodstream infection and septic arthritis: case report and review of the literature. *Infection,* 40 (2), 185-190.

Kallonen, T., Brodrick, H.J., Harris, S.R., Corander, J., Brown, N.M., Martin, V., Peacock, S.J. and Parkhill, J., 2017. Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Research,* 27, 1437-1449.

Kang, C.I., Cha, M.K., Kim, S.H., Ko, K.S., Wi, Y.M., Chung, D.R., Peck, K.R., Lee, N.Y. and Song, J.H., 2013. Clinical and molecular epidemiology of community-onset bacteremia caused by extended-spectrum beta-lactamase-producing *Escherichia coli* over a 6-year period. *Journal of Korean Medical Science,* 28 (7), 998-1004.

Kangas, S., Takkinen, J., Hakkinen, M., Nakari, U., Johansson, T., Henttonen, H., Virtaluoto, L., Siitonen, A., Ollgren, J. and Kuusi, M., 2008. *Yersinia pseudotuberculosis* O:1 Traced to Raw Carrots, Finland. *Emerging Infectious Diseases,* 14 (12), 1959-1961.

Kaper, J.B., Nataro, J.P. and Mobley, H.L., 2004. Pathogenic *Escherichia coli*. *Nature Reviews.Microbiology,* 2 (2), 123-140.

Kappell, A.D., DeNies, M.S., Ahuja, N.H., Ledeboer, N.A., Newton, R.J. and Hristova, K.R., 2015. Detection of multi-drug resistant *Escherichia coli* in the urban waterways of Milwaukee, WI. *Frontiers in Microbiology,* 6, 336.

Karim, A., Poirel, L., Nagarajan, S. and Nordmann, P., 2001. Plasmid-mediated extended-spectrum β-lactamase (CTX-M-3 like) from India and gene association with insertion sequence ISEcp1. *FEMS Microbiology Letters,* 201 (2), 237-241.

Karmali, M.A., Steele, B.T., Petric, M. and Lim, C., 1983. Sporadic cases of haemolytic-uraemic syndrome associated with faecal cytotoxin and cytotoxin-producing *Escherichia coli* in stools. *Lancet (London, England),* 1 (8325), 619-620.

Kleinheinz, K.A., Joensen, K.G. and Larsen, M.V., 2014. Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and *E. coli*

virulence genes in bacteriophage and prophage nucleotide sequences. *Bacteriophage,* 4 (1), e27943.

Knothe, H., Shah, P., Krcmery, V., Antal, M. and Mitsuhashi, S., 1983. Transferable resistance to cefotaxime, cefoxitin, cefamandole and cefuroxime in clinical isolates of *Klebsiella pneumoniae* and Serratia marcescens. *Infection,* 11 (6), 315-317.

Koskela, K.A., Mattinen, L., Kalin-Manttari, L., Vergnaud, G., Gorge, O., Nikkari, S. and Skurnik, M., 2015. Generation of a CRISPR database for *Yersinia pseudotuberculosis* complex and role of CRISPR-based immunity in conjugation. *Environmental Microbiology,* 17 (11), 4306-4321.

Kupczok, A., Landan, G. and Dagan, T., 2015. The contribution of genetic recombination to CRISPR array evolution. *Genome Biology and Evolution,* 7 (7), 1925-1939.

Lane, M.C., Lockatell, V., Monterosso, G., Lamphier, D., Weinert, J., Hebel, J.R., Johnson, D.E. and Mobley, H.L.T., 2005. Role of Motility in the Colonization of Uropathogenic *Escherichia coli* in the Urinary Tract. *Infection and Immunity,* 73 (11), 7644-7656.

Lau, S.H., Kaufmann, M.E., Livermore, D.M., Woodford, N., Willshaw, G.A., Cheasty, T., Stamper, K., Reddy, S., Cheesbrough, J., Bolton, F.J., Fox, A.J. and Upton, M., 2008. UK epidemic *Escherichia coli* strains A-E, with CTX-M-15 beta-lactamase, all belong to the international O25:H4-ST131 clone. *The Journal of Antimicrobial Chemotherapy,* 62 (6), 1241-1244.

Laukkanen-Ninios, R., Didelot, X., Jolley, K., Morelli, G., Sangal, V., Kristo, P., Brehony, C., Imori, P., Fukushima, H., Siitonen, A., Tseneva, G., Voskressenskaya, E., Falcao, J., Korkeala, H., Maiden, M., Mazzoni, C., Carniel, E., Skurnik, M. and Achtman, M., 2011. Population structure of the *Yersinia pseudotuberculosis* complex according to multilocus sequence typing. *Environ.Microbiol.,* 13, 3114-27.

Laver, T., Harrison, J., O'Neill, P.A., Moore, K., Farbos, A., Paszkiewicz, K. and Studholme, D.J., 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification,* 3, 1-8.

Lawrence, J.G. and Ochman, H., 1998. Molecular archaeology of the *Escherichia coli* genome. *Proceedings of the National Academy of Sciences of the United States of America,* 95 (16), 9413-9417.

Lazarus, B., Paterson, D.L., Mollinger, J.L. and Rogers, B.A., 2015. Do human extraintestinal *Escherichia coli* infections resistant to expanded-spectrum cephalosporins originate from food-

producing animals? A systematic review. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America,* 60 (3), 439-452.

Letunic, I. and Bork, P., 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research,* 44 (W1), W242-5.

Leverstein-van Hall, M.A., Dierikx, C.M., Cohen Stuart, J., Voets, G.M., van den Munckhof, M.P., van Essen-Zandbergen, A., Platteel, T., Fluit, A.C., van de Sande-Bruinsma, N., Scharinga, J., Bonten, M.J., Mevius, D.J. and National ESBL surveillance group, 2011. Dutch patients, retail chicken meat and poultry share the same ESBL genes, plasmids and strains. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases,* 17 (6), 873-880.

Livermore, D.M., 1995. Beta-Lactamases in laboratory and clinical resistance. *Clinical Microbiology Reviews,* 8 (4), 557-584.

Livermore, D.M. and Woodford, N., 2006. The beta-lactamase threat in Enterobacteriaceae, *Pseudomonas* and *Acinetobacter*. *Trends in Microbiology,* 14 (9), 413-420.

Logue, C.M., Wannemuehler, Y., Nicholson, B.A., Doetkott, C., Barbieri, N.L. and Nolan, L.K., 2017. Comparative Analysis of Phylogenetic Assignment of Human and Avian ExPEC and Fecal Commensal *Escherichia coli* Using the (Previous and Revised) Clermont Phylogenetic Typing Methods and its Impact on Avian Pathogenic *Escherichia coli* (APEC) Classification. *Frontiers in Microbiology,* 8, 283.

Luo, C., Walk, S.T., Gordon, D.M., Feldgarden, M., Tiedje, J.M. and Konstantinidis, K.T., 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences of the United States of America,* 108 (17), 7200-7205.

Mabbett, A.N., Ulett, G.C., Watts, R.E., Tree, J.J., Totsika, M., Ong, C.Y., Wood, J.M., Monaghan, W., Looke, D.F., Nimmo, G.R., Svanborg, C. and Schembri, M.A., 2009. Virulence properties of asymptomatic bacteriuria *Escherichia coli*. *International Journal of Medical Microbiology,* 299 (1), 53-63.

Magiorakos, A.P., Srinivasan, A., Carey, R.B., Carmeli, Y., Falagas, M.E., Giske, C.G., Harbarth, S., Hindler, J.F., Kahlmeter, G., Olsson-Liljequist, B., Paterson, D.L., Rice, L.B., Stelling, J., Struelens, M.J., Vatopoulos, A., Weber, J.T. and Monnet, D.L., 2012. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard

definitions for acquired resistance. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases,* 18 (3), 268-281.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M. and Spratt, B.G., 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America,* 95 (6), 3140-3145.

Maluta, R.P., Logue, C.M., Casas, M.R., Meng, T., Guastalli, E.A., Rojas, T.C., Montelli, A.C., Sadatsune, T., de Carvalho Ramos, M., Nolan, L.K. and da Silveira, W.D., 2014. Overlapped sequence types (STs) and serogroups of avian pathogenic (APEC) and human extra-intestinal pathogenic (ExPEC) *Escherichia coli* isolated in Brazil. *PloS One,* 9 (8), e105016.

Manges, A.R., 2016. *Escherichia coli* and urinary tract infections: the role of poultry-meat. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases,* 22 (2), 122-129.

Manges, A.R. and Johnson, J.R., 2012. Food-borne origins of *Escherichia coli* causing extraintestinal infections. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America,* 55 (5), 712-719.

Manges, A.R., Tabor, H., Tellis, P., Vincent, C. and Tellier, P., 2008. Endemic and Epidemic Lineages of *Escherichia coli* that Cause Urinary Tract Infections. *Emerging Infectious Diseases,* 14 (10), 1575-1583.

Mann, S. and Chen, Y.P., 2010. Bacterial genomic G+C composition-eliciting environmental adaptation. *Genomics,* 95 (1), 7-15.

Mansan-Almeida, R., Pereira, A.L. and Giugliano, L.G., 2013. Diffusely adherent *Escherichia coli* strains isolated from children and adults constitute two different populations. *BMC Microbiology,* 13, 22-22.

Marti, E., Variatza, E. and Balcazar, J.L., 2014. The role of aquatic ecosystems as reservoirs of antibiotic resistance. *Trends in Microbiology,* 22 (1), 36-41.

Martins, C.H., Bauab, T.M. and Falcao, D.P., 1998. Characteristics of *Yersinia pseudotuberculosis* isolated from animals in Brazil. *Journal of Applied Microbiology,* 85 (4), 703-707.

Marttinen, P., Hanage, W.P., Croucher, N.J., Connor, T.R., Harris, S.R., Bentley, S.D. and Corander, J., 2012. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Research,* 40, e6.

Mathewson, J.J., Jiang, Z.D., Zumla, A., Chintu, C., Luo, N., Calamari, S.R., Genta, R.M., Steephen, A., Schwartz, P. and DuPont, H.L., 1995. HEp-2 cell-adherent *Escherichia coli* in patients with human immunodeficiency virus-associated diarrhea. *The Journal of Infectious Diseases,* 171 (6), 1636-1639.

Maxam, A.M. and Gilbert, W., 1977. A new method for sequencing DNA. *Proc Natl Acad Sci USA,* 74 (2), 560.

McNally A, Cheng L, Harris SR, Corander J., 2013. The evolutionary path to extra intestinal pathogenic, drug resistant *Escherichia coli* is marked by drastic reduction in detectable recombination within the core genome. *Genome Biology and Evolution,* 5, 699-710.

McNally, A., Oren, Y., Kelly, D., Pascoe, B., Dunn, S., Sreecharan, T., Vehkala, M., Valimaki, N., Prentice, M.B., Ashour, A., Avram, O., Pupko, T., Dobrindt, U., Literak, I., Guenther, S., Schaufler, K., Wieler, L.H., Zhiyong, Z., Sheppard, S.K., McInerney, J.O. and Corander, J., 2016a. Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations. *PLoS Genetics,* 12 (9), e1006280.

McNally, A., Thomson, N.R., Reuter, S. and Wren, B.W., 2016b. 'Add, stir and reduce': *Yersinia* spp. as model bacteria for pathogen evolution. *Nature Reviews Microbiology,* 14 (3), 177-190.

Memariani, M., Najar Peerayeh, S., Zahraei Salehi, T. and Shokouhi Mostafavi, S.K., 2015. Occurrence of SHV, TEM and CTX-M beta-Lactamase Genes Among Enteropathogenic *Escherichia coli* Strains Isolated From Children With Diarrhea. *Jundishapur Journal of Microbiology,* 8 (4), e15620.

Merhej, V., Adekambi, T., Pagnier, I., Raoult, D. and Drancourt, M., 2008. *Yersinia massiliensis* sp. nov., isolated from fresh water. *International Journal of Systematic and Evolutionary Microbiology,* 58 (Pt 4), 779-784.

Metzker, M.L., 2010. Sequencing technologies - the next generation. *Nature Reviews Genetics,* 11 (1), 31-46.

Milkman, R., 1973. Electrophoretic Variation in *Escherichia coli* from natural sources. *Science,* 182, 1024-1026.

Moquet, O., Bouchiat, C., Kinana, A., Seck, A., Arouna, O., Bercion, R., Breurec, S. and Garin, B., 2011. Class D OXA-48 Carbapenemase in Multidrug-Resistant Enterobacteria, Senegal. *Emerging Infectious Diseases,* 17 (1), 143-144.

Mora, A., Lopez, C., Dabhi, G., Blanco, M., Blanco, J.E., Alonso, M.P., Herrera, A., Mamani, R., Bonacorsi, S., Moulin-Schouleur, M. and Blanco, J., 2009. Extraintestinal pathogenic *Escherichia coli* O1:K1:H7/NM from human and avian origin: detection of clonal groups B2 ST95 and D ST59 with different host distribution. *BMC Microbiology,* 9, 132-2180-9-132.

Moran, R.A., Anantham, S., Holt, K.E. and Hall, R.M., 2017. Prediction of antibiotic resistance from antibiotic resistance genes detected in antibiotic-resistant commensal *Escherichia coli* using PCR or WGS. *The Journal of Antimicrobial Chemotherapy,* 72 (3), 700-704.

Morelli, G., Song, Y., Mazzoni, C.J., Eppinger, M., Roumagnac, P., Wagner, D.M., Feldkamp, M., Kusecek, B., Vogler, A.J., Li, Y., Cui, Y., Thomson, N.R., Jombart, T., Leblois, R., Lichtner, P., Rahalison, L., Petersen, J.M., Balloux, F., Keim, P., Wirth, T., Ravel, J., Yang, R., Carniel, E. and Achtman, M., 2010. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genetics,* 42 (12), 1140-1143.

Moriarty, E.M., Mackenzie, M.L., Karki, N. and Sinton, L.W., 2010. Survival of *Escherichia coli*, *Enterococci*, and *Campylobacter* spp. in Sheep Feces on Pastures. *Applied and Environmental Microbiology,* 77 (5), 1797-1803.

Mosquito, S., Pons, M.J., Riveros, M., Ruiz, J. and Ochoa, T.J., 2015. Diarrheagenic *Escherichia coli* Phylogroups Are Associated with Antibiotic Resistance and Duration of Diarrheal Episode. *The Scientific World Journal,* 2015, 610403.

Moulin-Schouleur, M., Schouler, C., Tailliez, P., Kao, M.R., Bree, A., Germon, P., Oswald, E., Mainil, J., Blanco, M. and Blanco, J., 2006. Common virulence factors and genetic relationships between O18:K1:H7 *Escherichia coli* isolates of human and avian origin. *Journal of Clinical Microbiology,* 44 (10), 3484-3492.

Muller, A., Stephan, R. and Nuesch-Inderbinen, M., 2016. Distribution of virulence factors in ESBL-producing *Escherichia coli* isolated from the environment, livestock, food and humans. *The Science of the Total Environment,* 541, 667-672.

Murros-Kontiainen, A., Fredriksson-Ahomaa, M., Korkeala, H., Johansson, P., Rahkila, R. and Björkroth, J., 2011a. *Yersinia nurmii* sp. nov. *Int J Syst Evol Micro,* 61, 2368-72.

Murros-Kontiainen, A., Johansson, P., Niskanen, T., Fredriksson-Ahomaa, M., Korkeala, H. and Björkroth, J., 2011b. *Yersinia pekkanenii* sp. nov**.** *Int J Syst Evol Microbiol,* 61, 2363-7.

Nataro, J.P. and Kaper, J.B., 1998. Diarrheagenic *Escherichia coli*. *Clinical Microbiology Reviews,* 11 (1), 142-201.

Nataro, J.P., Kaper, J.B., Robins-Browne, R., Prado, V., Vial, P. and Levine, M.M., 1987. Patterns of adherence of diarrheagenic *Escherichia coli* to HEp-2 cells. *The Pediatric Infectious Disease Journal,* 6 (9), 829-831.

Nelson, K., Wang, F.S., Boyd, E.F. and Selander, R.K., 1997. Size and sequence polymorphism in the isocitrate dehydrogenase kinase/phosphatase gene (aceK) and flanking regions in *Salmonella enterica* and *Escherichia coli*. *Genetics,* 147 (4), 1509-1520.

Nemergut, D.R., Schmidt, S.K., Fukami, T., O'Neill, S.P., Bilinski, T.M., Stanish, L.F., Knelman, J.E., Darcy, J.L., Lynch, R.C., Wickey, P. and Ferrenberg, S., 2013. Patterns and Processes of Microbial Community Assembly. *Microbiology and Molecular Biology Reviews: MMBR*, 77 (3), 342-356.

Nicolas-Chanoine, M.H., Bertrand, X. and Madec, J.Y., 2014. *Escherichia coli* ST131, an intriguing clonal group. *Clinical Microbiology Reviews,* 27 (3), 543-574.

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

Niskanen, T., Fredriksson-Ahomaa, M. and Korkeala, H., 2002. *Yersinia pseudotuberculosis* with limited genetic diversity is a common finding in tonsils of fattening pigs. *Journal of Food Protection,* 65 (3), 540-545.

Nowgesic, E., Fyfe, M., Hockin, J., King, A., Ng, H., Paccagnella, A., Trinidad, A., Wilcott, L., Smith, R., Denney, A., Struck, L., Embree, G., Higo, K., Chan, J.I., Markey, P., Martin, S. and Bush, D., 1999. Outbreak of *Yersinia pseudotuberculosis* in British Columbia--November 1998. *Canada Communicable Disease Report = Releve Des Maladies Transmissibles Au Canada,* 25 (11), 97-100.

Nuorti, J.P., Niskanen, T., Hallanvuo, S., Mikkola, J., Kela, E., Hatakka, M., Fredriksson-Ahomaa, M., Lyytikainen, O., Siitonen, A., Korkeala, H. and Ruutu, P., 2004. A widespread outbreak of *Yersinia pseudotuberculosis* O:3 infection from iceberg lettuce. *The Journal of Infectious Diseases,* 189 (5), 766-774.

Ochman, H. and Selander, R.K., 1984. Standard reference strains of *Escherichia coli* from natural populations. *Journal of Bacteriology,* 157 (2), 690-693.

Ojer-Usoz, E., Gonzã¡lez, D. and Vitas, A.I., 2017. Clonal Diversity of ESBL-Producing *Escherichia coli* Isolated from Environmental, Human and Food Samples. *International Journal of Environmental Research and Public Health,* 14 (7), 676.

Oliveira, M., Vinas, I., Usall, J., Anguera, M. and Abadias, M., 2012. Presence and survival of *Escherichia coli* O157:H7 on lettuce leaves and in soil treated with contaminated compost and irrigation water. *International Journal of Food Microbiology,* 156 (2), 133-140.

Oteo, J., Cercenado, E., Cuevas, O., Bautista, V., Delgado-Iribarren, A., Orden, B., Perez-Vazquez, M., Garcia-Cobos, S. and Campos, J., 2010. AmpC beta-lactamases in *Escherichia coli*: emergence of CMY-2-producing virulent phylogroup D isolates belonging mainly to STs 57, 115, 354, 393, and 420, and phylogroup B2 isolates belonging to the international clone O25b-ST131. *Diagnostic Microbiology and Infectious Disease,* 67 (3), 270-276.

Overdevest, I., Willemsen, I., Rijnsburger, M., Eustace, A., Xu, L., Hawkey, P., Heck, M., Savelkoul, P., Vandenbroucke-Grauls, C., van der Zwaluw, K., Huijsdens, X. and Kluytmans, J., 2011. Extended-spectrum beta-lactamase genes of *Escherichia coli* in chicken meat and humans, The Netherlands. *Emerging Infectious Diseases,* 17 (7), 1216-1222.

Owens, R.C., Jr, Johnson, J.R., Stogsdill, P., Yarmus, L., Lolans, K. and Quinn, J., 2011. Community transmission in the United States of a CTX-M-15-producing sequence type ST131 *Escherichia coli* strain resulting in death. *Journal of Clinical Microbiology,* 49 (9), 3406-3408.

Paff, J.R., Triplett, D.A. and Saari, T.N., 1976. Clinical and laboratory aspects of *Yersinia pseudotuberculosis* infections, with a report of two cases. *American Journal of Clinical Pathology,* 66 (1), 101-110.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T., Fookes, M., Falush, D., Keane, J.A. and Parkhill, J., 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics (Oxford, England),* 31 (22), 3691-3693.

Paterson, D.L. and Bonomo, R.A., 2005. Extended-spectrum beta-lactamases: a clinical update. *Clinical Microbiology Reviews,* 18 (4), 657-686.

Paterson, D.L., Hujer, K.M., Hujer, A.M., Yeiser, B., Bonomo, M.D., Rice, L.B., Bonomo, R.A. and International *Klebsiella* Study Group, 2003. Extended-spectrum beta-lactamases in *Klebsiella pneumoniae* bloodstream isolates from seven countries: dominance and widespread prevalence

of SHV- and CTX-M-type beta-lactamases. *Antimicrobial Agents and Chemotherapy,* 47 (11), 3554-3560.

Pavlickova, S., Dolezalova, M. and Holko, I., 2015. Resistance and virulence factors of *Escherichia coli* isolated from chicken. *Journal of Environmental Science and Health, Part B. Pesticides, Food Contaminants, and Agricultural Wastes,* 50 (6), 417-421.

Peirano, G., van der Bij, A.K., Gregson, D.B. and Pitout, J.D., 2012. Molecular epidemiology over an 11-year period (2000 to 2010) of extended-spectrum β-lactamase-producing *Escherichia coli* causing bacteremia in a centralized Canadian region. *Journal of Clinical Microbiology,* 50, 294-9.

Peirano, G. and Pitout, J.D.D., 2010. Molecular epidemiology of *Escherichia coli* producing CTX-M β-lactamases: the worldwide emergence of clone ST131 O25:H4. *International Journal of Antimicrobial Agents,* 35 (4), 316-321.

Pepe, J.C. and Miller, V.L., 1993. *Yersinia enterocolitica* invasin: A primary role in the initiation of infection. *PNAS,* 90 (14), 6473-6477.

Perreten, V. and Boerlin, P., 2003. A new sulfonamide resistance gene (sul3) in *Escherichia coli* is widespread in the pig population of Switzerland. *Antimicrobial Agents and Chemotherapy,* 47 (3), 1169-1172.

Pessia, A., Grad, Y., Cobey, S., Puranen, J.S. and Corander, J., 2015. K-Pax2: Bayesian identification of cluster-defining amino acid positions in large sequence datasets. *Microbial Genomics,* 1 (1), e000025.

Petty, N.K., Ben Zakour, N.L., Stanton-Cook, M., Skippington, E., Totsika, M., Forde, B.M., Phan, M.D., Gomes Moriel, D., Peters, K.M., Davies, M., Rogers, B.A., Dougan, G., Rodriguez-Bano, J., Pascual, A., Pitout, J.D., Upton, M., Paterson, D.L., Walsh, T.R., Schembri, M.A. and Beatson, S.A., 2014. Global dissemination of a multidrug resistant *Escherichia coli* clone. *Proceedings of the National Academy of Sciences of the United States of America,* 111 (15), 5694-5699.

Picard, B., Garcia, J.S., Gouriou, S., Duriez, P., Brahimi, N., Bingen, E., Elion, J. and Denamur, E., 1999. The Link between Phylogeny and Virulence in *Escherichia coli* Extraintestinal Infection. *Infection and Immunity,* 67 (2), 546-553.

Pitout, J.D., Church, D.L., Gregson, D.B., Chow, B.L., McCracken, M., Mulvey, M.R. and Laupland, K.B., 2007. Molecular epidemiology of CTX-M-producing *Escherichia coli* in the Calgary Health Region: emergence of CTX-M-15-producing isolates. *Antimicrobial Agents and Chemotherapy,* 51 (4), 1281-1286.

Pitout, J.D.D., 2012. Extraintestinal Pathogenic *Escherichia coli*: A Combination of Virulence with Antibiotic Resistance. *Frontiers in Microbiology,* 3 (9).

Pitout, J.D.D., Nordmann, P., Laupland, K.B. and Poirel, L., 2005. Emergence of Enterobacteriaceae producing extended-spectrum β-lactamases (ESBLs) in the community. *Journal of Antimicrobial Chemotherapy,* 56 (1), 52-59.

Platell, J.L., Cobbold, R.N., Johnson, J.R., Heisig, A., Heisig, P., Clabots, C., Kuskowski, M.A. and Trott, D.J., 2011a. Commonality among fluoroquinolone-resistant sequence type ST131 extraintestinal *Escherichia coli* isolates from humans and companion animals in Australia. *Antimicrobial Agents and Chemotherapy,* 55 (8), 3782-3787.

Platell, J.L., Johnson, J.R., Cobbold, R.N. and Trott, D.J., 2011b. Multidrug-resistant extraintestinal pathogenic *Escherichia coli* of sequence type ST131 in animals and foods. *Veterinary Microbiology,* doi:10.1016/j.vetmic.2011.05.007.

Poirel, L., Ros, A., Carrer, A., Fortineau, N., Carricajo, A., Berthelot, P. and Nordmann, P., 2011a. Cross-border transmission of OXA-48-producing *Enterobacter cloacae* from Morocco to France. *The Journal of Antimicrobial Chemotherapy,* 66 (5), 1181-1182.

Poirel, L., Bernabeu, S., Fortineau, N., Podglajen, I., Lawrence, C. and Nordmann, P., 2011b. Emergence of OXA-48-Producing *Escherichia coli* Clone ST38 in France. *Antimicrobial Agents and Chemotherapy,* 55 (10), 4937-4938.

Poirel, L., Héritier, C., Tolün, V. and Nordmann, P., 2004. Emergence of Oxacillinase-Mediated Resistance to Imipenem in *Klebsiella pneumoniae*. *Antimicrobial Agents and Chemotherapy,* 48 (1), 15-22.

Polz, M.F., Alm, E.J. and Hanage, W.P., 2013. Horizontal Gene Transfer and the Evolution of Bacterial and Archaeal Population Structure. *Trends in Genetics: TIG,* 29 (3), 170-175.

Portnoy, D.A. and Falkow, S., 1981. Virulence-associated plasmids from *Yersinia enterocolitica* and *Yersinia pestis*. *Journal of Bacteriology,* 148 (3), 877-883.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Dore, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., MetaHIT Consortium, Bork, P., Ehrlich,

S.D. and Wang, J., 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature,* 464 (7285), 59-65.

Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J. and Segata, N., 2017. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology,* 35 (9), 833-844.

Ramamurthy, T., Yoshino, K., Abe, J., Ikeda, N. and Takeda, T., 1997. Purification, characterization and cloning of a novel variant of the superantigen *Yersinia pseudotuberculosis*-derived mitogen. *FEBS Letters,* 413 (1), 174-176.

Rasko, D.A., Rosovitz, M.J., Myers, G.S.A., Mongodin, E.F., Fricke, W.F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N.R., Chaudhuri, R., Henderson, I.R., Sperandio, V. and Ravel, J., 2008. The Pangenome Structure of *Escherichia coli*: Comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *Journal of Bacteriology,* 190 (20), 6881-6893.

Ratiner, Y.A., 1985. Two genetic arrangements determining flagellar antigen specificities in two diphasic *Escherichia coli* strains. *FEMS Microbiology Letters,* 29 (3), 317-323.

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: www.R-project.org.

Reuter, S., Connor, T., Barquist, L., Walker, D., Feltwell, T., Harris, S., Fookes, M., Hall, M., Petty, N., Fuchs, T., Corander, J., Dufour, M., Ringwood, T., Savin, C., Bouchier, C., Martin, L., Miettinen, M., Shubin, M., Riehm, J., Laukkanen-Ninios, R., Sihvonen, L., Siitonen, A., Skurnik, M., Falcão, J., Fukushima, H., Scholz, H., Prentice, M., Wren, B., Parkhill, J., Carniel, E., Achtman, M., McNally, A. and Thomson, N., 2014. Parallel independent evolution of pathogenicity within the genus *Yersinia*. *Proc Natl Acad Sci U S A,* 111, 6768-73.

Reuter, S., Corander, J., de Been, M., Harris, S., Cheng, L., Hall, M., Thomson, N.R. and McNally, A., 2015. Directional gene flow and ecological separation in *Yersinia enterocolitica*. *Microbial Genomics,* 1 (3), e000030.

Riesenfeld, C.S., Schloss, P.D. and Handelsman, J., 2004. Metagenomics: genomic analysis of microbial communities. *Annual Review of Genetics,* 38, 525-552.

Riley, L.W., 2014. Pandemic lineages of extraintestinal pathogenic *Escherichia coli*. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases,* 20 (5), 380-390.

Riley, L.W., Remis, R.S., Helgerson, S.D., McGee, H.B., Wells, J.G., Davis, B.R., Hebert, R.J., Olcott, E.S., Johnson, L.M., Hargrett, N.T., Blake, P.A. and Cohen, M.L., 1983. Hemorrhagic colitis

associated with a rare *Escherichia coli* serotype. *The New England Journal of Medicine,* 308 (12), 681-685.

Ron, E.Z., 2010. Distribution and evolution of virulence factors in septicemic *Escherichia coli*. *International Journal of Medical Microbiology: IJMM,* 300 (6), 367-370.

Russo, T.A. and Johnson, J.R., 2003. Medical and economic impact of extraintestinal infections due to *Escherichia coli*: focus on an increasingly important endemic problem. *Microbes and Infection,* 5 (5), 449-456.

Sahl, J.W., Caporaso, J.G., Rasko, D.A. and Keim, P., 2014. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. *PeerJ,* 2, e332.

Sanger, F. and Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology,* 94 (3), 441-448.

Sato, T., Yokota, S., Okubo, T., Usui, M., Fujii, N. and Tamura, Y., 2014. Phylogenetic association of fluoroquinolone and cephalosporin resistance of D-O1-ST648 *Escherichia coli* carrying blaCMY-2 from faecal samples of dogs in Japan. *Journal of Medical Microbiology,* 63 (Pt 2), 263-270.

Savageau, M.A., 1983. *Escherichia coli* Habitats, Cell Types, and Molecular Mechanisms of Gene Control. *The American Naturalist,* 122 (6), 732-744.

Savin, C., Martin, L., Bouchier, C., Filali, S., Chenau, J., Zhou, Z., Becher, F., Fukushima, H., Thomson, N.R., Scholz, H.C. and Carniel, E., 2014. The *Yersinia pseudotuberculosis* complex: characterization and delineation of a new species, *Yersinia wautersii*. *International Journal of Medical Microbiology: IJMM,* 304 (3-4), 452-463.

Schadt, E.E., Turner, S. and Kasarskis, A., 2010. A window into third-generation sequencing. *Human Molecular Genetics,* 19 (R2), R227-R240.

Schatz, M.C., Delcher, A.L. and Salzberg, S.L., 2010. Assembly of large genomes using second-generation sequencing. *Genome Research,* 20 (9), 1165-1173.

Schiano, C.A., Bellows, L.E. and Lathem, W.W., 2010. The small RNA chaperone Hfq is required for the virulence of *Yersinia pseudotuberculosis*. *Infection and Immunity,* 78 (5), 2034-2044.

Seecharran, T., Kalin-Manttari, L., Koskela, K., Nikkari, S., Dickins, B., Corander, J., Skurnik, M. and McNally, A., 2017. Phylogeographic separation and formation of sexually discrete lineages in a global population of *Yersinia pseudotuberculosis*. *Microbial Genomics,* 3 (10), e000133.

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics,* 10.1093/bioinformatics/btu153.

Seputiene, V., Povilonis, J., Ruzauskas, M., Pavilonis, A. and Suziedeliene, E., 2010. Prevalence of trimethoprim resistance genes in *Escherichia coli* isolates of human and animal origin in Lithuania. *Journal of Medical Microbiology,* 59 (Pt 3), 315-322.

Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabó, G., Polz, M.F. and Alm, E.J., 2012. Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science (New York, N.Y.),* 336 (6077), 48-51.

Sheppard, S., Cheng, L., Méric, G., de Haan, C., Llarena, A., Marttinen, P., Vidal, A., Ridley, A., Clifton-Hadley, F., Connor, T., Strachan, N., Forbes, K., Colles, F., Jolley, K., Bentley, S., Maiden, M., Hänninen, M., Parkhill, J., Hanage, W. and Corander, J., 2014. Cryptic ecology among host generalist *Campylobacter jejuni* in domestic animals. *Mol.Ecol.,* 23, 2442-51.

Sheppard, S.K., McCarthy, N.D., Falush, D. and Maiden, M.C., 2008. Convergence of *Campylobacter* species: implications for bacterial evolution. *Science (New York, N.Y.),* 320 (5873), 237-239.

Sidjabat, H.E., Paterson, D.L., Adams-Haduch, J.M., Ewan, L., Pasculle, A.W., Muto, C.A., Tian, G.B. and Doi, Y., 2009. Molecular epidemiology of CTX-M-producing *Escherichia coli* isolates at a tertiary medical center in western Pennsylvania. *Antimicrobial Agents and Chemotherapy,* 53 (11), 4733-4739.

Singer, R.S., 2015. Urinary tract infections attributed to diverse ExPEC strains in food animals: evidence and data gaps. *Frontiers in Microbiology,* 6, 28.

Skurnik, D., Clermont, O., Guillard, T., Launay, A., Danilchanka, O., Pons, S., Diancourt, L., Lebreton, F., Kadlec, K., Roux, D., Jiang, D., Dion, S., Aschard, H., Denamur, M., Cywes-Bentley, C., Schwarz, S., Tenaillon, O., Andremont, A., Picard, B., Mekalanos, J., Brisse, S. and Denamur, E., 2016. Emergence of Antimicrobial-Resistant *Escherichia coli* of Animal Origin Spreading in Humans. *Molecular Biology and Evolution,* 33 (4), 898-914.

Skurnik, M., Peippo, A. and Ervela, E., 2000. Characterization of the O-antigen gene clusters of *Yersinia pseudotuberculosis* and the cryptic O-antigen gene cluster of *Yersinia pestis* shows that

the plague bacillus is most closely related to and has evolved from *Y. pseudotuberculosis* serotype O:1b. *Mol Microbiol,* 37 (2), 316-30.

Sloan, W.T., Lunn, M., Woodcock, S., Head, I.M., Nee, S. and Curtis, T.P., 2006. Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environmental Microbiology*, 8 (4), 732-740.

Smith, J.L., Fratamico, P.M. and Gunther, N.W., 2007. Extraintestinal pathogenic *Escherichia coli*. *Foodborne Pathogens and Disease,* 4 (2), 134-163.

Sprague, L.D., and Neubauer, H., 2005. *Yersinia aleksiciae* sp. nov. *International Journal of Systematic and Evolutionary Microbiology,* 55 (Pt 2), 831-835.

Sprague, L.D., Scholz, H.C., Amann, S., Busse, H.J. and Neubauer, H., 2008. *Yersinia similis* sp. nov. *International Journal of Systematic and Evolutionary Microbiology,* 58 (Pt 4), 952-958.

Stine, O.C., Sozhamannan, S., Gou, Q., Zheng, S., Morris, J.G., and Johnson, J.A., 2000. Phylogeny of *Vibrio cholerae* Based on *recA* Sequence. *Infection and Immunity*, 68 (12), 7180-7185.

Su, H.C., Ying, G.G., Tao, R., Zhang, R.Q., Zhao, J.L. and Liu, Y.S., 2012. Class 1 and 2 integrons, sul resistance genes and antibiotic resistance in *Escherichia coli* isolated from Dongjiang River, South China. *Environmental Pollution (Barking, Essex: 1987),* 169, 42-49.

Sulakvelidze, A., 2000. *Yersinia*e other than *Y. enterocolitica*, *Y. pseudotuberculosis*, and *Y. pestis*: the ignored species. *Microbes and Infection,* 2 (5), 497-513.

Sunahara, C., Yamanaka, Y. and Yamanishi, S., 2000. Sporadic cases of *Yersinia pseudotuberculosis* serotype 5 infection in Shodo Island, Kagawa Prefecture. *Japanese Journal of Infectious Diseases,* 53 (2), 74-75.

Tartof, S.Y., Solberg, O.D., Manges, A.R. and Riley, L.W., 2005. Analysis of a uropathogenic *Escherichia coli* clonal group by multilocus sequence typing. *Journal of Clinical Microbiology,* 43 (12), 5860-5864.

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

Terlizzi, M.E., Gribaudo, G. and Maffei, M.E., 2017. UroPathogenic *Escherichia coli* (UPEC) Infections: Virulence Factors, Bladder Responses, Antibiotic, and Non-antibiotic Antimicrobial Strategies. *Frontiers in Microbiology,* 8, 1566.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R. and Fraser, C.M., 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America,* 102 (39), 13950-13955.

Tettelin, H., Riley, D., Cattuto, C. and Medini, D., 2008. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology,* 11 (5), 472-477.

Tivendale, K.A., Logue, C.M., Kariyawasam, S., Jordan, D., Hussein, A., Li, G., Wannemuehler, Y.M. and Nolan, L.K., 2010. Avian-Pathogenic *Escherichia coli* Strains Are Similar to Neonatal Meningitis *E. coli* Strains and Are Able To Cause Meningitis in the Rat Model of Human Disease. *Infection and Immunity,* 78 (8), 3412-3412-3419.

Trabulsi, L.R., Keller, R. and Tardelli Gomes, T.A., 2002. Typical and atypical enteropathogenic *Escherichia coli*. *Emerging Infectious Diseases,* 8 (5), 508-513.

Treangen, T.J., Ondov, B.D., Koren, S. and Phillippy, A.M., 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology,* 15 (11), 524.

Tseneva, G.Y., Chesnokova, M.V., Timofeevich, K.V., Aleksandrovna, V.E., Burgasova, O.A., Sayapina, L.V., Aleksandrovna, T.K. and Karimova, T.V., 2012. *Pseudotuberculosis* in the Russian federation. *Advances in Experimental Medicine and Biology,* 954, 63-68.

Tumbarello, M., Sanguinetti, M., Montuori, E., Trecarichi, E.M., Posteraro, B., Fiori, B., Citton, R., D'Inzeo, T., Fadda, G., Cauda, R. and Spanu, T., 2007. Predictors of mortality in patients with bloodstream infections caused by extended-spectrum-beta-lactamase-producing Enterobacteriaceae: importance of inadequate initial antimicrobial treatment. *Antimicrobial Agents and Chemotherapy,* 51 (6), 1987-1994.

Tzouvelekis, L.S., Tzelepi, E., Tassios, P.T. and Legakis, N.J., 2000. CTX-M-type beta-lactamases: an emerging group of extended-spectrum enzymes. *International Journal of Antimicrobial Agents,* 14 (2), 137-142.

Ueshiba, H., Kato, H., Miyoshi-Akiyama, T., Tsubokura, M., Nagano, T., Kaneko, S. and Uchiyama, T., 1998. Analysis of the superantigen-producing ability of *Yersinia pseudotuberculosis* strains of various serotypes isolated from patients with systemic or gastroenteric infections, wildlife animals and natural environments. *Zentralblatt Fur Bakteriologie: International Journal of Medical Microbiology,* 288 (2), 277-291.

van Belkum, A., Soriaga, L.B., LaFave, M.C., Akella, S., Veyrieras, J.B., Barbu, E.M., Shortridge, D., Blanc, B., Hannum, G., Zambardi, G., Miller, K., Enright, M.C., Mugnier, N., Brami, D., Schicklin, S., Felderman, M., Schwartz, A.S., Richardson, T.H., Peterson, T.C., Hubby, B. and Cady, K.C., 2015. Phylogenetic Distribution of CRISPR-Cas Systems in Antibiotic-Resistant *Pseudomonas aeruginosa*. *MBio,* 6 (6), e01796-15.

van Elsas, J.D., Semenov, A.V., Costa, R. and Trevors, J.T., 2011. Survival of *Escherichia coli* in the environment: fundamental and public health aspects. *The ISME Journal,* 5 (2), 173-183.

van Overbeek, L.S., Wellington, E.M., Egan, S., Smalla, K., Heuer, H., Collard, J.M., Guillaume, G., Karagouni, A.D., Nikolakopoulou, T.L. and van Elsas, J.D., 2002. Prevalence of streptomycin-resistance genes in bacterial populations in European habitats. *FEMS Microbiology Ecology,* 42 (2), 277-288.

Vejborg, R.M., Hancock, V., Schembri, M.A. and Klemm, P., 2011. Comparative Genomics of *Escherichia coli* Strains Causing Urinary Tract Infections. *Applied and Environmental Microbiology,* 77 (10), 3268-3278.

Vincent, C., Boerlin, P., Daignault, D., Dozois, C.M., Dutil, L., Galanakis, C., Reid-Smith, R.J., Tellier, P.P., Tellis, P.A., Ziebell, K. and Manges, A.R., 2010. Food reservoir for *Escherichia coli* causing urinary tract infections. *Emerging Infectious Diseases,* 16 (1), 88-95.

Virtanen, S., Laukkanen-Ninios, R., Ortiz Martinez, P., Siitonen, A., Fredriksson-Ahomaa, M. and Korkeala, H., 2013. Multiple-locus variable-number tandem-repeat analysis in genotyping *Yersinia enterocolitica* strains from human and porcine origins. *Journal of Clinical Microbiology,* 51 (7), 2154-2159.

Walk, S.T., Alm, E.W., Gordon, D.M., Ram, J.L., Toranzos, G.A., Tiedje, J.M. and Whittam, T.S., 2009. Cryptic lineages of the genus *Escherichia*. *Applied and Environmental Microbiology,* 75 (20), 6534-6544.

Wang, X., Li, Y., Jing, H., Ren, Y., Zhou, Z., Wang, S., Kan, B., Xu, J. and Wang, L., 2011. Complete genome sequence of a *Yersinia enterocolitica* "Old World" (3/O:9) strain and comparison with the "New World" (1B/O:8) strain. *Journal of Clinical Microbiology,* 49 (4), 1251-1259.

Watt, S., Lanotte, P., Mereghetti, L., Moulin-Schouleur, M., Picard, B. and Quentin, R., 2003. *Escherichia coli* strains from pregnant women and neonates: intraspecies genetic distribution and prevalence of virulence factors. *Journal of Clinical Microbiology,* 41 (5), 1929-1935.

Weinert, L.A. and Welch, J.J., 2017. Why Might Bacterial Pathogens Have Small Genomes? *Trends in Ecology & Evolution,* 32 (12), 936-947.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham, H. (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse.

Willems, R.J., Top, J., van Schaik, W., Leavis, H., Bonten, M., Siren, J., Hanage, W.P. and Corander, J., 2012. Restricted gene flow among hospital subpopulations of *Enterococcus faecium*. *MBio,* 3 (4), e00151-12.

Williamson, D.A., Baines, S.L., Carter, G.P., da Silva, A.G., Ren, X., Sherwood, J., Dufour, M., Schultz, M.B., French, N.P., Seemann, T., Stinear, T.P. and Howden, B.P., 2016. Genomic Insights into a Sustained National Outbreak of *Yersinia pseudotuberculosis*. *Genome Biology and Evolution,* 8 (12), 3806-3814.

Winn, W.J., Allen, S., Janda, W., Koneman, E., Procop, G., Schreckenberger, P.C. and Woods, G., 2006. *Koneman's Color Atlas and Textbook of Diagnostic Microbiology.* 6th edition ed. USA: Lippincott, Williams & Wilkins.

Winokur, P.L., Vonstein, D.L., Hoffman, L.J., Uhlenhopp, E.K. and Doern, G.V., 2001. Evidence for transfer of CMY-2 AmpC beta-lactamase plasmids between *Escherichia coli* and *Salmonella* isolates from food animals and humans. *Antimicrobial Agents and Chemotherapy,* 45 (10), 2716-2722.

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L.H., Karch, H., Reeves, P.R., Maiden, M.C.J., Ochman, H. and Achtman, M., 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Molecular Microbiology,* 60 (5), 1136-1151.

Woodcock, S., van der Gast, C.J., Bell, T., Lunn, M., Curtis, T.P., Head, I.M. and Sloan, W.T., 2007. Neutral assembly of bacterial communities. *FEMS Microbiology Ecology*, 62 (2), 171-180.

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

Woodford, N., Ward, M.E., Kaufmann, M.E., Turton, J., Fagan, E.J., James, D., Johnson, A.P., Pike, R., Warner, M., Cheasty, T., Pearson, A., Harry, S., Leach, J.B., Loughrey, A., Lowes, J.A., Warren, R.E. and Livermore, D.M., 2004. Community and hospital spread of *Escherichia coli* producing CTX-M extended-spectrum beta-lactamases in the UK. *Journal of Antimicrobial Chemotherapy,* 54, 735-743.

Wren, B.W., 2003. The *Yersiniae* - a model genus to study the rapid evolution of bacterial pathogens. *Nature Reviews Microbiology,* 1 (1), 55-64.

Xu, J., 2006. Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Molecular Ecology,* 15 (7), 1713-1731.

Yang, C., Li, P., Su, W., Li, H., Liu, H., Yang, G., Xie, J., Yi, S., Wang, J., Cui, X., Wu, Z., Wang, L., Hao, R., Jia, L., Qiu, S. and Song, H., 2015. Polymorphism of CRISPR shows separated natural groupings of *Shigella* subtypes and evidence of horizontal transfer of CRISPR. *RNA Biology,* 12 (10), 1109-1120.

Yoshino, K., Ramamurthy, T., Nair, G.B., Fukushima, H., Ohtomo, Y., Takeda, N., Kaneko, S. and Takeda, T., 1995. Geographical heterogeneity between Far East and Europe in prevalence of *ypm* gene encoding the novel superantigen among *Yersinia pseudotuberculosis* strains. *Journal of Clinical Microbiology,* 33 (12), 3356-3358.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F.M. and Larsen, M.V., 2012. Identification of acquired antimicrobial resistance genes. *The Journal of Antimicrobial Chemotherapy,* 67 (11), 2640-2644.

Zerbino, D.R. and Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research,* 18 (5), 821-829.

Zhang, H., Gao, Y. and Chang, W., 2016. Comparison of Extended-Spectrum β-Lactamase-Producing *Escherichia coli* Isolates from Drinking Well Water and Pit Latrine Wastewater in a Rural Area of China. *BioMed Research International,* 2016, 4343564.

Zhao, L., Gao, S., Huan, H., Xu, X., Zhu, X., Yang, W., Gao, Q. and Liu, X., 2009. Comparison of virulence factors and expression of specific genes between uropathogenic *Escherichia coli* and avian pathogenic *E. coli* in a murine urinary tract infection model and a chicken challenge model. *Microbiology (Reading, England),* 155 (Pt 5), 1634-1644.

Zhao, S., White, D.G., McDermott, P.F., Friedman, S., English, L., Ayers, S., Meng, J., Maurer, J.J., Holland, R. and Walker, R.D., 2001. Identification and Expression of Cephamycinase bla(CMY)

Genes in *Escherichia coli* and *Salmonella* Isolates from Food Animals and Ground Meat. *Antimicrobial Agents and Chemotherapy,* 45 (12), 3647-3650.

Zong, Z. and Yu, R., 2010. *Escherichia coli* carrying the blaCTX-M-15 gene of ST648. *Journal of Medical Microbiology,* 59 (12), 1536-1537.

Appendix

Appendix 1. Full printed copy of the published journal article containing information and data from chapter 3 (provided in pocket at the end of the thesis):

Title:

Phylogeographic separation of sexually discrete lineages in a global population of *Yersinia pseudotuberculosis*

Authors:

Seecharran T., Kalin-Manttari L., Koskela K., Nikkari S., Dickins B., Corander J., Skurnik M., McNally A.

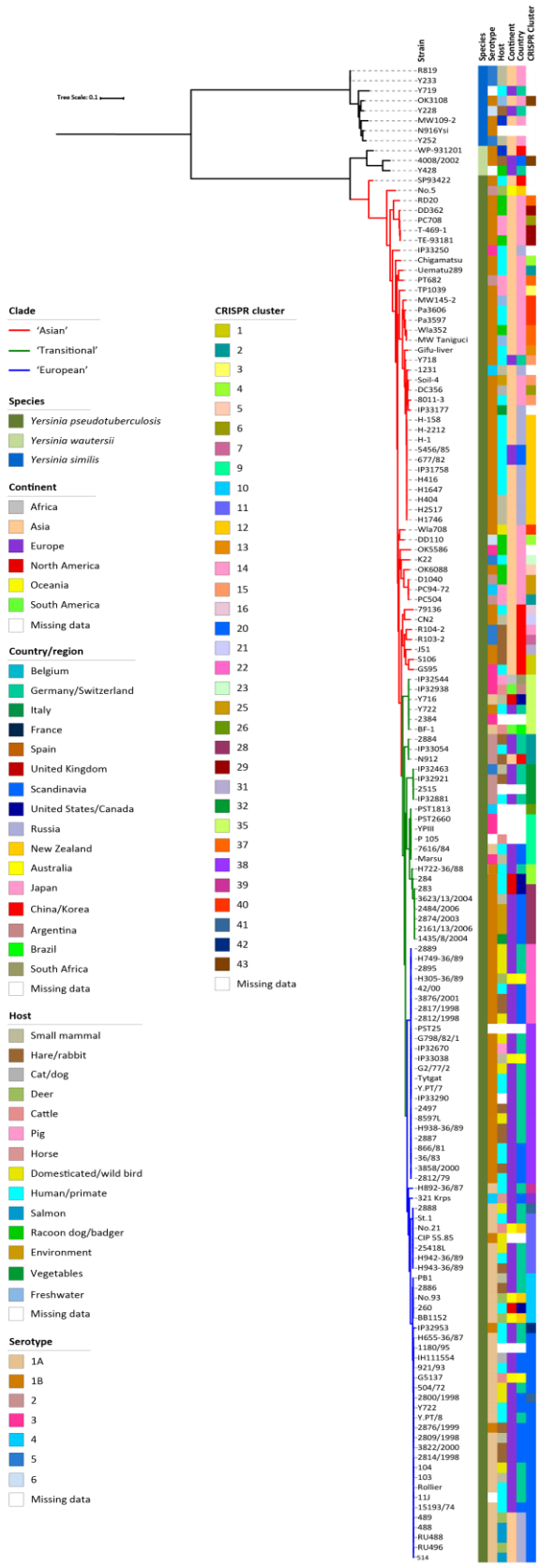
Journal:

Microbial Genomics. 2017;3 (10): e000133. http://doi.org/10.1099/mgen.0.000133

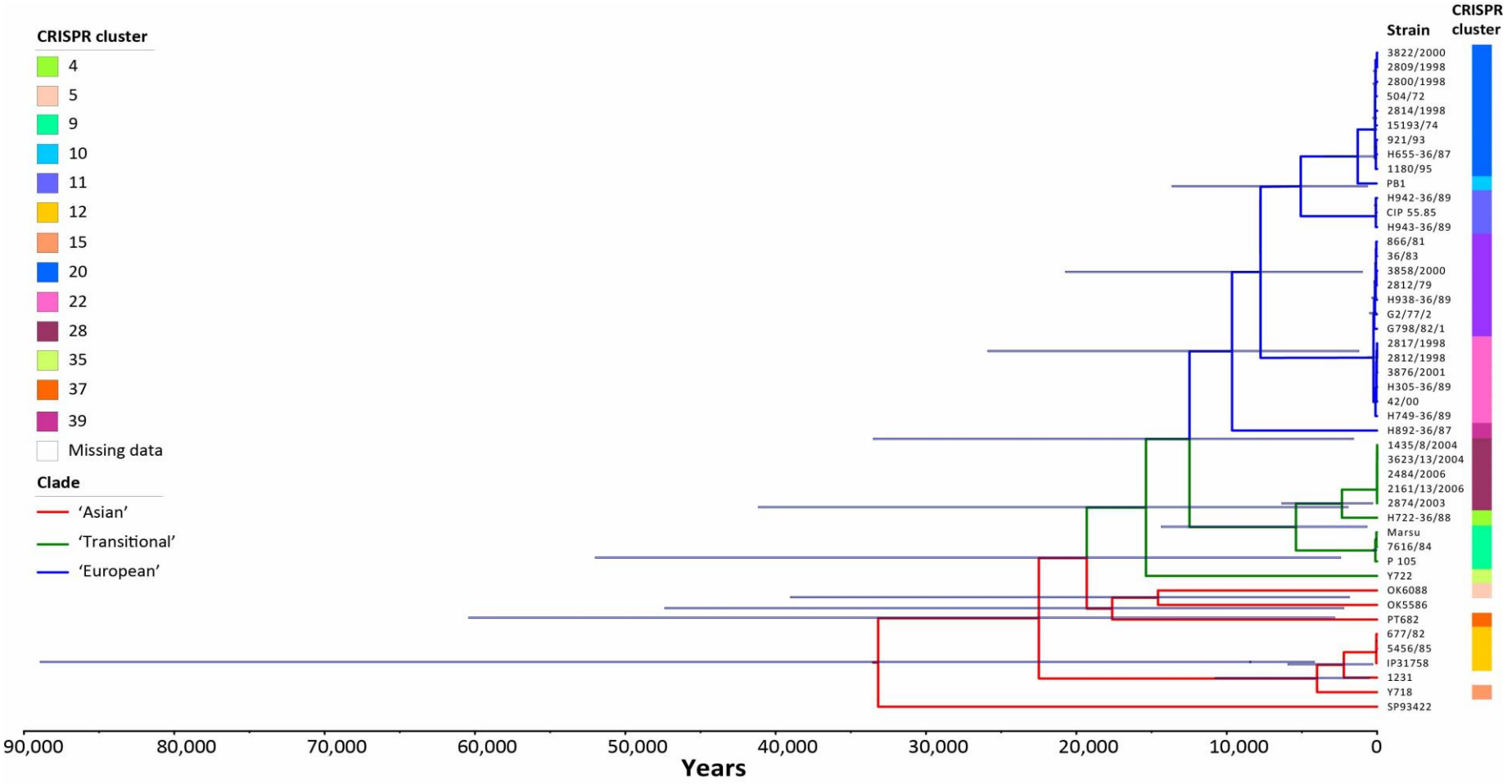
Appendix 2. *Y. similis* and *Y. wautersii* genomes used for the outgroup in Appendix 3.

Strain name	Species	Year	ST	Host	Continent	Country	Accession number
R819	*Y. similis*	1988	75	Mouse	Asia	Japan	ERR024895
Y233	*Y. similis*	1991	75	Mole	Asia	Japan	ERR024899
Y719	*Y. similis*	1991	75	Human	Europe	Germany	ERR024904
OK3108	*Y. similis*	-	-	River	Asia	Japan	ERR1448034
Y228	*Y. similis*	1990	92	Rabbit	Europe	Germany	ERR024898
MW109-2	*Y. similis*	1990	71	Water	Asia	Japan	ERR024915
N916Ysi	*Y. similis*	-	-	Animal	Europe	Finland	ERR027410
Y252	*Y. similis*	1990	71	Mole	Asia	Japan	ERR024900
WP-931201	*Y. wautersii*	1993	21	Water	Asia	Korea	ERR024893
4008/2002	*Y. wautersii*	2002	-	Hare	Europe	Finland	ERR1448055
Y428	*Y. wautersii*	-	96	Badger	Europe	Germany	ERR024901

Eight *Y. similis* and three *Y. wautersii* genomes were included as an outgroup on the phylogenetic tree of the *Y. pseudotuberculosis* study population (Appendix 3). *De novo* assemblies of *Y. similis* and *Y. wautersii* genomes are available on Enterobase (https://enterobase.warwick.ac.uk/species/index/yersinia), searchable by the strain name or accession number indicated in the table.



Appendix 3. Maximum-likelihood phylogenetic tree of 134 *Y. pseudotuberculosis*, 3 *Y. wautersii*, and 8 *Y. similis* isolates annotated with all available metadata. *Y. wautersii* and *Y. similis* are separate species of the *Y. pseudotuberculosis* complex and were included in the phylogenetic tree as an outgroup. Inclusion of these genomes (accession numbers shown in Appendix 2) indicated that the "Asian" clade represents the ancestral clade for the study population.



Appendix 4. Dated maximum clade credibility (MCC) tree produced from BEAST 2 analysis performed on the 46 *Y. pseudotuberculosis* strains for which isolation dates are available. The tree was visualised and manipulated using FigTree. Error bars are displayed at each node representing the upper and lower values within the 95% HPD (highest probability density) from the BEAST analysis.

Appendix 5. Assembly statistics for the 180 sequenced non-human *E. coli* genomes.

Strain	N50 (bp)	L50 (bp)	Genome size (bp)	GC content (%)	Number of N's per 100 kbp
I2-15	1,107	951	3,359,765	50.28	8.04
I2-8	1,135	894	3,276,310	50.51	15.26
AFR-38	1,226	819	3,334,302	50.49	15.9
I1-29	1,263	829	3,311,016	50.09	25.67
I2-7	1,367	895	3,884,517	50.36	18.74
AFR-26	1,426	826	3,801,326	50.53	18.41
M3-26	1,462	821	3,930,731	50.86	32.49
TFR-7	1,528	722	3,690,087	50.81	25.47
AFR-36	1,555	733	3,827,893	51.31	19.04
I1-6	1,640	726	3,743,495	50.5	28.05
AFR-28	1,814	779	4,627,632	50.32	23.9
AFR-34	1,863	654	4,241,859	50.4	30.36
M3-24	1,939	712	4,514,416	50.91	31.45
M2-8	2,163	610	4,349,702	51.48	33.54
I2-6	2,188	632	4,804,256	50.48	16.86
AFR-4	2,206	625	4,722,035	51.07	35.66
I1-33	2,285	601	4,399,093	50.77	32.05
AFR-12	2,309	599	4,741,109	50.77	27.21
AFR-20	2,403	486	4,467,222	51.03	25.07
M3-31	2,458	500	4,631,908	51	7.77
T3-18	2,512	543	4,616,073	50.92	21.01
AFR-16	2,532	588	4,854,746	50.85	36.97
M3-30	2,628	560	4,902,181	51.18	35.07
AFR-18	2,790	528	4,941,315	50.86	57.84
TFR-6	2,993	454	4,624,797	51.04	30.68
GU1	3,207	394	4,433,701	50.83	2.26
I1-11	3,326	450	5,181,407	50.16	21.62
AFR-22	3,346	465	5,200,256	50.7	50.42
M2-6	3,647	434	5,111,765	50.93	40.59
M2-2	3,696	402	4,905,557	51.02	27.52
AFR-14	3,755	421	5,235,395	50.70	33.73
M3-36	4,047	385	5,328,946	50.85	26.27
M2-3	4,080	398	5,334,950	51.03	41.26
T1-53	4,149	371	5,525,329	50.87	31.4
M3-23	4,231	394	5,697,090	50.42	33.96
AFR-30	4,480	338	5,164,501	50.68	20.14
M2-9	4,785	316	4,990,196	51.37	14.83
T3-21	4,819	317	5,435,770	50.34	25.13
I2-18	5,422	318	5,610,195	50.55	19.59
GU2	6,468	219	4,898,346	50.60	3.06
AFR-10	6,675	254	5,756,792	50.52	20.06
M2-7	6,694	263	5,466,211	50.76	16.28
GD138	7,567	189	4,785,679	50.67	2.93
M2-5	7,946	218	5,606,004	50.60	16.05
I1-5	8,208	196	5,038,692	50.37	13.61
T3-1	8,235	194	5,569,609	50.78	20.47
M3-21	8,262	214	5,808,794	50.75	27.17
I1-24	8,272	201	5,238,932	50.80	13.93

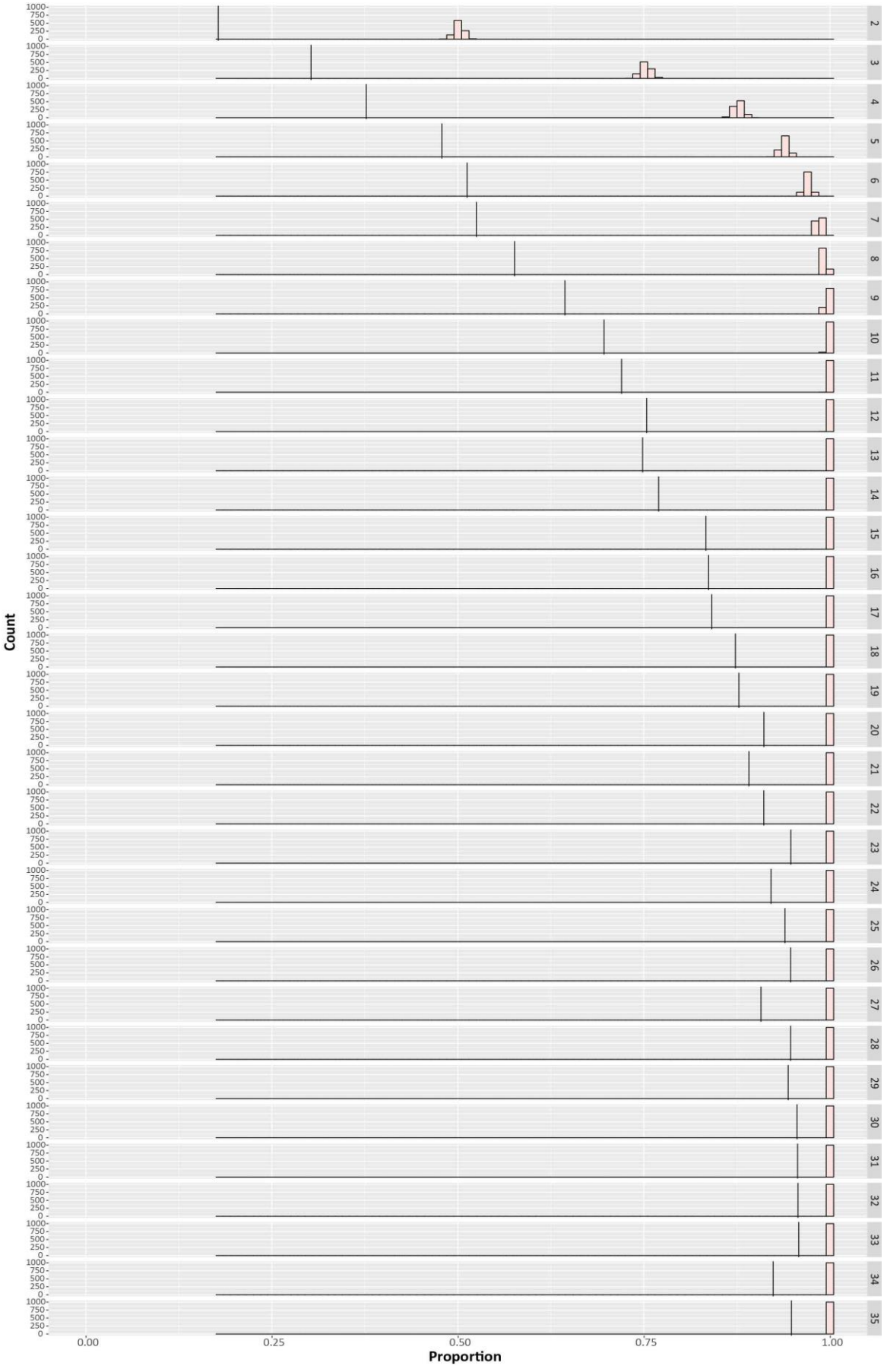
M2-1	8,539	199	5,474,492	50.61	15.11
M3-27	8,644	198	5,585,928	50.77	7.88
I2-20	8,857	201	5,781,737	50.46	13.3
TFR-2	8,869	160	5,265,664	50.72	5.32
M3-29	8,894	158	4,797,545	50.71	13.34
AFR-40	9,378	181	5,598,833	50.87	48.62
M3-22	9,745	175	5,457,271	50.84	16.68
T1-7	10,079	154	4,995,381	50.76	13.41
AFR-32	10,583	151	5,365,379	50.76	15.66
M3-25	10,645	157	5,634,273	50.71	10.65
M3-20	11,200	141	5,593,407	50.59	19.31
I1-25	11,346	160	5,913,968	50.34	41.28
M3-34	11,788	148	5,689,566	50.60	7.73
AFR-8	12,613	131	5,396,156	50.71	8.9
I1-15	12,990	117	5,018,581	50.57	11.34
AFR-24	13,191	121	5,370,200	50.85	10.8
T3-7	13,614	114	5,151,232	50.61	10.72
T1-11	14,911	101	5,015,163	50.65	13.06
I2-5	14,954	117	5,588,523	50.30	9.3
M2-4	14,989	107	5,521,688	51.01	44.03
S2-5	17,101	102	5,585,248	50.56	4.83
M3-28	18,091	97	5,754,056	50.55	9.54
T3-19	19,015	95	5,856,858	50.62	12.96
T1-52	20,742	84	5,700,106	50.31	24.61
AFR-2	20,861	85	5,752,069	50.69	41.95
M2-10	22,332	75	5,582,390	50.93	35.4
GD109	22,711	67	4,905,714	50.47	1.2
I2-1	23,125	68	5,632,497	50.89	65.64
S2-10	25,291	63	6,002,178	50.66	12.1
M3-18	27,148	47	4,800,461	50.77	9.73
S2-7	27,148	64	5,852,125	50.75	37.63
T1-49	28,588	57	6,028,358	50.78	1.82
AFR-6	28,746	58	5,636,541	50.57	10.88
S2-8	29,947	56	6,014,939	50.5	25.7
T1-9	35,181	44	5,123,419	50.69	12.49
SFR-6	36,805	44	5,024,664	50.65	16.54
TFR-15	39,555	40	4,704,649	50.89	19.7
M3-32	39,818	45	5,995,208	50.44	18.35
ELU65	46,408	35	4,893,644	50.65	10.3
T1-27	48,560	32	5,339,120	50.97	8.52
GU77	50,825	34	5,600,130	51.21	31.55
ELU103	51,772	34	5,629,601	50.78	19.18
I1-21	52,253	29	5,664,842	50.92	36.19
T1-32	52,361	33	5,362,438	50.47	9.75
T1-73	55,462	29	5,089,386	50.58	27.96
M3-19	56,424	28	4,866,004	50.56	11.04
EPU5	56,609	30	4,875,919	50.83	30
T1-56	57,551	30	5,761,442	50.4	17.32
I1-16	57,612	32	5,404,494	50.72	18.65
EPU62	58,039	27	4,979,186	50.73	0.8
GU15	59,632	31	5,437,433	50.56	16.53
GU41	60,562	26	5,019,387	50.75	14.56
ELU7	63,109	25	4,972,394	50.73	1.41

GU13	65,278	23	4,777,286	50.47	3.24
I1-28	66,591	26	5,420,730	50.76	33.21
GD3	69,188	25	5,096,017	50.85	22.27
T1-30	70,461	22	5,423,956	50.63	23.27
I1-12	70,756	26	5,170,047	50.44	15.40
SFR-11	76,811	22	5,227,403	50.66	31.99
T3-6	80,594	22	5,433,335	50.7	27.00
GU27	80,995	19	4,913,402	50.51	28.09
ELU71	82,060	20	4,979,055	50.74	0.4
TFR-1	82,704	23	5,629,233	50.67	65.6
I1-19	84,958	22	5,385,365	50.73	15.39
ELU98	85,575	21	5,411,422	50.39	9.68
T3-3	87,089	19	5,186,480	50.65	26.61
EPD5	87,284	21	5,404,635	50.56	28.46
ELU17	87,482	19	4,978,547	50.74	0.4
ELU67	87,606	19	4,977,900	50.74	0.8
EPU22	88,228	20	5,156,329	50.76	18.71
EPU51	90,598	18	4,964,100	50.69	11.18
GD131	91,113	15	5,230,245	50.57	90.38
T3-4	91,242	22	5,708,849	50.62	39.52
GU53	92,610	19	5,406,222	50.48	9.53
S2-4	96,735	20	5,519,072	50.43	19.99
T1-57	100,004	18	5,415,934	50.37	19.33
I1-17	100,097	20	5,392,204	50.74	18.69
ELU122	100,420	17	4,967,713	50.69	28.26
T1-25	102,293	18	5,408,044	50.35	26.65
T1-3	103,224	17	5,858,729	50.65	42.94
ELU88	103,785	17	4,976,742	50.59	10.13
T1-1	103,939	16	4,946,744	50.68	28.46
T1-61	104,032	18	5,121,170	50.67	8.98
ELU21	104,242	18	4,974,449	50.73	0.2
EPD30	104,925	15	4,787,052	50.75	0.21
ELU24	105,056	14	4,981,484	50.72	15.38
GU80	106,390	16	5,211,996	50.37	67.15
EPU17	106,671	15	4,923,598	50.68	9.22
ELU39	107,069	14	4,730,043	50.62	9.01
ELU72	107,272	17	4,858,247	50.69	10.27
GU34	107,488	15	4,846,293	50.78	0.41
T1-39	107,683	16	5,400,960	50.62	7.54
ELU28	108,043	14	4,745,339	50.89	2.84
T3-14	109,000	12	5,199,403	50.63	23.75
ELU22	109,377	15	4,979,149	50.72	0
T1-5	109,634	13	4,748,761	50.68	9.6
I1-32	110,485	16	5,190,226	50.37	13.64
ELU20	111,954	16	4,980,229	50.73	0
S2-2	112,253	15	5,289,074	50.25	0.19
GU47	114,419	12	4,846,073	50.7	14.84
SFR-4	114,470	15	4,975,197	50.78	0.20
SFR-15	114,554	14	4,754,370	50.58	12.85
GU45	117,001	16	5,319,647	50.79	20.7
ELU16	119,583	13	4,980,732	50.72	0
S2-3	122,183	16	5,236,836	50.72	8.13
T1-35	123,140	14	4,908,868	50.63	11.53

GU39	124,074	12	4,948,309	50.33	11.24
TFR-13	126,831	16	5,742,821	50.67	34.48
GU49	128,774	13	5,219,251	50.77	21.02
GU51	135,819	11	5,057,578	50.66	0
GU24	142,610	12	5,004,195	50.71	20.48
ELU87	152,411	11	5,253,977	50.45	20.61
GU35	154,804	10	5,063,753	50.61	10.64
GU46	163,299	12	5,218,231	50.72	53.03
GD46	179,117	10	5,154,659	50.66	0
GU48	180,109	12	5,334,803	50.5	23.45
ELU29	188,138	10	4,861,206	50.69	9.94
GU70	189,883	9	5,009,043	50.26	10.62
GU10	200,504	8	4,931,939	50.62	8.23
GU5	203,129	9	4,673,999	50.86	10.38
GD49	209,107	8	5,154,459	50.63	26.21
GU82	220,014	7	4,482,513	50.38	21.37
GD93	227,685	7	4,784,503	50.67	0
GU6	234,263	8	5,218,733	50.49	26.88
GU43	243,045	7	4,856,399	50.58	0
GU52	248,372	9	5,236,603	50.72	24.58
GU31	252,882	8	4,950,178	50.27	20.10
GD162	257,260	7	4,788,765	50.64	0
GD45	340,187	5	5,097,407	50.6	12.34
GU87	348,597	6	5,115,851	50.5	21.03
GU50	367,693	4	4,869,944	50.61	9.26
ELU34	401,570	5	4,991,917	50.55	9.76

Appendix 5. Assembly statistics for the 180 sequenced non-human *E. coli* genomes.

One hundred and eighty *E. coli* isolates were selected for sequencing, so as to represent the full diversity of sampled isolates. Assembly statistics were obtained from running the QUAST quality assessment tool (Gurevich *et al*., 2013). N50 values indicate the length for which the collection of all contigs of that length or longer covers at least half an assembly; L50 is the minimal number of contigs that cover half of the assembly; GC content refers to the total number of G and C nucleotides in the assembly, divided by the total length of the assembly; number of N's per 100 kb is the total number of uncalled bases per 100,000 assembled bases. The first 12 strains highlighted in red in the table, with N50 values < 1,900 bp and genome sizes < 4.3 Mbp, were excluded from further genomic analyses as they represent incomplete assembled genomes. Genome sizes ranging from ~4.4 Mbp to ~6.0 Mbp within the study population represent the average *E. coli* genome size of 5.19 Mbp and an average GC content of ~50% would be within the narrow range that is normal an *Escherichia* genome (Mann and Chen, 2010).



Appendix 6. Observed and expected proportions of genes shared between the human-clinical and non-human populations of *E. coli*.

Appendix 6. Observed and expected proportions of genes shared between the human-clinical and non-human populations of *E. coli*.

Appendix 6. Observed and expected proportions of genes shared between the human-clinical and non-human populations of *E. coli*.

Appendix 6. Observed and expected proportions of genes shared between the human-clinical and non-human populations of *E. coli*.

Appendix 6. Observed and expected proportions of genes shared between the human-clinical and non-human populations of *E. coli*.

Appendix 6. Observed and expected proportions of genes shared between the human-clinical and non-human populations of *E. coli*. The expected proportions were determined by carrying out a permutation test with random re-sampling of the population without replacement of genomes. Strain-specific genes were excluded from the data set. Permutations were run 1,000 times repeating for each gene category. The simulated proportions are plotted as histograms and the observed proportions are mapped onto the graphs for comparison. The permutation script was written by Ben Dickins (NTU) and the graphs were generated using the tidyverse package of the R statistical software.