# The microbial ecology of *Escherichia coli* in the vertebrate gut

By

Ebenezer Foster-Nyarko, BSc, MSc

A thesis submitted for the degree of Doctor of Philosophy

University of East Anglia

in the Faculty of Medicine and Health Sciences

Norwich Medical School

February 2021

# ABSTRACT

*Escherichia coli* has a rich history as biology's favourite organism, driving advances across many fields. In the wild, *E. coli* often resides innocuously in the human and animal gut but is also a common pathogen linked to intestinal and extraintestinal infections and antimicrobial resistance. A literature review exposed gaps in our knowledge of the ecology and evolution of this organism in the vertebrate gut. I therefore investigated the genomic diversity and burden of antimicrobial resistance of *E. coli* in three vertebrate hosts from the Gambia.

First, I explored the population structure of *E. coli* in non-human primates from the Gambia, where they interact with human communities. Here, I found strains closely related to those causing human extraintestinal infection, together with novel strains specific to the intestinal ecosystems of non-human primates.

Next, I investigated the population structure of *E. coli* in backyard chickens and guinea fowl, which are commonly reared in Gambian homes as affordable sources of protein. I identified a clade of *E. coli* sequence type ST155 that includes closely related isolates from poultry and livestock from sub-Saharan Africa, suggesting that poultry and livestock exchange strains of *E. coli* on this continent. I compared the prevalence of antimicrobial resistance genes in *E. coli* isolates from Gambian poultry to that seen in poultry isolates from around the world.

Finally, I used genomic analysis to shed light on the relative contributions of immigration and within-host evolution in the generation of diversity among commensal strains of *E. coli* in the guts of healthy children from rural Gambia.

In closing, I discuss the implications and prospects of these findings.

## LIST OF CONTENTS

**LIST OF TABLES**

**LIST OF FIGURES**

## LIST OF ABBREVIATIONS

| **Abbreviation** | **Meaning** |
|---|---|
| AIEC | Adherent invasive *E. coli* |
| AmpC | Class C (or group 1) beta-lactamases |
| AMR | Antimicrobial resistance |
| APEC | Avian pathogenic *E. coli* |
| ARIBA | Antimicrobial resistance identification by assembly |
| ATP | Adenosine triphosphate |
| BWA-MEM | Burrows-Wheeler alignment maximal exact match |
| cgMLST | Core-genome multi-locus sequence typing |
| CLIMB | Cloud infrastructure for microbial bioinformatics |
| CMY | Class C carbapenemases |
| CTX-M | Class A beta-lactamases named for their greater activity against cefotaxime (than the other oxyimino-beta-lactam substrates) |
| DAEC | Diffusely adherent *E. coli* |
| DBD | Daily dose per 100 bed days |
| DEC | Diarrheagenic *E. coli* |
| DID | Daily dose per 1000 clients |
| EAEC | Enteroaggregative *E. coli* |
| eBURST | Online version of the "based upon related sequence type" clustering algorithm |
| ECOR | *E. coli* reference |
| EDTA | Ethylenediaminetetraacetic acid |
| EHEC | Enterohaemorrhagic *E. coli* |
| EIEC | Enteroinvasive *E. coli* |
| EnPEC | Endometrial pathogenic *E. coli* |
| EPEC | Enteropathogenic *E. coli* |
| ESBL | Extended-spectrum beta-lactamase |
| ET | Electrophoretic type |
| ETEC | Enterotoxigenic *E. coli* |
| ExPEC | Extraintestinal pathogenic *E. coli* |
| GEMS | The global enteric multicenter study |
| GTDB | Genome taxonomy database toolkit |
| GTDB-Tk | Genome taxonomy database toolkit (GTDB-Tk) |
| HC | Hierarchical cluster |
| HGT | Horizontal gene transfer |
| HierCC | Hierarchical clustering |

**LIST OF ABBREVIATIONS (continued)**

| | |
|---|---|
| IgA | Immunoglobulin A |
| MDR | Multiple-drug resistance |
| MIC | Minimum inhibitory concentrations |
| MLEE | Multilocus enzyme electrophoresis |
| MLST | Multi-locus sequence typing |
| MPEC | Mammary pathogenic *E. coli* |
| MRSA | Methicillin resistant *Staphylococcus aureus* |
| NAUTICA | North American urinary tract infection collaborative alliance survey |
| NCBI | National center for biotechnology information |
| NDM | New Delhi metallo-beta-lactamase |
| NGS | Next generation sequencing |
| NMEC | Neonatal meningitis *E. coli* |
| NTEC | Necrotoxigenic *E. coli* |
| OXA | Oxacillinase group of β-lactamases |
| PCR | Polymerase chain reaction |
| QUAST | Quality assessment tool |
| RPM | Revolutions per minute |
| SAM | Sequence alignment/map |
| SCC | Scientific coordinating committee |
| SDS | Sodium dodecyl-sulphate |
| SENTRY | Global antimicrobial surveillance programme |
| SePEC | Human sepsis-associated *E. coli* |
| SHV | Class A beta-lactamases |
| SNP | Single nucleotide polymorphism |
| SNP | Single nucleotide polymorphism |
| SPATE | Serine protease autotransporters of *Enterobacteriaceae* |
| SPRI | Solid phase reversible immobilisation |
| SRA | Sequence read archive |
| ST | Sequence type |
| STEAEC | Shiga-toxigenic enteraggregative *E. coli* |
| STGG | Skim-milk-tryptone-glucose-glycerol |
| TEM | Class A beta-lactamases named after Temoniera (patient from whom it was discovered) |
| UPEC | Uropathogenic *E. coli* |
| UPGMA | Unweighted pair group method with arithmetic mean |

## LIST OF ABBREVIATIONS (continued)

| | |
|---|---|
| VCF | Variant call format |
| VFDB | Virulence factors database |
| WHO | World health organisation |

## DECLARATION OF AUTHORSHIP

I, Ebenezer Foster-Nyarko, declare that the work presented in this thesis is my own. I confirm that:

- This work was done wholly or mainly while in candidature for the degree of Doctor of Philosophy at the University of East Anglia.
- No part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution.
- Where I have described the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and by me, as outlined in the cases below:
    - Dr Jennifer D. Cramer, a collaborator at the American Public University System, Charles Town, WV, USA collected the non-human primate stool samples analysed in Chapter Three.
    - Dr Arss Secka euthanised and removed the caecal contents of the birds analysed in Chapter Four.
    - Colleagues at the Medical Research Council Unit the Gambia at London School of Hygiene and Tropical Medicine collected the human stool samples analysed in Chapter Five as part of the Global Enteric Multicenter Study (GEMS) and provided aliquots of the samples.
    - Contributions to the library preparation and sequencing of the isolates presented in this thesis were as follows:

      Short-read sequencing: author (40%) and Mr David Baker (60%)
      Long-read sequencing: author (70%) and Mr Steven Rudder (30%)
    - Dr George Savva wrote the R script for plotting the map of the study sites described in Chapter 3 (Figure 3.1).
    - Dr Matt Bawn wrote the R script which I adapted for displaying my phylogenetic trees and accessory gene contents.
    - Dr Marianne Defernez carried out the statistical tests of associations between virulence and antimicrobial resistance factors and host sources from which they were derived (discussed in Chapters Three and Five).

*"Tout ce qui est vrai pour le Colibacille est vrai pour l'éléphant."*
*(All that is true of E. coli is also true of the elephant)*

Jacques Monod

## PUBLICATIONS

The work presented in Chapters Three to Five of this thesis have been published in the following papers:

1. **Foster-Nyarko E**, Alikhan NF, Ravi A, Thilliez G, Thomson NM *et al*. Genomic diversity of *Escherichia coli* isolates from non-human primates in the Gambia. *Microb Genom* 2020 Sep 14;6(9). PMID: 32924917; DOI: 10.1099/mgen.0.000428 (**Chapter Three**).
2. **Foster-Nyarko E**, Alikhan NF, Ravi A, Thomson NM, Jarju S *et al*. Genomic diversity of *Escherichia coli* isolates from backyard chickens and guinea fowl in the Gambia. *Microb Genom* 2020 Nov 30 (online ahead of print). PMID: 33253086; DOI: 10.1099/mgen.0.000484 (**Chapter Four**).
3. **Foster-Nyarko E,** Ravi A, Ikumapayi UN, Sarwar G, Okoi C *et al*. Genomic diversity of *Escherichia coli* from healthy children in rural Gambia. *PeerJ* 2021 Jan 6;9:e10572. https://doi.org/10.7717/peerj.10572 (**Chapter Five**)

## ACKNOWLEDGEMENTS

My PhD journey has been remarkable and life changing. Coming from a clinical microbiology background, I started this PhD with zero bioinformatics or genomics skills. I recall the excitement when I executed my first command-line analysis (downloading reads from Illumina BaseSpace and concatenating them). This has been a steep learning curve, with highs and lows, but generally enjoyable. I would not have come this far without the generous support and wise guidance from my supervisory team. I would like to thank them sincerely for the progress I have made these past three years.

First and foremost, I am extremely grateful to Professor Mark Pallen, my primary supervisor, for taking me on. The meetings we spent together interrogating data and chatting through inferences and interpretations have honed my skills in making sound scientific judgements. I want to thank the Quadram Institute for the studentship that allowed me to conduct this thesis. I am indebted to Professor Martin Antonio, my line manager of seven years at the Medical Research Council Unit the Gambia at London School of Hygiene and Tropical Medicine and external supervisor. His scholarly example inspired me to reach for the stars. I express my sincere thanks to Dr Justin O'Grady, my secondary supervisor, for his unwavering support and belief in me. Thank you for encouraging me when I felt like giving up. To Dr Nabil-Fareed Alikhan: you have taught me so much in the past two years as part of my supervisory team and I am proud to have learnt from you. Dr Anuradha Ravi guided my baby steps in genomic analysis and interpretation of the results. Her rich insights and thoroughness helped me a great deal. I want to express my sincere thanks to Dr Nicholas Thomson, Dr Rachel Gilroy and Dr Gemma Kay for proofreading Chapter 1 of this thesis and members of the Pallen Group at the Quadram Institute for insightful comments and suggestions during group meetings. Dr Andrea Telatin and Dr Gaëtan Thilliez: thank you for your patience each time I came to you with bioinformatics queries. Dr Matt Bawn first introduced me to the concept of hierarchical clustering for resolving population structures. Thanks for your brilliance in bioinformatics. I dedicate this thesis to Dr Martin Ota, an outstanding mentor and friend, for his tremendous impact on my career.

I appreciate the love of my family and friends in Ghana and abroad. It is their prayers that have kept me going. I am also profoundly grateful for the friendships I have made on this journey; Federico, Gregory and Prithika stand out among them.

Finally, without the exceptional support and encouragement of my darling wife and lovely children (the four Cs), it would be impossible to complete my study. Thanks for putting up with my busy schedule and many long hours spent working on this dissertation. Love you forever!

# 1 CHAPTER ONE: INTRODUCTION

## 1.1 *Escherichia coli*: a versatile organism

"It is a truth universally acknowledged that there are only two kinds of bacteria. One is *Escherichia coli* and the other is not." [1]. This widely quoted epigram reflects the exalted status of *E. coli* in microbiology as the favourite model organism. Many advances in biology have been driven by studies with *E. coli*, particularly, using the strain designated K-12 and its derivatives [2, 3] (reviewed in Section 1.1.1).

Theodor Escherich, a Bavarian doctor, first described *Escherichia coli* on 14 July 1885 in a lecture to the Society for Morphology and Physiology in Munich [4]. Over fifteen months, Escherich observed and isolated nineteen different bacteria in Otto von Bollinger's bacteriology laboratory [4], including the *Bacterium coli commune* ("the common colon bacillus" [5, 6])—which we now know as *E. coli* and *Bacterium lactis aërogenes* (now *Klebsiella pneumoniae*). Aided by Christian Gram's new staining technique [3, 7, 8], Escherich described the organism as a Gram-negative bacillus of approximately 1.1-1.5 µM x 2.0-6.0 µM. The name *Escherichia coli* was proposed by Castellani and Chalmers in 1919 to honour Escherich and was officially adopted in 1958 by the Judicial Commission of the International Committee on Systematic Bacteriology. It was subsequently included in the Approved Lists of Bacterial Names in 1980 [7, 9-11].

In the wild, *E. coli* exists as a common resident of the vertebrate gut and non-host associated habitats such as water, soil, manure and food [12-15]. Due to its remarkable metabolic and regulatory abilities, *E. coli* can also survive under prolonged periods of non-growth [16]. Thus, *E. coli* represents a highly versatile species, capable of adapting to many different ecological habitats [16]. *E. coli* is also a versatile pathogen, eliciting a broad spectrum of diseases and responsible for at least two million human deaths per year [17]. The organism's role in intestinal and extraintestinal disease was recognised not long after its discovery [3]. Ørskov and Ørskov note, "Any *E. coli* strain can probably cause invasive disease given the right opportunities, and *E. coli* has therefore aptly been called an opportunistic pathogen" [18].

*E. coli* is a facultative anaerobe, meaning that it can grow in aerobic and anaerobic conditions. It is the most common aerobe in the lower intestine of mammals [2]; however, it typically constitutes only 0.1-5% of a gut microbial community comprised of over 500 other bacterial species [2]. This community is typically dominated by obligate anaerobes including members of Firmicutes and Bacteroidetes—which make up at least 90% of the gut microbial population [2, 19, 20]. Nevertheless, *E. coli* can hold its ground in this highly competitive, ever-changing niche, existing in a life-long relationship with its host.

### 1.1.1 Impact of studies with *E. coli*

The following selected examples illustrate the contribution of studies using *E. coli* to many advances across a variety of fields (Table 1.1).

***Table 1.1:*** *The contribution of studies using E. coli across various fields.*

| **Field(s)** | **Contribution of studies with *E. coli*** | **Reference(s)** |
|---|---|---|
| Molecular biology, physiology and genetics | The explanation of the genetic code | [21] |
| | DNA replication | [22] |
| | Transcription | [23] |
| | The life cycle of bacterial viruses | [24, 25] |
| | The elucidation of the molecular basis of antibiotic tolerance | [26] |
| | The discovery of restriction enzymes | [27, 28] |
| | Swarming motility | [29, 30] |
| | Gene regulation | [31] |
| | Elucidation of the structure and function of ATP synthase | [32] |
| Pharmaceuticals | The synthesis of recombinant proteins in vivo, such as insulin, which is used to treat millions of people with diabetes worldwide | [50] |
| | The synthesis of several other biopharmaceuticals, such as human interferon-β, interleukin-2, human growth hormone and human blood clotting factors | [33, 34] |
| Evolution | *E. coli* is the model organism of choice in experimental evolution studies; for example, in Lenski's long-term evolution experiment, on-going since February 1988 and spanning over 60,000 generations | [35] |
| | The demonstration of the stochastic nature of mutations | [36, 37] |
| | Mapping the trajectory of long-term fitness | [38] |
| | The elucidation of how sexual recombination influences adaptation | [39] |
| | Insights into predator-prey interactions | [40] |
| | The evolution of a novel trait, aerobic citrate utilization | [41] |
| Genetic engineering and biotechnology | The development of recombinant DNA techniques and molecular cloning, the production of biofuels and industrial chemicals such as phenol | [42] |
| | Mannitol production | [43] |
| | Ethanol production | [44, 45] |

### 1.1.2 *E. coli* pathotypes

Pathogenic *E. coli* are classified into "pathotypes" or "pathovars" [46, 47] (Table 1.2) based on several criteria, including:

- Site of infection (e.g., uropathogenic strains, named for their impact on the urinary tract, extraintestinal pathogenic *E. coli* (ExPEC), which cause infections in organs outside the gut)
- Host (e.g., avian pathogenic *E. coli* (APEC), named after infections in avian species)
- Site and host (e.g., neonatal meningitis *E. coli* (NMEC) which infect the cerebrospinal fluid in new-borns)
- Pathogenesis (e.g., Shiga-toxigenic *E. coli* (STEC)).

The pathogenic strains of *E. coli* have acquired specific virulence factors that enable them to adapt to new niches and cause a wide variety of diseases [46, 48]. These include adhesion/colonisation factors, toxins and effectors enabling pathogenic strains to colonise sites such as the urethra and small intestine and affect various fundamental eukaryotic processes [46] (Tables 1.3 and 1.4). For example, uropathogenic *E. coli* strains are equipped with type I fimbriae, *AfA/Dr* adhesins and pyelonephritis-associated pili (PAP) that enable them to colonise and infect the urinary tract. NMEC and sepsis-associated *E. coli* are armed with the K1 polysaccharide capsule that facilitates their evasion from host complement-mediated killing [16, 49].

Despite its designation as a separate genus, *Shigella* is classified as an intestinal pathogenic *E. coli* (InPEC) pathotype. It possesses virulence traits and pathogenicity that closely resemble enteroinvasive *E. coli* (EIEC) and is thus regarded as an EIEC pathotype [50, 51]. Phylogenomic data support the classification of *Shigella* as an *E. coli* pathotype [52, 53], even though its nomenclature has been retained solely for historical reasons and to avoid confusion in the clinical setting. Chaudhuri and Henderson [54] advocate that the clinical and academic community working on *E. coli/Shigella* adopts a similar approach as was used in the classification of species such as *Salmonella typhimurium, S. enteritidis* and *S. typhi* as *Salmonella enterica* subgroups, so that the *Shigella* species are re-designated as subspecies within *E. coli* to avoid their neglect in *E. coli* studies.

***Table 1.2**: Classification and examples of the pathotypes of E. coli.*

| **Pathotype** | **Pathovar** | **Virulence mechanism(s)** | **Host range** | **Reference** |
|---|---|---|---|---|
| InPEC | EPEC | Locus of enterocyte effacement; pathogenicity island 1 | Humans, all mammals | [55] |
| InPEC | EAEC | Small fimbrial adhesins; toxins; transcriptional activator gene; aggregative adhesion | Humans | [55] |
| InPEC | EHEC | Shiga toxin or verotoxins; afimbrial and fimbrial adhesins | Humans, piglets | [55] |
| InPEC | ETEC | Heat labile and heat-stable enterotoxins | Humans, ruminants, pigs, dogs | [55] |
| InPEC | EIEC/Shigella | Invasion and multiplication in enterocytes | Humans, primates | [55] |
| InPEC | DAEC | Adhesins | Humans, animals | [55] |
| InPEC | STEAEC | Shiga toxin | Humans | [56] |
| InPEC | AIEC | Adherent invasive phenotype | Humans and animals | [55] |
| ExPEC | APEC | Adhesins, secretion and iron uptake systems, increased serum survival and cytotoxic proteins | Birds | [57] |
| ExPEC | UPEC | Fimbrial adhesins; siderophores, resistance to complement | Humans, animals (especially dogs and cats) | [58] |
| ExPEC | NMEC | Iron acquisition systems, degradation of interferon-gamma and cleavage of the human defensin LL-37 | Humans | [59] |
| ExPEC | SePEC | Fimbrial adhesins; siderophores; resistance to complement | All mammals and birds (especially poultry) | [58] |
| ExPEC | MPEC | Unknown | Animals | [60] |
| ExPEC | ExPEC | Type II, IV and VI secretion systems, long polar fimbriae (lpfA) and iron acquisition | Animals | [61] |
| ExPEC | NTEC | Cytotoxic Necrotizing Factors 1 or 2 and α haemolysin; fimbrial and/or afimbrial adhesins; siderophores; resistance to complement | Humans, animals and ruminants | [58] |

InPEC, Intestinal pathogenic *E. coli*; ExPEC, Extraintestinal pathogenic *E. coli*; DEC, diarrheagenic *E. coli*, EPEC, Enteropathogenic *E. coli;* EAEC, Enteroaggregative *E. coli;* EHEC, Enterohaemorrhagic *E. coli;* ETEC, Enterotoxigenic *E. coli;* EIEC, Enteroinvasive *E. coli;* DAEC, Diffusely adherent *E. coli;* STEAEC, Shiga-toxigenic Enteraggregative *E. coli;* AIEC, Adherent invasive *E. coli*; APEC, Avian pathogenic *E. coli;* UPEC, Uropathogenic *E. coli;* NMEC, Neonatal meningitis *E. coli;* SePEC, Human sepsis-associated *E. coli;* MPEC, Mammary pathogenic *E. coli;* EnPEC, Endometrial pathogenic *E. coli;* NTEC, Necrotoxigenic *E. coli.*

***Table 1.3:*** *Colonisation and fitness factors.*
*(Adapted from [46]).*

| **Virulence factor** | **Pathotype** | **Effect(s)** |
|---|---|---|
| *icsA/virG* | EIEC | Nucleates actin filaments |
| Intimin | EPEC/EHEC | Adhesin, inducing TH1 response |
| Dr adhesins | DAEC/UPEC | Adhesin, binds to decay-accelerating factor and activates phosphatidylinositol 3-kinase, induces MHC class I chain-related gene A |
| P (*Pap*) fimbriae | UPEC | Adhesin, also induces cytokine expression |
| Colonisation factor antigens | ETEC | Adhesin |
| S fimbriae | UPEC/NMEC | Adhesin |
| Bundle-forming pili (BFP) | EPEC | Type IV pili |
| Aggregative adherence fimbriae | EAEC | Adhesin |
| *paa* | EPEC/EHEC | Adhesin |
| *toxB* | EHEC | Adhesin |
| *Efa-1/LifA* | EHEC | Adhesin |
| Long polar filaments | EHEC/EPEC | Adhesin |
| *saa* | EHEC | Adhesin |
| *ompA* | NMEC/EHEC | Adhesin |
| Curli | Various | Adhesin, binds to fibronectin |
| *ibeA/B/C* | NMEC | Stimulates invasion |
| *aslA* | NMEC | Stimulates invasion |
| Dispersin | EAEC | Stimulates colonisation; facilitates mucous penetration |
| K antigen capsules | MNEC | Antiphagocytic activity |
| Aerobactin | EIEC | Siderophore, iron acquisition |
| Yersiniabactin | Various | Siderophore, iron acquisition |
| *ireA* | UPEC | Siderophore, iron acquisition |
| *iroN* | UPEC | Siderophore, iron acquisition |
| *chu* (*shu*) | EIEC/UPEC/NMEC | Siderophore, iron acquisition |
| Flagellin | All | Motility, inducing cytokine expression through Toll-like receptors |
| Liposaccharides | All | Inducing cytokine expression through Toll-like receptors |

The full names of the pathotypes are as provided in the footnote under Table 1.2.

***Table 1.4:*** *E. coli toxins and effectors.*
*(Adapted from [46]).*

| Virulence factor | Pathotype | Toxin class | Effect(s) |
|---|---|---|---|
| Heat-labile enterotoxin | ETEC | AB subunit/type II effector | ADP ribosylates and activation of adenylate cyclase, leading to ion secretion |
| Shiga toxin | EHEC | AB subunit | Depurination of rRNA, inhibiting protein synthesis and inducing apoptosis |
| Cytolethal distending toxin | Various | ABC subunit | DNase activity, blocks mitosin in G2/M phase |
| Shigella enterotoxin 1 | EAEC/ EIEC | AB subunit | Ion secretion |
| Urease | EHEC | ABC subunit | Cleaves urea to $NH_3$ and $CO_3$ |
| *EspC* | EPEC | Autotransporter | Serine protease, cleavage of coagulation |
| *EspP* | EHEC | Autotransporter | Serine protease, cleavage of coagulation factor V |
| Haemoglobin-binding protease | ExPEC, APEC | Autotransporter | Degradation of haemoglobin to release haem/iron |
| *pet* | EAEC | Autotransporter | Serine protease; ion secretion and cytotoxicity |
| *pic* | UPEC, EAEC, EIEC | Autotransporter | Protease/mucinase |
| *sat* | UPEC | Autotransporter | Vacuolation |
| *sepA* | EIEC | Autotransporter | Serine protease |
| *sigA* | EIEC | Autotransporter | Ion secretion |
| Cycle-inhibiting factor | EPEC, EHEC | Type III effector | Blocks mitosis in G2/M phase, resulting in the inactivation of cdk1 |
| *espF* | EPEC, EHEC | Type III effector | Opens tight junctions and reduces apoptosis |
| *espH* | EPEC, EHEC | Type III effector | Modulates filopodia and pedestal formation |
| *map* | EPEC, EHEC | Type III effector | Disrupts mitochondrial membrane potential |
| *tir* | EPEC, EHEC | Type III effector | Nucleates cytoskeletal proteins, loss of microvilli and GAP-like activity |
| *ipaA* | EIEC | Type III effector | Actin depolymerisation |
| *ipaB* | EIEC | Type III effector | Apoptosis, Interleukin-1 release and membrane insertion |
| *ipaC* | EIEC | Type III effector | Actin polymerisation |
| *ipaH* | EIEC | Type III effector | Modulation of inflammation |
| *ipgD* | EIEC | Type III effector | Inositol 4-phosphatase and membrane blebbing |
| *VirA* | EIEC | Type III effector | Microtubule destabilisation and membrane ruffling |
| *stcE* | EHEC | Type II effector | Cleavage of C1-esterase inhibitor and disruption of the complement cascade |
| *hlyA* | UPEC | Repeats-in-toxin (RTX) toxin | Cell lysis |
| *ehx* | EHEC | RTX toxin | Cell lysis |
| Cytotoxic necrotising factors (1 and 2) | NMEC, UPEC, NTEC | - | Altered cytoskeleton and necrosis |
| *LifA/Efa* | EPEC, EHEC | - | Inhibits lymphocyte activation and adhesion |
| Shigella enterotoxin 2 | EIEC, ETEC | | Ion secretion |
| Heat-stable enterotoxin a | ETEC | Heat-stable enterotoxins | Activating guanylate cyclase, leading to ion secretion |
| Heat-stable enterotoxin b | ETEC | Heat-stable enterotoxins | Ion secretion via increasing intracellular calcium |
| Enteroaggregative *E. coli* heat-stable enterotoxin | Various | Heat-stable enterotoxin | Activating guanylate cyclase, leading to ion secretion |

The full names of the pathotypes are as provided in the footnote under Table 1.2.

Despite the importance of pathotyping to the epidemiology and pathogenesis of strains, pathotype classification "has been rendered difficult to follow" [48] due to the recent discovery of complex hybrid pathotypes:

- The strain that caused an outbreak of foodborne illness in Germany in 2011 [62], which swept across most of Europe, killing fifty-four individuals and causing approximately 4,000 infections including 900 cases of haemolytic uraemic syndrome (HUS), was, in fact, an EHEC-EAEC hybrid strain [48, 63] belonging to serotype O104:H4 and sequence type ST678. This strain combined virulence characteristics of EHEC (Shiga toxin production) and adherence typical of EAEC strains, despite lacking the Type III secretion and *tir*/intimin system. It also possessed virulence factors commonly found in ExPEC strains, such as yersiniabactin and aerobactin (iron acquisition factors) and demonstrated expanded-spectrum beta-lactamase resistance [63].
- A hybrid clone belonging to serotype O80:H2 and sequence type ST301 (clonal complex 165) has emerged in France, Belgium and Switzerland, capable of causing HUS and bacteraemia [64-68]. This strain possesses all the virulence factors typical of EHEC strains, such as intimin, Shiga toxin production and enterohaemolysin, yet it belongs to the phylogenetic group A, unlike other EHEC strains [69]. It possesses a large plasmid (>100 kb), which encodes a resistance cassette, providing resistance characteristics to a wide range of antimicrobials, including cotrimoxazole, tetracyclines, streptomycin and penicillin [64, 70]. Carriage of large plasmids is commonly associated with ExPEC strains [59, 71].
- Hybrid clones exhibiting characteristics of the B2 phylogenetic backbone typical of ExPEC strains and EPEC and STEC attributes have been recently reported in the literature, further blurring the lines of pathotype boundaries that have been used to define *E. coli* pathogens for so long. Isolates belonging to serotype O153:H10 and harbouring *eae* (depicting an atypical EPEC-ExPEC hybrid pathotype) have been detected in meat, poultry farms, human diarrhoeagenic samples and wildlife from northwest Spain [72]. Similarly, isolates belonging to serotype O137:H6 and ST2678 and positive for *eae*, *bfpA* and *stx2f* genes—demonstrating typical EPEC-STEC hybrid pathotype—have been isolated in exotic psittacine birds [73]. In addition, strains belonging to serotype O2:H4 and ST141 and demonstrating uropathogenic traits have been found with some characteristics of the EHEC pathotype [74, 75]. Also, an ST12 strain belonging to serotype O4:H1 was found to harbour the locus of enterocyte effacement pathogenicity island and the bundle forming pili protein (encoded by the *bfp* gene*)* typical of the EPEC pathotype but simultaneously causing diarrhoea and bacteraemia in the same patient [76].

- Some EPEC isolates that harbour the heat-stable enterotoxin produced by ETEC strains have been described [77].

These examples amply demonstrate that pathotyping is limited in its capacity to adapt to new strains that fail to respect the currently utilised pathotype boundaries. A further limitation lies in the use of negative criteria to identify pathogens. For instance, a strain is described as a typical EPEC if it presents the locus of enterocyte effacement but does not produce Shiga toxin. On the other hand, if a strain both lacks Shiga toxin and *bfp,* it is classified as an atypical EPEC [46]. As Robins-Browne *et al.* [55] point out, "characterising pathogens based on their lack of one or more virulence determinants may group several types of distantly related or unrelated bacteria together and cause some distinct pathogenic categories with uncharacterised virulence determinants to be overlooked".

### 1.1.3 *E. coli* serotyping

The first efforts at unravelling the diversity of *E. coli* in humans was based on serotyping [78-81]—which distinguishes isolates based on their agglutination patterns when reacted against antisera raised against three surface antigens: O (oligosaccharides), K (capsule) and H (flagella) [82]. Kauffmann's classification of *E. coli* into somatic, capsular and flagellar serotypes in 1947 was pivotal to early serotyping studies [83]. Kauffmann and Vahlne [84] (reviewed in [81] and [85]) described the K antigen to represent the cell envelope that masks the O antigen, thus rendering some strains O-non-typable. As very few labs could perform K typing, it was infrequently used, while O- and H-typing quickly became the gold standard [82].

Using their newly developed agglutination scheme, Kauffmann and Dupont serotyped strains sourced from infantile diarrhoeal cases across several centres and observed that most of them belonged to serogroup O55 and O111 [86]. Reports from investigations of infantile diarrhoea from all over the world quickly confirmed the critical role of O111 and soon other serogroups emerged as causative agents of paediatric diarrhoea [87-90]. By 2016, at least 186 O-antigens (numbered O1-O188, excluding O31, O47, O67, O72, O94 and O122 which were left out) and 53 H-antigens (numbered H1-H56, but less H13, H22 and H50 which are no longer in use) had been defined [82].

Traditionally, serotyping techniques for identifying O, H and K groups relied upon bacterial agglutination (O and H) and gel immunoprecipitation or phage typing (K antigen) [91]. Whole-genome sequencing (WGS)-based *in silico* serotyping tools have been developed for short reads as of 2015 [92], digitising and significantly simplifying the serotyping procedure, as "one WGS run can replace multiple assays for bacterial typing" [93]. (WGS is discussed in Section 1.5.1).

The *E. coli* serotyping method is fraught with several challenges. First, it is a highly complex system, as evidenced by the high number of O and H groups that make up the scheme (the final number of serotypes is at least 100,000, based on the possible combinations of the O, H and K antigens found in nature [18]). Thus, it is labour-intensive and time-consuming [82]. Further, cross-reactivity among the antisera is common (coupled with batch-to-batch variations in antibodies) and a large number of strains remain non-typable—particularly, Shiga-toxigenic *E. coli* [94]. Furthermore, evidence from multilocus enzyme electrophoresis (discussed in Section 1.2.1) signalled that serotyping is less discriminatory and tends to group distantly-related strains under the same O:H type [95].

## 1.2 Investigating the population structure of *E. coli*

The population structure of *E. coli* was an early target of investigations [96, 97], owing to its ease of propagation and laboratory manipulation, short generation time, broad spectrum of phenotypes and lifestyles and haploid chromosome [98]. The first quantitative *E. coli* population genetics study was by Milkman in 1973 [97], who studied variations in the frequency distribution of classes of electrophoretic mobility at each of five loci in 829 natural *E. coli* clones (between ten to twenty clones isolated from a single faecal sample) sourced from 156 samples of diverse sources. Milkman investigated the "neutral hypothesis" that within large bacterial populations, selective neutrality of the different alleles of an enzyme would result in many electrophoretic variants. "The neutral hypothesis attributes most observed electrophoretic variation and most amino acid substitutions over the course of evolution, to the random genetic drift of the frequencies of various alleles at a locus, all of practically equivalent adaptive value." [97]. However, Milkman observed very few electrophoretically different alleles among the samples he studied and concluded that the neutral hypothesis was incorrect. He interpreted his results to favour the selection hypothesis, which "rejects the notion of many neutral alleles at a locus and would predict a small effective number of alleles and great genetic similarity among many individuals." [97]

Levin subsequently suggested that Milkman's conclusion relied on the assumption of widespread recombination within a large population, which influences the observed evolutionary relationships among related strains [99]. Milkman's work resulted in multi-locus enzyme electrophoresis being widely adopted for bacterial diversity studies [54], spawning several similar studies over the next few years with extensive strain collections (discussed in Section 1.2.1).

### 1.2.1 Multilocus enzyme electrophoresis (MLEE)

The high level of variation in the electrophoretic mobilities of enzymes vital to the normal functions of metabolism in eukaryotes inspired the use of MLEE to explain the metabolic differences in prokaryotes. The critical principle of MLEE is that electrophoretic mobilities, also known as electromorphs or allozymes "directly equated with alleles of the corresponding structural gene and that electromorph profiles over the sample of different enzymes (frequently termed electrophoretic types or ETs) correspond to multilocus chromosomal genotypes" [98]. Metabolic enzymes that were expressed in all isolates of a specific strain (e.g., enzymes involved in glycolysis) were targeted, based on the fact that allelic variation at these loci was not affected by environmental conditions (e.g., laboratory medium or storage) and the observed variation at these loci was selectively neutral (or nearly so). Thus, convergence to the same allele via adaptive evolution was minimal [100, 101]. In this way, the genetic variation at multiple chromosomal loci could be analysed rapidly and the information used to infer the genetic relationship among strains.

Given the large number of alleles at a particular locus in a bacterial population, generation of identical strains via recombination was taken to be rare and strains of the same electrophoretic type were considered to be descended from a common ancestor. Each enzyme in a chosen set of multiple core metabolic genes is electrophoresed on an agar gel: the differences in how far a band travels on the gel are indicative of mutations resulting in substitutions of amino acids and thus affecting the net charge of the enzyme [102]. The matrix of pair-wise differences between the electrophoretic types can be used to construct a dendrogram to depict the genetic relatedness among isolates.

Based on MLEE data, it was demonstrated that certain allelic combinations occurred multiple times, suggestive of a clonal population structure with limited recombination and that the diversity within individuals resulted primarily from the independent immigration of new strains [95, 96, 103, 104]. By analysing an extensive collection of twelve enzyme loci in 1,705 clones of *E. coli* drawn from various human and animal sources, Whittam *et al.* [103] observed that three subspecific groups existed within the *E. coli* population (designated I-III), as indexed by MLEE. Through these studies, it became apparent that O, H and K serotypes did not correlate well with genetic diversity, as closely related strains may be assigned different serotypes [95, 105, 106].

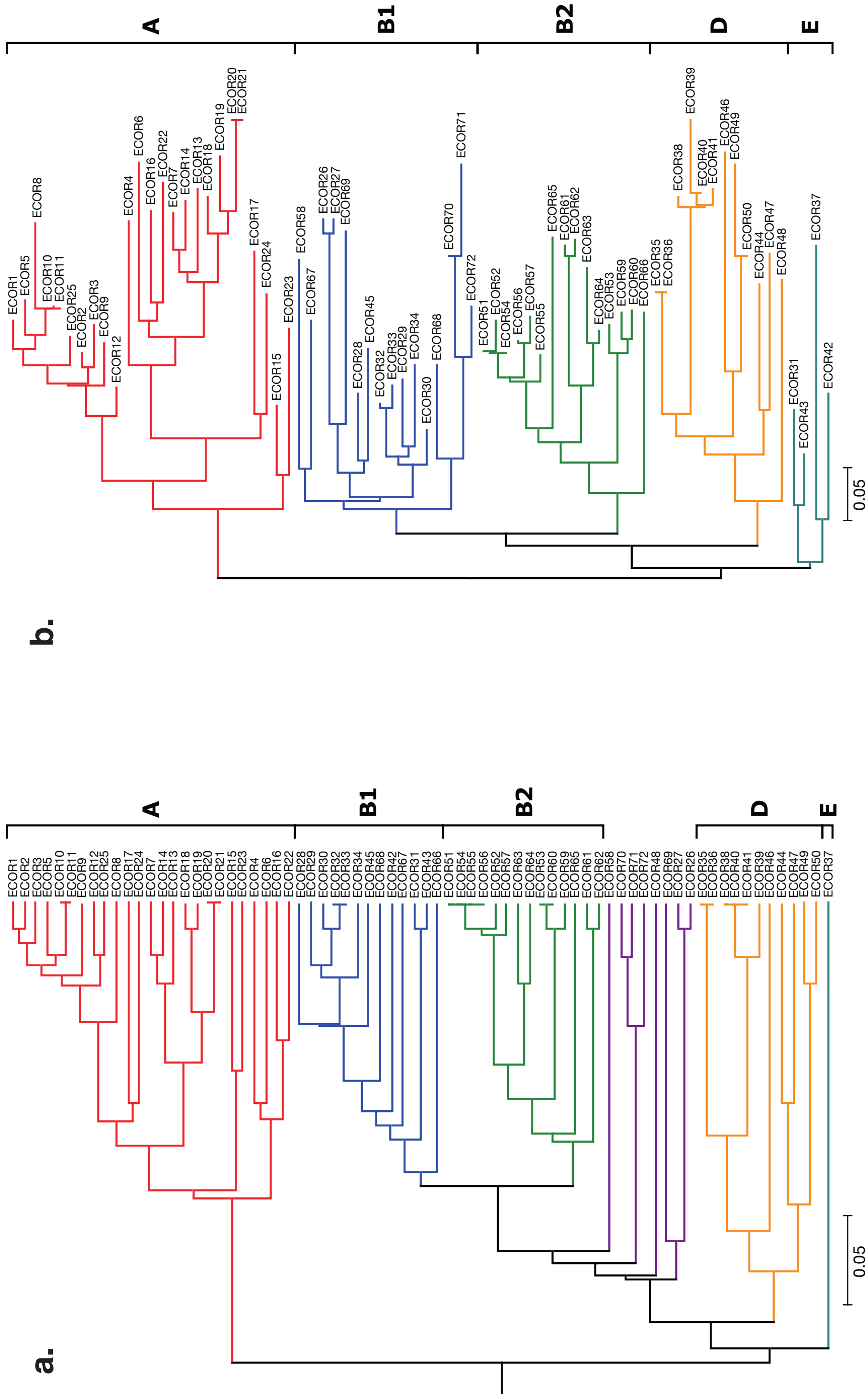
Subsequently, in 1984, Ochman and Selander established a reference collection of 72 strains sourced from a variety of hosts (human and sixteen other mammalian hosts) across various geographical locations to represent the genotypic diversity in the *E. coli* species as characterised by MLEE, "for use in studies of variation and genetic structure in natural populations" [105]. Selander *et al.* utilised this strain collection, which became known as the ECOR (*E. coli* reference) collection, to classify *E. coli* into six significant lineages or haplogroups (also known as phylogroups/phylotypes): A, B1, B2, C, D and E, based on cluster analysis of MLEE data derived for 35 loci [107]. Although their phylogenetic tree was based on multilocus enzyme electrophoresis (MLEE), improved methods (including rDNA restriction fragment length polymorphism, random amplified polymorphic DNA, multilocus sequence typing and recently, whole-genome sequencing) have confirmed the topology it describes [54, 108-111]. Selandar *et al.* utilised the unweighted pair group method with arithmetic mean (UPGMA) tree-building algorithm [112] to reconstruct their phylogenetic tree. This method is based on a strict molecular clock, which assumes a constant evolution rate across all the lineages. Herzer *et al.* [113] applied a neighbour-joining algorithm [114] based on the assumption of a relaxed molecular clock (more robust than the earlier method by [107]) to analyse the ECOR collection, with the resultant tree confirming the four significant groups A, B1, B2 and D described earlier (Figure 1.1).

Phylogroup E was identified as a new group; however, phylogroup C was not identified by Herzer *et al.* and is subsequently not used. Mid-point rooting of the tree from this study (without an outgroup species) suggested that phylogroup A was the first lineage to diverge, with phylogroups B1 and B2 being sister clades [113]. Two more phylogroups (F and G) have been recently described [115-117]. The currently accepted hypothesis is that phylogroups B2, F and D appear to be the most basal taxa, with phylogroup E emerging before phylogroups C, B1 and A, which are considered to be the most recently diverged lineages [116]. The accurate detection of *E. coli* phylogroups is useful in predicting the ecological niche, lifestyle and pathogenic potential of strains [14, 118], with the most anciently diverged encompassing mostly extraintestinal pathogenic strains. In contrast, the most recently diverged lineages span strains that are associated with life-threatening intestinal diseases such as dysentery and haemolytic uraemic syndrome [109, 119].

### 1.2.2 *E. coli* phylotyping

*E. coli* isolates can be assigned to phylogroups using the "Clermont typing" method—based on the presence or absence of the genes, *chuA, yjaA* and *arpA*, and the DNA fragment, TspE4.C2

[120, 121]. Clermont typing, published in 2000 [120, 121], consisted of a triplex PCR method for differentiating *E. coli* into the seven phylogroups*:* A, B1, B2, C, D, E and F. Besides the classic *E. coli* strains described in the seven phylogroups, five *Escherichia* clades were described in 2009, designated clades I-V.



**a.** A UPGMA tree similar to that reconstructed by Selander et al. in 1987 (Figure 1 of Reference 107), using MLEE data from 35-enzyme loci.
**b.** A neighbour-joining tree reconstructed using data from 38-enzyme loci, similar to that reported by Herzer et al. in 1990 (Reference 112).
Both trees gratefully obtained from Dr Roy Chaudhuri (personal communication, January 2021), with very minor edits by author and used with permission.

***Figure 1.1:*** *Phylogenetic trees depicting the genetic relationships among the 72 strains in the ECOR collection.*

These clades are otherwise referred to as cryptic *Escherichia* clades, as they could not be phenotypically distinguished from *E. coli* [118]; however, Clermont *et al.* [122] have observed some differential utilisation of lysine and ornithine between *E. coli* and the cryptic clades. Accumulating data suggests that these cryptic clades are overabundant in environmental samples (water, soil and aquatic sediments), although some clades have been associated with birds and non-human mammals [118, 122-124].

Clermont *et al.* updated their typing scheme in 2013 to include a quadruplex reaction, encompassing sub-groups within the phylotypes and the cryptic clades. The expanded typing scheme now detects seven phylogroups of *E. coli*: A, B1, B2, C, D, E and F and cryptic clades I-V within the species *Escherichia* [121]. An eighth phylogroup (phylogroup G) was recently proposed for a lineage of *E. coli* characterised by high extraintestinal virulence and antimicrobial resistance [117]. Members of this clade are intermediate between B2 and F and are typed phylogroup F by *in vitro* and *in silico* Clermont methods [117].

Recently, phylotyping has been made more accessible, with the advent of *in-silico* sequence-based methods. Two such methods have been published: the ClermonTyper [116] and EzClermont typing tool [125]. Both methods were highly congruent with the classical typing by PCR (99.4% and 94% concordance respectively). The advantages of these sequence-based methods over the traditional method are their ease of use and quick turn-around time. A further advantage is that these methods can be updated as our knowledge of the *Escherichia* genus phylogeny continues to grow.

The *in-silico* methods require a genome in a DNA FASTA or multi-FASTA format as input. A BLAST database is created with the input as the query genome. Using BLASTn user-defined parameters, matches against the same set of primers used in the PCR assay are sought and the presence or absence of each primer pair used to predict the phylotypes as for the PCR method. Both the ClermonTyper and EzClermont typing methods are available as command-line and web-interface platforms and are open-source.

### 1.2.3 Multilocus Sequence Typing

Although MLEE was seminal in estimating the population structure for the bacterial kingdom, it had its limitations. Target enzymes could be phenotypically modified in response to environmental conditions such as cofactor binding and cleavage of transport sequences, or phosphorylation, thus hampering the reproducibility of MLEE results. Further, similar electromorphs could be derived from an enzyme with different amino acid sequences. In addition, silent mutations that change the DNA but not the protein sequence will yield the same electromorph profile.

The development of multilocus sequence typing (MLST) represented an equivalent typing approach that enabled the unambiguous classification of bacterial isolates in a portable, standardised and reproducible manner with accompanying comparable DNA sequence database inter-laboratory comparisons [110, 126, 127]. MLST owed much to the pioneering technique of MLEE, as it adapted the proven concepts and methods of MLEE to identify alleles directly from the nucleotide sequences of internal fragments of housekeeping genes instead of comparing the electrophoretic mobilities of the enzymes they encode [126]. As the name indicates, it is a method that uses DNA sequences from multiple loci to characterise strains in populations. The loci are chosen from six to eight housekeeping genes (400-500 nucleotides long—a suitable length for direct sequencing of a DNA fragment with a single primer) which are likely to be under strong purifying selection and thus the detected variations are likely to be selectively neutral [54]. The term "MLST" was coined by Maiden *et al.* in 1998 [126] and since then, MLST has been used to elucidate clonal expansions of different pathogens based on variations within seven housekeeping genes highly conserved across the individual species. Vital to the conceptual development of MLST was the recognition that the bacterial population structure is not essentially clonal. Thus, the patterns of genetic exchange among bacteria and their descent could only be determined by analysing nucleotide sequence data from multiple locations of the chromosome [128-130]. Although MLEE provided a rapid and inexpensive approach to investigating bacterial populations' genetic structure, many research groups were beginning to directly sequence the genes encoding virulence factors or the enzymes utilised for MLEE [98]. DNA sequencing was advantageous over MLEE in that all the allelic variation at a particular locus could be detected, including the detection of intragenic recombination events.

Furthermore, nucleotide sequences availed themselves for the portable, unambiguous identification of alleles and facilitated inter-laboratory comparisons. MLST capitalised on the increasing use of sequence data to detect variation at the DNA level that was not apparent by previous approaches, such as serotyping and MLEE [127]. The initial application by Maiden *et al.* [126] involved using the nucleotide sequences of PCR-amplified fragments from 11 housekeeping genes drawn from a total of 107 globally representing isolates of *Neisseria meningitidis*.

MLST facilitates microbial genetic diversity analysis at a sufficient scale and can be applied to investigate several ecological issues within natural populations, such as clonality, recombination, gene flow, divergence, host and niche switching and adaptive evolution, among others [110, 131]. Given that ST numbers are arbitrary, it is possible to have several STs within a population related to each other [132]. This justifies the creation of clonal complexes, which clusters related ST variants in *E. coli* / *Shigella* using an eBURST approach [110, 132]. Thus, a clonal complex depicts many STs that have very recently diversified from a joint founder [132].

### 1.2.4 Designation of MLST sequence types for *E. coli*

In MLST, unique numerical designations are applied to sequence variants of each of the genes employed in a particular scheme. The seven-allele MLST comprises seven integers representing the alleles in seven housekeeping gene fragments [126]. There are currently three primary MLST schemes available for *E. coli* [133], with the corresponding databases hosted at Warwick Medical School, UK [110], Michigan State University, USA [134] and Pasteur Institute, France [111]. These are based on three different gene combinations: with only the *icd* gene in common. The rationale for the choice of genes within each of the MLST schemas is unclear, except that they are all housekeeping genes. The seven housekeeping gene sets across the three schemes vary in their nucleotide diversity, with the highest diversity observed among the genes used in the Pasteur Institute scheme [135] and the lowest diversity found among the genes in the Achtman scheme based at the University of Warwick [136].

### 1.2.5 Application of MLST to the study of *E. coli* populations

Reid *et al.* [134] were among the first to pioneer the application of MLST to study the population structure of *E. coli*. They considered the genetic relationship between many EPEC and EHEC isolates, using the housekeeping genes, *arcA, aroE, icd*, *mdh*, *mtlD*, *pgi* and *rpoS*. The phylogeny they obtained was similar to what had been found earlier with the ECOR collection [100], with the EPEC strains split into two distinct clades, designated EPEC-1 and EPEC-2. EPEC-1 clustered with the uropathogenic strain 536 (belonging to phylogroup B2), while EPEC-2 fell in phylogroup B1. The EHEC strains similarly separated into two clades, with EHEC-1 containing the O157:H7 strain and O55:H7 in phylogroup E, while EHEC-2 fell into phylogroup A, encompassing strain K-12 and isolates belonging to serotype O111:H8. These findings pointed to the parallel evolution of EHEC and EPEC pathotypes on multiple occasions, indicative of the possible acquisition of specific virulence factors via the acquisition of mobile genetic elements. Escobar-Paramo *et al.* [109] also developed an MLST scheme based on the seven housekeeping genes *trpA, trpB, pabB, putP, icd* and *polB*, which they applied to some 98 non-pathogenic and pathogenic strains selected to represent the commensal and pathogenic diversity within *E. coli* /

*Shigella*. The authors explored the relationship between the genetic background and *E. coli* pathovars' virulence genes and reported roughly the same phylogenetic groups as had been seen earlier. Crucially, they identified a new phylogroup, C (not the same as phylogroup C from the earlier MLEE studies), which occurred between phylogroups A, B1 and D. A striking observation from this study was that certain pathovars were restricted to particular phylotypes, which suggested that the expression and maintenance of virulence genes were associated with a specific genetic background. Notably, the pathovars associated with severe pathologies, such as EHEC, ETEC and *Shigella* / EIEC were restricted to phylogroups A, B1 and E and that all strains possessed virulence factors that were linked with mild and chronic diarrhoea.

Subsequently, Wirth *et al.* [110] published a third MLST scheme, involving the genes *adk*, *icd*, *fumC*, *recA*, *mdh*, *gyrB* and *purA*, which has become known as the Achtman scheme. Here, the authors selected a total of 462 isolates, representing both pathogenic and commensal strains from a wide range of hosts and geographical locations. Based on the phylogenetic tree they obtained from maximum-likelihood analysis (following the removal of recombinant regions), Wirth *et al.* [110] suggested that recombination was frequent in *E. coli* and challenged the view that the evolution of *E. coli* was clonal. This was evidenced by the detection of some hybrids between phylogroup A and B1 (termed AxB1 lineage) and others with multiple ancestry sources. The authors concluded that phylogenetic methods might be unsuitable for resolving *E. coli* strains' relationships, given widespread recombination. However, Chaudhuri and Henderson [54] point out that the findings in [110] may have been influenced by the arbitrary choice of genes in their scheme, in particular some genes, such as *gyrB*, may have been recombination hotspots and thus may not be representative of the whole genome. Another significant conclusion from the study by Wirth *et al.* was the observation of frequent recombination in *Shigella*/EIEC, which led them to posit a link between recombination and virulence ("sex and virulence" [110])—a pathogenic lifestyle would result in an increased exposure to host immune defences, which in turn would result in diversifying selection of escape variants and thus an increase in the prevalence of recombinant variants.

### 1.2.6 Emergence of core-genome MLST (cgMLST) and hierarchical clustering of cgMLST

A drawback of the seven-allele MLST scheme is the low resolution, in light of the high quantity of output from NGS [137, 138]. Moreover, since 2007, seven-allele clonal complexes have not been updated, as they were merging into each other (https://enterobase.readthedocs.io/en/latest/mlst/mlst-legacy-info-ecoli.html). However, the nomenclature within the classic (seven-allele) MLST scheme is easy to remember (for example, *E.*

*coli* ST131) and is thus commonly used by microbiologists and other public health partners [138]. Also, MLST classifications of the bacterial population still hold for many organisms, such as *E. coli* and *Salmonella* and roughly compare with what is achieved with serotyping [139]. Furthermore, the existence of well-established databases is an attraction of this method. A solution to the low discriminatory power of the classical MLST technique is to increase the number of genes or resort to single nucleotide polymorphisms (SNPs) based on the core genome as the informative sites to classify populations [139].  Consequently, core genome MLST (cgMLST) has emerged as a highly discriminatory and powerful tool for investigating the microbial population structure. cgMLST is based on the typing of up to 3,002 genes in the core genome (2,512 genes in *E. coli*) [139].

Zhou *et al.* [139] recently published an expansion to the ECOR collection to include 9,479 genomes, which they named the 'ECORPlus' collection, in homage to the work of Selander and colleagues [107]. Based on the EnteroBase integrated software environment, their work provides critical insight into the diversity of *E. coli*. EnteroBase contains 561,732 genomes assembled from Illumina short reads of *Salmonella, Escherichia/Shigella, Clostridiodes*, *Vibrio, Yersinia* and *Helicobacter* (as of 05 January 2021). As a result of missing data in draft genomes, almost every cgMLST is unique, making visual comparisons of cgMLST a rather complicated and laborious exercise. Consequently, the Hierarchical Clustering concept of cgMLST STs (HierCC) was introduced in EnteroBase to facilitate the study of population structures based on the cgMLST distances between genomes. cgMLST profiles for *E. coli* facilitate single-linkage hierarchical clustering according to fixed cgMLST allelic distances. Allelic distance matrices were calculated for all existing pairs of cgMLSTs at several levels: for example, HC0, HC1, HC5, HC10, up to HC2521. For *E. coli*, HC1100 was found to correspond to ST complexes based on the 7-gene MLSTs, while HC1100 corresponds to the seven-allele MLST clonal complexes [139]. Genomic relationships at HC5 to HC10 can be used to detect local transmission chains across genera. The cgST HierCC algorithm, therefore, lends itself as a handy and robust tool for analysing bacterial population structures at multiple levels of resolution. In a recent study of the population structure of *Clostridioides difficile*, Frentrup *et al.* [140] showed that HierCC allows closely-related neighbours to be detected at 89% consistency between cgMLST pair-wise allelic differences and SNPs.

Analysis of a maximum-likelihood tree reconstructed from the core SNPs derived from the 9,479 genomes in the ECORPlus collection confirmed clustering among the HC1100 groups within *E. coli* [139]. The other genera within *Escherichia* and the cryptic clades II-V, were similarly confirmed as distinct long branches of comparable lengths. The analytical tool, GrapeTree [141], incorporated within EnteroBase and available as a stand-alone package, allows for the analysis of cgMLSTs spanning thousands of genomes. The EnteroBase interactive platform presents an excellent and timely opportunity for performing comparative genomic studies with hundreds of thousands of *E. coli* genomes from around the world.

## 1.3 The taxonomy of *Escherichia*

In the pre-molecular era, several non-*coli* species were designated under the *Escherichia* genus based on DNA relatedness/hybridisation and overall phenotypic similarity:

- *E. blattae* (1973) [142],
- *E. fergusonii* (1985) [143],
- *E. hermannii* (1982) [144] and
- *E. vulneris* (1982) [145].

Lawrence *et al.* [146] utilised DNA sequences of the slowly evolving genes *gap* and *ompA* (encoding glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and outer membrane protein 3A respectively), representing conserved genes across eleven species of enteric bacteria, to map the phylogenetic relationships among the above mentioned five species. Their analysis revealed these species to be distantly related, rather than a monophyletic group. Apart from *E. fergusonii*, the other species were more divergent from *E. coli* than *Salmonella*. As *Salmonella* is traditionally considered most closely related to *E. coli*, the results justified the redesignation of *E. blattae, E. hermannii* and *E. vulneris* in alternative genera [146].

Subsequently, in 1991, Albert *et al.* [147] isolated a diarrhoeagenic isolate with EPEC-like phenotypic and genetic features from a nine-month old girl with watery diarrhoea in Bangladesh, initially designated *Hafnia alvei.* Subsequent identification and characterisation of five "*H. alvei*-like" isolates by DNA-DNA hybridization, phenotypic characterization and 16S rDNA sequencing led to their redesignation as *E. albertii* [148]. Members of this species were later found to be closely related to *Shigella boydii* serotype 13—a divergent lineage in *Escherichia*; the *E. albertii/Shigella boydii* serotype 13 lineage estimated to have diverged from an *E. coli*-like ancestor circa 28 million years ago [149].

Consequently, the genus *Escherichia's* taxonomy has been modified with the reclassification of *E. hermannii*, *E. blattae* and *E. vulneris* to other genera and the description of five cryptic clades within *Escherichia* [118, 150-152]. The genus is now comprised of three named species: *E. albertii, E. coli*, *E. fergusonii* and five cryptic clades, designated *Escherichia* clades I-V [48]. Of the three named species within the genus, *E. albertii* is the most divergent species, while *E. fergusonii* is closely related to *E. coli sensu stricto* [122].

The name *E. marmotae* has recently been validly published for Clade V [153, 154], although the species is not restricted to marmots [155, 156]. This builds on an earlier suggestion to classify clade V as a novel species and clades III and IV combined as a second novel species, based on digital DNA-DNA association and Average Nucleotide Identity (ANI) data [118] (an ANI of 95% between two genomes is generally taken as the standard for the demarcation of prokaryotic species [152, 157]). Furthermore, cryptic clade I strains possess the trademark virulence traits of *E. coli*, besides their potential to cause infections in humans [122, 158, 159]. Thus, *E. coli* and *Escherichia* clade I are now designated as *E. coli sensu lato* and the classic *E. coli* (phylogroups A-G) as *E. coli sensu stricto* [121]. Recently, Gilroy *et al.* [156] assigned the species name *E. whittamii* to clade II*,* in honour of the American bacteriologist, Thomas S. Whittam, for his contributions to the study of *E. coli*.

Walk and colleagues [118] estimated the lineages' divergence times that gave rise to each of the cryptic clades using a minimum evolution tree. Based on the assumption that *E. coli* split from *Salmonella enterica* between 100 to 160 million years ago, they estimated that the *Escherichia* lineages shared a common ancestor between 48-75 million years ago. They suggested that *E. albertii*, *E. fergusonii*, *Escherichia* clade II and clade V split between 38-75 million years ago, whereas *E. coli*, *Escherichia* clade I, clade III and clade IV split between 19-31 million years ago [118].

As evidenced by earlier efforts at classifying *E. blattae*, *E. hermanii* and *E. vulneris*, metabolic characteristics such as the utilisation of specific carbon sources or the production/catabolism of certain biochemical compounds are insufficient to delineate strains into species and species groups accurately. First, multiple genes are often required for the expression of a particular phenotype [146]. Second, convergent evolution in distantly related species confounds the delineation of species based on these phenetic characteristics. Similarly, DNA hybridisation fails to consider the relatedness among congeneric species reliably [146].

The Genome Taxonomy Database Toolkit (GTDB-Tk), a tool for the automatic classification of draft bacterial and archaeal genomes was published recently [160]. GTDB-Tk inputs genome assemblies in FASTA format and predicts the placements within domains based on identifying a set of 120 bacterial and archaeal marker genes and domain-specific reference trees. Then, species designations are computed using a GTDB reference tree, the relative evolutionary divergence parameter and ANI values. However, GTDB automatically assigns a random alphanumeric designation to novel species, which do not scale well to the increasing number of newly identified novel species and are often confusing and user-unfriendly [156]. To address this gap, Pallen *et al.* [161] recently developed an automated combinatorial approach to creating more than one million Linnaean binomials for Bacteria and Archaea.

## 1.4 The emergence of antimicrobial resistance in *E. coli*

*E. coli* has long been used as an indicator bacterium for monitoring the faecal contamination of food and water and antimicrobial resistance (AMR) in enteric bacteria of animals and humans [162]. For example, *E. coli* has been employed as the model organism in determining the level of AMR in bacteria from people in close contact with food animals, such as those who work at abattoirs and veterinarians—with the observation that such people harbour significantly higher levels of resistant *E. coli* compared to the overall community [163]. The ability to acquire and transfer AMR traits was recognised in *E. coli* as far back as the 1960s: 26% - 61% and 50% - 76% of human and pig multi-drug resistant *E. coli* strains respectively were found to transfer resistance to lab strains of *E. coli* K-12 in conjugation experiments [164-170].

According to the World Health Organisation (WHO), the ever-increasing levels of AMR is one of the most significant threats to human health (www.who.int/entity/drugresistance/en ), with significant economic implications [171]. Resistance to antimicrobials represents a daunting challenge to treating many infections, not least those caused by *E. coli*.

As a pathogen, *E. coli* is a prominent cause of urinary tract infections (UTIs), gastroenteritis and bloodstream infections, among others; and as such, antibiotics are frequently applied to treat *E. coli* infections [172-174]. The use (and abuse) of antibiotics in treatment is linked with antibiotic resistance development [175, 176]. Microorganisms exhibit a natural ability to resist antimicrobials' action—a phenomenon referred to as intrinsic resistance [177]. In addition, antimicrobial resistance can arise due to a gene mutation or the acquisition of resistance determinants via horizontal gene transfer [178]. Horizontal gene transfer can take the form of free DNA uptake (transformation), plasmid-mediated transfer of resistance gene traits (conjugation) or via phage-mediated transfer (transformation) [179, 180].

*E. coli* frequently displays resistance to multiple classes of antimicrobials (MDR)—with observed rates of MDR identified among strains that cause UTI and bacteraemia exceeding 50% [173, 181, 182]. The occurrence of MDR in *E. coli* has been focused on a small number of widely dispersed clones, mediated by extended-spectrum beta-lactamase (ESBL) and carbapenemase-encoding plasmids [183]. Worryingly, the rates of *E. coli* extraintestinal infections such as bacteraemia and UTIs have been rapidly increasing in recent years, due to an increase in the number of antibiotic-resistant infections caused by "superbugs" such as the ST131 clone [181, 184, 185].

Although several strains have been linked with the global emergence and dissemination of MDR in *E. coli* (for example strains belonging to ST88, ST410, ST648, ST405 and ST73), none equal ST131 in their extent [183]. The MDR ST131 clone belongs to a sublineage designated clade C which arose from two predominantly drug-susceptible clades by acquiring large MDR plasmids which confer ESBL and metallo-beta-lactamases encoding resistance to carbapenems and cephalosporins up to third generation; as well as point mutations leading to fluoroquinolone resistance [186, 187].

McNally *et al.* [184] recently analysed an extensive collection of more than 1,000 ST131 strains. They reported that the acquisition of colonisation and fitness factors and the accumulation of genes encoding dehydrogenase enzymes involved in anaerobic metabolism account, at least in part, for the successful expansion of this clone [184]. These results indicate hyper-resistant clones' ability to outcompete resident non-pathogenic strains of the same species and thus facilitate their long-term colonisation in the gut [183]. The fact that MDR strains such as ST131 are well suited to competitively colonise the gut suggests that antimicrobial resistance traits can be readily transferred from resistant strains to co-colonising susceptible strains, with the potential to hamper the treatment of future infections caused by such strains.

Accumulating data shows that ExPEC are frequently isolated from diseased companion animals and livestock—highlighting the potential for zoonotic as well as anthroponotic transmission [91, 188-192]. High rates of AMR among *E. coli* isolates from livestock and poultry have been documented, linked with the agricultural use of antimicrobials [193]. The use of antimicrobials as growth promoters in poultry feed was banned in Europe in 2006 [194], with a resultant sharp drop in AMR rates among isolates from livestock and poultry [193]. However, this may be less well controlled in other parts of the world, particularly in low-to-middle income countries. Worryingly, the usage of antimicrobials in developing countries is likely to increase as intensive farming practices are adopted [195].

Several studies have investigated the link between antimicrobial usage and antimicrobial resistance in animals and AMR in humans. In 2015, a systematic review [196] of the published literature to quantify the zoonotic transfer of ESBL-encoding extraintestinal *E. coli* infections found that:

- Six studies established the zoonotic transfer of AMR by whole bacterial transmission using molecular methods, in particular, via poultry in the Netherlands.
- Thirteen molecular studies suggested the mobile-genetic element-mediated transfer of AMR from animals to humans.
- Four observational epidemiological studies inferred the zoonotic transfer of AMR to humans.

Although the authors cautioned that their conclusions might not be geographically generalisable, it appears that a proportion of human extraintestinal infections caused by ESBL-encoding strains originate from food-producing animals. Further, *E. coli* strains from commercial broilers and backyard chickens have recently been reported to share resistance profiles with strains recovered from human extraintestinal infections [197]. This is worrying and highlights the need for further investigation, especially with whole-genome sequence-based studies [175].

### 1.4.1 Mechanisms of antibiotic resistance in *E. coli*

*E. coli* demonstrates resistance to antimicrobials through the following modes [175]:

- Modification/mutation of target sites (e.g., mutations in topoisomerase genes conferring resistance to quinolones)
- Enzymatic degradation of the antimicrobial agent (e.g., beta-lactamase genes conferring resistance to beta-lactam antibiotics)
- Active efflux or reduced cell permeability, leading to a decreased accumulation of antimicrobials (e.g., *Tet* efflux pumps conferring tetracycline resistance) and
- Tolerance to the antimicrobial agent due to novel insensitive antimicrobial alleles (e.g., tetracycline resistance arising from novel hydrofolate alleles).

### 1.4.2 Resistance to various classes of antibiotics

The resistance mechanisms employed by *E. coli* to various classes of antibiotics are discussed in the following section. For an excellent review on the subject, see [175].

#### 1.4.2.1 Beta-lactam resistance

Beta-lactams are probably the most widely prescribed antibiotics for a wide array of clinical indications, contributing up to 65% of the antibiotic market and about $15 billion in annual expenditure [198]. This class encompasses the penicillins, cephalosporins, carbapenems, cephamycins, monobactams and beta-lactamase inhibitors—with the common feature of members of this class being the presence of the beta-lactam ring (a highly reactive 3-carbon and 1-nitrogen ring) in their biochemical structure. The discovery of penicillin and its introduction into clinical use was immediately followed by reports of resistance in the 1940s and 1950s [199].

The production of ESBLs is the most common resistance to beta-lactams in *E. coli* and is considered the most common antibiotic resistance mechanism among Gram-negative bacteria, contributing to widespread resistance [200]. Over five hundred beta-lactamases have been described to date, either chromosomally-encoded or borne on plasmids [201]. These include different classes of the TEM, SHV, OXA, CTX-M, CMY and NDM enzymes, resulting in resistance to all the classes of beta-lactam antibiotics, including last-line antibiotics such as carbapenems and beta-lactamase inhibitors (e.g., amoxicillin-clavulanic acid and piperacillin-tazobactam combinations) in both food animals and clinical isolates [202-209]. Recently, isolates resistant to piperacillin-tazobactam but sensitive to third generation cephalosporins have been described, mediated by TEM overproduction [207]. Kot [210] recently reviewed the global prevalence of resistance among uropathogenic isolates and observed rates of resistance to amoxicillin-clavulanic acid of 3.1% - 40% in developed countries and 48% - 83% in developing countries.

Besides these ESBLs, AmpC beta-lactamases, chromosomally-located or acquired on plasmids, also hydrolyse cephalosporins up to third generation [211]. The induction or hyperexpression of chromosomal AmpCs (usually produced at deficient levels) leads to resistance [212]. Examples of AmpC beta-lactamase genes are the $bla_{CMY\text{-}like}$, $bla_{DHA\text{-}like}$ and $bla_{ACC\text{-}like}$ genes. Beta-lactamases are frequently associated with plasmids, integrons, insertion sequences and transposable elements, facilitating their dissemination among *E. coli* and between bacterial species [201, 206, 213].

### 1.4.2.2 Tetracycline resistance

Tetracyclines were developed in the 1940s and were the first antibiotics to be described as 'broad-spectrum' [214]. In 1953, the first tetracycline resistance was described in an *S. dysenteriae* isolate [214]. Tetracyclines have been used in human and veterinary medicine, particularly as a growth promoter in animal husbandry [215] but are not prescribed in children as they are deleterious to bone and teeth and never prescribed for *E. coli* infections. Nevertheless, tetracycline resistance in *E. coli* is widespread [216-222], suggesting the bystander effect on the commensal *E. coli* population from the treatment of other infections [223].

Tetracyclines are ribosome-binding protein inhibitors. The Tet family of efflux pumps are widely responsible for resistance to this class of antibiotics in *E. coli* [224]. Although over forty classes of tetracycline resistance genes have been described to date [225-227], only a few (e.g., *tet(A), tet(B), tet(C), tet(D)* and *tet(G)*) confer resistance in Gram-negative bacteria [228]. Of these, *tet(A)* and *tet(B)* remain the most common [216, 222, 229, 230], present in approximately 35% and 60% of *E. coli* isolates respectively [231]. Tet pumps' association with various transposons and plasmids have been documented [224], facilitating tetracycline resistance dissemination. For instance, *tet(A)* is associated with the transposon Tn*1721*, while IS*10*-bound Tn*10* can mobilise *tet(B)* [224]. Similarly, *tet(C)* and *tet(D)* have been associated with transposon-like structures, flanked by IS*26* genes (termed pseudo-compound transposons) [232-234]. Also, plasmids such as IncN [235], pBR322 [236], pCER1 and pCER2 [234], among others, have been associated with tetracycline resistance determinants.

#### 1.4.2.3 Aminoglycoside resistance

Aminoglycosides attack the ribosome; thus, resistance to this class of antibiotics involves ribosomal target site modifications via methylation or by chromosomal mutation, as well as active efflux and chemical modification by aminoglycoside-modifying enzymes [237]. The enzymatic modification of aminoglycosides is the most common resistance mode among clinical isolates [238-242]. Of the three classes of aminoglycoside-modifying enzymes that have been described (phosphotransferases, nucleotidyltransferases and acetyltransferases), acetyltransferases are the most common in *E. coli*, frequently associated with transposons and plasmids [239, 240, 243].

#### 1.4.2.4 Fluoroquinolone resistance

In *E. coli*, resistance to fluoroquinolone antibiotics is commonly mediated by mutations in DNA gyrase and topoisomerase IV, active efflux or decreased accumulation of the antibiotic [244]. Plasmid-mediated resistance to this class of antibiotics elicits only low-level resistance that tends to fall below the clinical breakpoint for resistance but facilitates the selection of higher-level resistance [175, 245]. Often, mutations in DNA gyrase (*gyrA*) occur with mutations in topoisomerase (*parE* or *parC*), resulting in highly resistant strains. All *E. coli* isolates are inherently capable of developing fluoroquinolone resistance, as the resistance determinants are chromosomally located. Fluoroquinolones are the drug of choice in the event of resistance to first-line antibiotics such as trimethoprim-sulfamethoxazole [246-248]. However, the increased use of fluoroquinolones has led to the emergence of increased resistance to this class of antibiotics among *E. coli* globally [247, 249, 250]. For example, data from the Czech republic's Olomouc region show a 7% and 5% increase in fluoroquinolone resistance among inpatients and outpatients respectively following increased consumption of fluoroquinolones from 2.52 daily dose per 100 bed days (DBD) to 4.29 DBD (inpatients) and from 0.14 daily dose per 1000 clients (DID) to 0.95 DID (outpatients) between 1997 and 2002 [250]. The NAUTICA (North American Urinary Tract Infection Collaborative Alliance) survey analysed 1,990 urinary tract isolates from 40 medical centres in the USA and Canada between 2003 and 2004 and found about 5% of the isolates to be resistant to ciprofloxacin and levofloxacin—the majority of fluoroquinolone resistance occurring in patients aged over 65 years [251]. Subsequently, between 2004 and 2005, data from the same surveillance program based on the analysis of 1,858 fluoroquinolone-resistant urinary tract *E. coli* isolates showed that 62% of the isolates were resistant against two or more other antimicrobial agents (multi-drug resistant) [252]. On the other hand, another survey of pathogen frequency and antimicrobial resistance, the SENTRY programme (on-going since 1997), in 2014 analysed a total of 3,537 isolates from paediatric patients from North America, Latin America and Europe and reported the susceptibility of *Enterobacteriaceae* to fluoroquinolones to be over 94% [253]. Simultaneously, data from the US indicates that the prevalence of fluoroquinolone resistance increased from 1% in 1998 [254] to 25% in 2012–2014 [255]. Kot [210] reports the rates of resistance to ciprofloxacin in developing countries to range from 56% to 86%.

#### 1.4.2.5 Sulfonamide-trimethoprim combination

Sulfonamides inhibit folic acid synthesis by targeting dihydropteroate synthase while trimethoprims targets dihydrofolate reductase; both enzymes are part of the folate biosynthetic pathway which is essential for the production of thymine and bacterial cell growth [256, 257]. Trimethoprim is widely used in combination with sulfonamides for treating UTIs, skin, respiratory and certain enteric diseases [258]. Two plasmid-borne genes, *sulI* and *sulII* mediate resistance to sulfonamides. On the other hand, resistance to trimethoprim is encoded by several plasmid-mediated dihydrofolate reductase (*dhfr*) genes, of which *dhfr-I* and *dhfr-II* are most common [259] and often associated with class 1 and class 2 integron cassettes [258, 260]. In the USA, national trimethoprim-sulfamethoxazole resistance in urinary tract *E. coli* isolates increased from 7% to 9% between 1989 and 1992, with a subsequent increase to 28% between 2009 and 2013 [261]. The resistance of uropathogenic *E. coli* to trimethoprim-sulfamethoxazole in developed countries is reported to be 15% - 37%, while that in developing countries ranges from 54% - 82% [210]. Chromosomal resistance to trimethoprim can occur via mutational changes in the intrinsic *dfr* gene [262]. Additionally, a mutation leading to the inability to methylate deoxyuridylic acid to thymidylic acid can result in low-level resistance [263]. Similarly, a mutational change in the dihydropteroate synthase gene, *folP,* can result in chromosomal sulfonamide resistance via an impaired affinity for sulfonamide [256].

### 1.4.2.6 Chloramphenicol and nitrofurantoin

Chloramphenicol is a broad-spectrum antibiotic used to treat several bacterial infections and is considered an essential medicine by the WHO (https://www.who.int/publications/i/item/WHOMVPEMPIAU2019.06). Due to toxicity concerns, chloramphenicol's use was substantially hampered in developed countries [264]; consequently, it remains active against several pathogens, particularly methicillin-resistant *Staphylococcus aureus* (MRSA) [265, 266]. Resistance to chloramphenicol is mediated by enzymatic inactivation by acetyltransferase, efflux pump activity and ribosomal protection [267-269]. Due to the decreased use of chloramphenicol in the developed world, resistance surveillance data are sparse. However, available data from developing countries indicate high resistance to this antibiotic:

- A survey of poultry isolates from different chicken farms in Taif, Saudi Arabia [270] reported 72% resistance to chloramphenicol, mediated by the *catA1/cmlA* genes. The authors concluded that this high rate of resistance reflected the use of antibiotics in agriculture. This was evidenced by very high resistance rates against several antibiotics, for example, oxacillin (99%), lincomycin (98%) and oxytetracycline (97%). Moreover, 99% of the 180 isolates analysed in that study were multi-drug resistant.
- On the other hand, a survey of antimicrobial resistance among poultry *E. coli* isolates from the formal and informal poultry farming sectors in South Africa [271] (a total of 264 isolates) reported a lower prevalence of chloramphenicol resistance (1.7% each) compared with that of aminoglycosides (41% v 32%), beta-lactams (20% v 45%), sulfonamides (22% v 27%) and tetracyclines (12% v 24%).
- A systematic review of chloramphenicol and florfenicol resistance among porcine *E. coli* isolates from China [272] showed resistance rates of 72% and 59% respectively, between 2000 and 2018. However, a rapid decline in chloramphenicol resistance was observed from 2012 onwards, following a ban on the veterinary use of chloramphenicol in 2002. Simultaneously, a survey of *E. coli* resistance among 103 healthy adults in Ho Chi Minh city, Vietnam reported 34% resistance to chloramphenicol and 43% of strains showed resistance to more than three antibiotic classes [273].

- A survey of antimicrobial resistance among *E. coli* from dogs and their owners in Shiraz, Iran between 2013 and 2014 reported resistance rates of 11% and 2% for chloramphenicol and florfenicol respectively, regardless of host.

Nitrofurans include nitrofurantoin and nitrofurazone and are usually prescribed to treat UTIs, particularly as first-line treatment for uncomplicated cystitis [175, 274]. Resistance to this class of antibiotics is generally low, thought to be the consequence of numerous action mechanisms produced by the reactive intermediates when attacked by nitroreductases [175]. A recent review of the global resistance rate of uropathogenic *E. coli* to nitrofurantoin shows that rates of resistance range from 0.9% (in the USA) to 13% (in India).

## 1.5 DNA sequencing

In 1977, two seminal DNA sequencing methods were published: Frederick Sanger's enzymatic dideoxy DNA sequencing technique (commonly known as Sanger sequencing) [275] and the Allan Maxam and Walter Gilbert's chemical degradation DNA sequencing method [276]. Around the same time (in 1979), Staden [277] described computer-based programs to analyse sequence gel readings and assemble sequences. Sanger sequencing relies on a DNA synthesis polymerase reaction with dideoxynucleotide chain terminators [278]. The initial protocol involved a quadruplicate reaction using the different base terminators, ddA, ddC, ddT or ddG and subsequent gel separation in separate lanes, with the nucleotide sequence at each position determined from the gel by the terminator base and fragment length [278]. The Maxam and Gilbert technique was based on the cleavage of terminally labelled DNA fragments at specific bases, followed by separation by gel electrophoresis [276]. In the 1970s, DNA sequencing was accomplished for organelles and small genomes like viruses (e.g., the genomes of cytomegalovirus and vaccinia (229 kb and 192 kb respectively)) [279] and involved substantial effort in the creation and subsequent mapping of lambda and cosmid libraries [280]. However, due to technical limitations and cost implications, bacterial genomes' complete sequencing was not achievable.

### 1.5.1 Whole-genome sequencing

The first complete bacterial genome sequences were published in 1995 for *Haemophilus influenzae* and *Mycoplasma genitalium* [281, 282]. Since then, we have seen an explosion in the number of sequenced bacterial genomes allowing large-scale genomic population studies of bacterial species such as *E. coli*. Consequently, our appreciation of bacterial evolution, function, biotic and abiotic interactions has been greatly enhanced as bacterial genome sequencing is now standard [283, 284]. We have seen the application of sequence-based information coupled with innovative bioinformatics tools used to resolve chains of transmission during outbreaks, as well as bedside applications, including the development of vaccines and drugs [283, 285]. These advancements have been made possible by the continuous improvements in sequencing efficiency, the decreasing costs of sequencing and expansive global sharing of sequence data [284]. We have witnessed three technological revolutions in bacterial genome sequencing: *(i)* whole-genome shotgun sequencing, *(ii)* next-generation or high throughput sequencing and *(iii)* single-molecule sequencing (long-read sequencing) (Figure 1.2). For an excellent review of the landmark scientific and cultural achievements covered under these revolutions, see [283].

#### 1.5.1.1 Whole-genome shotgun sequencing

The pioneering bacterial genome sequencing efforts in the 1990s involved a great deal of effort in creating and mapping large insert clones, from which small insert libraries were created and sequenced [283]. With the advent of whole-genome shotgun sequencing (a combination of Sanger sequencing and shotgun cloning) in 1995, bacterial genome sequencing was greatly simplified by the shotgun approach where the genome is sheared into many small fragments and the fragments are cloned and sequenced simultaneously, followed by electrophoresis using 96 or 384 well capillary machines. The output was then assembled into larger contiguous sequences using robust computer algorithms to produce a high-quality draft genome. More effort was required to finish these draft genomes, requiring a separate production line [284]. The development of bioinformatics tools such as Artemis and Glimmer for genome assembly was a fundamental breakthrough [286, 287].

The application of Sanger shotgun sequencing to several organisms—including model organisms like *E. coli* K-12 and *Bacillus subtilis* as well as fearsome human pathogens such as *Yersinia pestis*, *Mycobacterium leprae* and *Mycobacterium tuberculosis* [288-291] led to significant insights into pathogen biology and the identification of numerous novel genes. The availability of multiple genomes from the same genus or species also facilitated comparative genomics analyses and novel insights for organisms such as *E. coli, Campylobacter jejuni* and *Salmonella enterica* [289, 292-295]. A drawback of Sanger shotgun sequencing was that it was labour intensive [284]. It was also expensive: costing as much as $50,000 to produce a finished bacterial genome [296]. Furthermore, only the genome's clonable regions could be sequenced [283], meaning genes that proved toxic to the cloning host could not be sequenced [283].

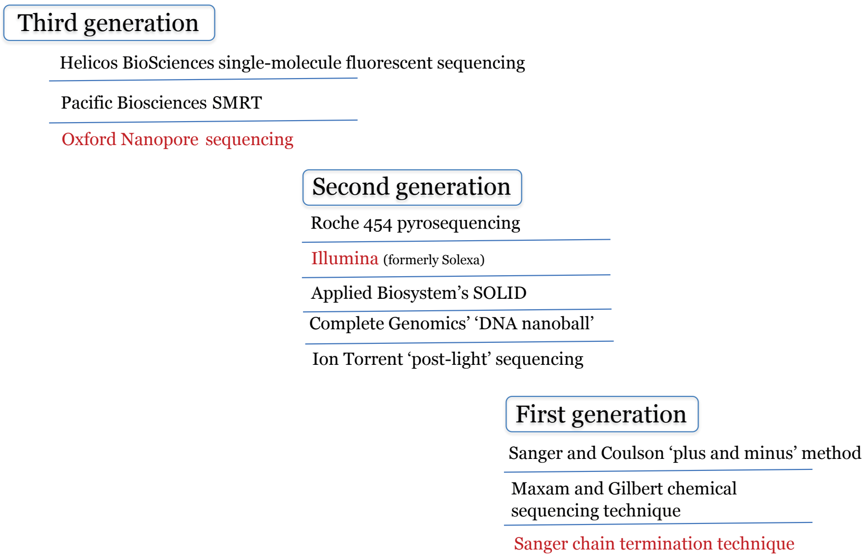

Third generation
Helicos BioSciences single-molecule fluorescent sequencing
Pacific Biosciences SMRT
Oxford Nanopore sequencing
Second generation
Roche 454 pyrosequencing
Illumina (formerly Solexa)
Applied Biosystem's SOLID
Complete Genomics' 'DNA nanoball'
Ion Torrent 'post-light' sequencing
First generation
Sanger and Coulson 'plus and minus' method
Maxam and Gilbert chemical sequencing technique
Sanger chain termination technique

***Figure 1.2:*** *DNA sequencing technologies: from first to third generation.*

### 1.5.1.2 Next-generation sequencing

Sanger shotgun sequencing is referred to as a first-generation technology [297]. 'Next-generation sequencing' was used to describe the second revolution in sequencing characterised by several technologies involving template preparation, sequencing and imaging, genome alignment and genome assembly (for an excellent review, see [297]). A significant difference between next-generation sequencing and first-generation sequencing is the use of chemistry for template generation instead of biological cloning, thus overcoming the limitations of the Sanger shotgun method (where non-clonable regions cannot be sequenced). Another significant advance was the substantial increase in throughput, generating more than one billion reads per instrument run [283, 297]. This era saw high-throughput sequencing applied to unravel disease transmission and screening pathogens for single nucleotide polymorphisms. This phase of sequencing evolution also coincided with the development of a number of bioinformatics approaches used in analysing bacterial diversity within patients, leading to the discovery that bacterial microevolution results in 'clouds of diversity' among closely related strains of a bacterial species within an individual host [137, 298]. A remarkable development in this sequencing era was the generation of metagenomic sequencing data [276], which has proved vital to, among others:

- Profiling the taxonomic composition of microbial communities [299],
- Elucidating the functional potential of microbial communities [300] and
- Recovering whole genome sequences from metagenomes without the need for culture [156].

Shotgun metagenomics involves the untargeted (shotgun) sequencing of all (meta) microbial genomes (genomics) present in an environmental sample.

Short-read sequencing has the advantage of high accuracy; however, a drawback of short-read technology is alluded to by its name: the short reads. Short read sequencers produce reads limited to a maximum of 600 bases [301], making it challenging to assemble genomes completely due to long repeats within the bacterial genome. Short read lengths also hampered the detection of large structural variations in genomes, e.g., large chromosomal insertions or duplications [283].

#### 1.5.1.3 Long-read sequencing

The limitations of short-read sequencing prompted the development of long-read sequencing. This era witnessed the advent of sequencing platforms capable of sequencing without the need for DNA amplification (required by first and second-generation sequencing). Unlike short-read technologies, long-read sequencing approaches do not involve chemical cycling for the addition of dNTPs. This sequencing technology also produces longer reads. Long-read sequencing is used to resolve long repetitive regions, structural variations and copy number alterations in genomes. When used in synergy with the high coverage of short-read sequencing, high-quality or even complete assemblies can be achieved.

Pacific Biosciences' RS II instrument can generate single polymerase reads of average lengths 10-15kb from a long insert library. This makes it ideal for *de novo* assemblies (discussed in Section 1.5.2.1). Nanopore sequencing has a potential niche in routine clinical diagnostics due to its ability to generate bacterial genome analysis data in real-time. Nanopore sequencers have the further advantage of transferring technology from ‘bench to bedside’. For example, Nanopore sequencing enables the rapid diagnosis of lower respiratory tract infections (clinical metagenomics) [302]. Deploying this technology to the field during outbreaks has significantly impacted how genomic epidemiological investigations are conducted. During an outbreak of *Salmonella enterica* in the UK, researchers capitalised on the rapid pathogen profiling abilities of the MinION (a sequencing device used for nanopore sequencing technologies) to identify the outbreak serovar within 50 minutes into the sequencing process [303]. Similarly, Nanopore technology has been used to rapidly sequence amplicon libraries of SARS-CoV-2 genomes to generate near real-time genomic and epidemiological analyses which were used to track healthcare-associated SARS-CoV-2 outbreaks in a hospital and community settings, thus rapidly optimising interventions [304].

Synthetic approaches incorporate modifications to library preparation that use barcodes for the computational assembly of large DNA fragments. They provide an advantage of deriving long-read information using short-read sequencing platforms. Synthetic long-read sequencing has found applications in the phasing of genomes, as demonstrated by Kuleshov *et al.* [305].

### 1.5.2 Analysing genome sequences

The term 'depth of coverage' is used to describe the average number of reads covering genomic positions (the mean value depicting the number of times each base is sequenced). It is often used as an indication of how good the overall quality of the genome sequences will be. A technical definition of coverage, according to Sim [306] states: "The theoretical or expected coverage is the average number of times that each nucleotide is expected to be sequenced given a certain number of reads of a given length and the assumption that reads are randomly distributed across an idealized genome. Actual empirical per-base coverage represents the exact number of times that a base in the reference is covered by a high-quality aligned read from a given sequencing experiment". The uniformity of coverage and sequence quality affects the accuracy of many downstream analyses including variant calling techniques [306].

#### 1.5.2.1 Genome assembly

Genome assembly describes the process of merging genomic sequences into longer contiguous sequences (contigs) in an attempt to reconstruct the original genome [307]. There are two approaches to sequence assembly:

1. *de novo* assembly and
2. Reference-based assembly.

*De novo* assembly involves reconstructing the genome without using a reference genome and assuming no prior knowledge of the length, composition or layout of the source DNA sequence [308]. Typically, reconstructed contigs do not span the entirety of the genome and assemblies will often include multiple unassembled regions. Popular tools for carrying out *de novo* assembly include SPAdes, MEGAHIT and Velvet (Table 1.5). The alternative approach is reference-based assembly, where short-reads are aligned to a reference genome and overlapping fragments are assembled into contigs. Minimap2 and the Reference-based Genome Assembly and Annotation Tool (RGAAT) are examples of reference-guided assemblers [309, 310]. The major drawback for reference-based assembly methods is largely due to the reliance on completeness of the reference genome, with regions that are unmapped to the reference being excluded. The choice of assembly strategy is informed by the intended biological application and other factors such as cost and time limitations [297]. For example, reference-based assembly is used to detect and catalogue genetic variation in several strains of highly related genomes, as was applied in *S.* Typhi [311]. This approach's strength is that all assembled genomes are directly comparable with each other as each is aligned to the same reference and thus against each other. The drawback of reference-based assembly is the requirement of a previously sequenced reference genome and prior knowledge of the taxonomic identity of the genome in question—making this approach of little use to novel genomes. Again, regions not found in the reference genome cannot be assembled—which hampers a study of the accessory genome.

Often, high-quality assemblies are achieved using a hybrid approach, where the advantages of short-read sequencing (high depth) and long-read sequencing (longer reads) are combined [284]. A widespread tool for generating accurate and complete assemblies of bacterial genomes using the hybrid approach is Unicycler [312]. Improvements in genome reconstruction have also been achieved using a reference-guided *de novo* assembly approach [313].

***Table 1.5:*** *Relevant tools for mining genomic data.*

| **Activity** | **Tool** | **Brief description** |
|---|---|---|
| Genome Assembly | Spades [314] | Assembly of short reads |
| | MEGAHIT [315] | Assembly of complex metagenomics data |
| | SSPACE [316] | Fast scaffolding of pre-assembled contigs |
| | Velvet [317] | Fast assembling of small genomes |
| | Miniasm [318] | Fast *de novo* assembly of long reads |
| | Canu [319] | *De novo* assembly of long reads |
| Mapping | Bowtie2 [320] | Reads alignment against a reference genome |
| | Burrows Wheeler Aligner (BWA) [321] | Mapping of short reads against a reference genome |
| | Minimap2 [310] | Mapping PacBio or Oxford Nanopore genomic reads against a reference genome |
| Genome Annotation | Prokka [322] | Annotation of bacteria, archaea and viruses |
| | Rapid Annotation using Subsystem Technology (RAST) server [323] | Annotation of bacteria and archaea |
| | GeneMark [324] | Gene prediction of metagenomes, bacteria, archaea, eukaryotes and, viruses |
| | MetaGeneMark [325] | Metagenomics gene identification |
| Pathogen screening | MetaMLST [326] | MLST calling |
| | Antimicrobial Resistance Identification by Assembly (ARIBA) | Identification of antibiotic resistance genes from assemblies and MLST calling |
| | PanPhlAn [327] | Metagenomics profiling at strain level, strain identification and characterisation, identifying strain diversity among hundreds of strains |
| Phylogenetic analysis | Roary [328] | Pan-genome building, identification of core and accessory genes |
| | FastTree / RAxML [329, 330] | Construction of phylogenetic trees from aligned nucleotides |
| | FigTree / TreeView / Forrester [331, 332] | Phylogenetic tree visualisation |
| | GrapeTree [333] | Interactive visualisation of large phylogenetic trees |

#### 1.5.2.2 Mapping and variant calling

As already discussed, a previously sequenced reference genome can often be used as a scaffold onto which we can map query sequencing reads [334]. This creates what is known as a 'pileup', whereby each read of high similarity is aligned against the reference genome and retained. Common tools to do this include BWA, SMALT, Stampy and Bowtie (Table 1.5).

Variant calling is done by determining whether the mapped reads align to the reference genome at each nucleotide position. Popular algorithms for variant calling include Samtools [335] and FreeBayes [336]. The alignment of the query reads at each position against the reference is carefully considered. Fair agreements between the aligned reads and the positions on the reference genome result in the nucleotides being called; regions that do not map to the reference, show high disagreements between reads, or insufficient read mapping are left undetermined. Reasons for the latter may include:

- Regions that are present in the reference genome but not in the target genome.
- Repetitive regions.
- Poor sequencing quality.

#### 1.5.2.3 Phylogenetic analysis

Phylogenetics aims at determining the evolutionary relatedness among genes, traits or organisms [337, 338]. A phylogenetic tree is a useful way to diagrammatically express the genetic relationship between a set of genomes (e.g. ancestry) and is widely used in published research [339]. The convention is to read a tree from the root, along the axis (x-axis) to the leaves. Closely related genomes will cluster together on the tree. The branch's length separating two genomes indicates the distance between them (i.e., how closely related they are to each other). The y-axis is arbitrary; thus, two genomes close to each other on the y-axis do not mean they are closely related [339].

Phylogenetic trees are based on computational algorithms and are traditionally reconstructed using genome alignment files as input—however, in recent years, several alignment-free methods such as the k-tuple and string-based distance measure methods have been published [340]. There are two approaches to phylogenetic trees:

1. Methods that assume no recombination events and
2. Methods that take into account recombination.

The former approach is usually applied to bacterial genomes for which recombination does not frequently occur, for example, *M. tuberculosis*, or in instances where recombination has been detected and recombinant regions removed. Examples of such methods include RAxML and FastTree (Table 1.5), MrBayes, RevBayes and BEAST [341].

Software based on the latter approach to phylogenetic trees include ClonalFrameML [342] and Gubbins [343]. Recombination in bacteria may occur as a gene conversion process or as a 'crossing-over-like' process. In gene conversion, the recipient cell contributes the bulk of the resulting genome of the recombination; the donor cell contributing only a short fragment. Methods such as ClonalFrame [344] and ClonalFrameML [342] take this into account and depict the recombined fragments on every tree branch. This is known as clonal genealogy. In the 'crossing-over-like' type of recombination, both parents contribute large amounts of DNA to the resulting genome. Here, phylogenetic reconstructions cannot depict clonal genealogy (recipient and donor cells for recombination events cannot be easily identified). The employed approach relies on identifying the breakpoints along the alignment where recombination occurred and representing the regions between breakpoints by separate phylogenies [341].

For microorganisms with high recombination rates, for example, *H. pylori*, phylogenetic trees are not suitable for depicting genome ancestry. Instead, algorithms exist that determine the number of ancestral populations (designated *K*) and individual samples are analysed as belonging to one or a mixture of the populations. Examples of such algorithms are STRUCTURE and ADMIXTURE [341].

Phylogenetic tree construction methods may also be categorised as character-based or distance based [338, 345]. Character-based methods include maximum parsimony and maximum-likelihood methods that compare all sequences simultaneously, considering one character or site at a time (e.g., RAxML) [329]. These approaches consider the tree with the best score, requiring the smallest number of changes to perform alignment and have the advantage of enabling the hypothesis about evolutionary relationships to be devised [345]. However, they are computationally intensive, time-consuming and do not scale well to very large datasets. On the other hand, distance-based methods rely on the distance (or dissimilarity) between all possible pair-wise sequences to construct trees. Examples include neighbour-joining methods such as NINJA [346] and have the advantage of being fast and suitable for large datasets. However, the conversion of pair-wise sequence alignment to distance data tends to lose information [338, 345].

#### 1.5.2.4 Pan-genome analysis

Following genome assembly, a pan-genome can be constructed using software such as Roary (Table 1.5). Alternatively, pan-genomes may be constructed from the raw sequences using tools such as Mauve [347]. This is critical for the analysis of genomic non-core/accessory regions. Roary builds pan-genomes using genes as its unit and is useful for exploring data across various diversities. Methods that utilise sequences have the advantage of providing information about non-coding regions, such as promoters.

Following the construction of the core and pan-genome, a gene presence and absence matrix can be plotted against a phylogenetic tree to visualise the ancestry of genomes. This plot also informs on gene gain and loss, as reported by Touchon *et al.* [348] in their study of the evolution of *E. coli*.

## 1.6 *E. coli* genomics

Frederick Blattner first conceived the *E. coli* K-12 genome sequence project in 1983 [349]; however, due to funding and technological challenges, it was not until fifteen years later that the project (which was six-year-long) was finally completed [289]. "*E. coli* was the obvious choice for a sequencing effort", noted Frederick Neidhardt in [350], since more was known about the organism than any other [351] and so much information had been gleaned from studies with *E. coli*, that, as Blattner stated, "Figuring out the microbe's genetic code would help integrate all those years of study" [350]. Blattner and colleagues submitted the final 2.0 Mb of the 4.6 Mb *E. coli* genome to GenBank on 16 January 1997 [350]. This was closely followed by the deposition in GenBank of an incomplete genome sequence of the closely related strain, W3110, seven days later [350] (the complete genome sequence of which was published in 2006 [351]). Based on atypical codon usage and base composition, Lawrence and Ochman [352] inferred 18% of the K-12 genome to have arisen from horizontal transfer (earlier studies had shown that horizontally-acquired genes exhibited atypical codon usage, base composition and dinucleotide frequencies [353, 354]). This estimate was later revised to 24.5%, using improved methods [355].

Next, the complete genome sequences of two EHEC O157:H7 isolates were swiftly published, expanding the number of *E. coli* genomes for comparative studies. First came the EDL933 strain isolated from Michigan ground beef connected to the 1982 multi-state outbreak by Perna *et al.* [356], then RIMD 0508992—the strain that caused a large outbreak in 1996 in Sakai city in Osaka prefecture, Japan, involving at least 6,000 schoolchildren [357]—the latter sequenced by Hayashi *et al.* [294]. Comparisons with K-12 revealed a shared sequence of 4.1 Mb, representing a common chromosomal 'backbone' of *E. coli*, with the remaining sequence comprising strain-specific clusters varying in size, encoding putative virulence factors, prophages and prophage-like elements [294, 356]. These analyses provided evidence of extensive horizontal gene transfer, with the description of many "K" and "O" islands—depicting introgressed DNA present only in K-12 but not in O157:H7 or only in O157:H7 respectively. The existence of a shared *E. coli* backbone was confirmed by a three-way comparison of the K-12 and O157:H7 strains with the genomic sequence of the third *E. coli* strain to be completed: the ExPEC strain CFT073 [295]. A surprising finding was that all three strains shared only 39.2% of the combined non-redundant set of proteins. The role of horizontal gene transfer in the pathogenic strains' evolution was also highlighted by the presence of several pathogenicity islands exhibiting atypical codon usage interrupting the common backbone [295].

With the availability of more *E. coli* genome sequences, the core and accessory genome concept was defined to represent a conserved set of roughly 2,200 genes common to all *E. coli* strains and strain-specific sequences respectively [348, 358, 359]. The core genomic sequence lends itself as a useful tool for the phylogenomic comparison of isolates, provided the effect of homologous recombination (estimated to affect about a tenth of the *E. coli* core genome [360]) is accounted for [358].

### 1.6.1 The pan-genome concept

The concept of a "pan-genome" was proposed by Tettelin *et al.* in 2005 [361], during an attempt to utilise genome sequence information from Group B streptococcus to predict proteins that might be exposed on the organism's surface and could be exploited as vaccine candidates [361, 362]. By this concept, the pangenome of each bacterial species is defined by three distinct components: namely, its:

- Core genome, representing the genes found in each isolate of the species,
- Accessory genome, depicting the genes present in several but not all isolates of the species and
- Strain-specific genes detected in one isolate only.

Through genomic comparisons of nineteen GBS isolates, Tettelin and colleagues uncovered the first evidence that closely related isolates differed significantly in their gene content. A single isolate of a particular species was insufficient to capture the species' genome [361, 362].

In a study that compared sixty-one sequenced genomes of *E. coli*, Lukjancenko *et al.* [363] predicted a pangenome comprised of 15,741 gene families, with only 993 (6%) of the gene families present in every genome (the core genome)—indicating an accessory genome of more than 90% in *E. coli*. This equates to an accessory or variable gene content of approximately four-fifths of any given *E. coli* genome [363]. These results corroborate those reported by Rasko *et al.* [359] and others before them [364, 365]. Rasko *et al.* identified a pangenome comprised of more than 13,000 genes via comparison of seventeen *E. coli* reference genome sequences encompassing human commensal and distinct clinical groups of *E. coli* and a core gene set of about 2,200 genes conserved in all isolates. It has become apparent that the more *E. coli* genomes are sequenced and compared, the more the pangenome continues to increase—what has been referred to as an 'open pangenome' [359]—and the core genome shrinks [55]. Thus, the accessory genome content contributes crucial insights which, coupled with single nucleotide polymorphisms (SNPs) in the core genome, can be employed to track the evolutionary history of natural isolates, as has been recently demonstrated by McNally *et al.* and others [366, 367]. The accessory genome content has arisen from repeated gene acquisition and the contemporaneous loss of sequences is thought to account for the distinctions between divergent lineages within the same species [365]. They include genes encoding virulence determinants, bacteriophages, virulence factors and acquired antimicrobial resistance determinants [55].

WGS offers many advantages for the diagnosis and understanding of the pathobiology of *E. coli* strains, in that it is possible to predict most of the subtypes of *E. coli* based on the presence of well-recognised virulence factors, as well as elucidating the full array of virulence factors possessed by individual strains within a particular pathotype. Furthermore, typing schemes that combine several genes within the core genome (e.g., core genome MLST available on platforms such as EnteroBase [139]) and others that would incorporate the accessory genome are expected to become the mainstay of *E. coli* analysis, particularly as the accessibility of whole-genome data continues to increase [55].

## 1.7 Ecology of *E. coli* in the gut

Microbial ecology is the study of the diversity, distribution and abundance of microorganisms and how microorganisms interact with each other and their environment to generate and preserve such diversities [368]. There are two areas of focus that have encapsulated microbial ecological studies to date: namely, the

1. Examination of microbial diversity—'who is there,' i.e., the identification and characterisation and estimation of abundance across a variety of niches.

2. Study of microbial activity, i.e., what microorganisms are doing—including their biotic and abiotic interactions and how they impact the observed diversity and the ecosystem.

Members of the gut microbiota have co-diversified with their hosts over millions of years and exist in a mutualistic relationship with their host, in which the gut microbiota carry out vital functions for their hosts and in return occupy a nutrient-rich environment [369, 370]. The gut microbial communities' composition is thought to be influenced by factors such as diet, physiology of the gut, host phylogeny and diet [371-374]. A healthy gut microbiota plays critical roles in developing the host immune system and is required for homeostasis in adult life [375, 376]. For example, the gut microbiota cells help maintain the balance between host metabolism and the immune system and in the large intestine, metabolise the indigestible components of the diet [19, 377]. The gut microbiota also detoxify toxic products and serve as a barrier against the colonisation of opportunistic pathogens (termed as colonisation resistance) [370, 378]. The mechanisms by which the resident intestinal microbiota elicit resistance against the colonisation and invasion of pathogens include [379, 380]:

1. The direct competition for nutrients,
2. The modification of metabolites such as bile salts and short-chain fatty acids that render them toxic to invading pathogens,
3. The alteration of pH and oxygen tension,
4. The induction of host antimicrobial peptides,
5. The expression of a dense mucous, IgA and cellular immunity,
6. Direct attacks through the production of bacteriocins or Type IV secretion.

However, intestinal microorganisms constitute a persistent invasion threat, considering their vast numbers and the large intestinal surface area [376]. Emerging evidence suggests that dysbiosis—defined as a shift in the composition of the intestinal microbiota and thus an alteration of the relationship between the host and the gut microbiota—is linked to the development of various diseases, such as inflammatory bowel disease, obesity, allergy, autoimmune disease and irritable bowel syndrome [370, 381-387].

Estimates show that anaerobic bacteria outnumber *E. coli* anywhere from 100:1 up to 10,000:1 [388]. The prevalence of *E. coli* in the various hosts they colonise varies widely (0-100%), being influenced by host characteristics such as body size, microbiota, diet and digesta retention times [2]. Over 90% of humans carry *E. coli*, while about 25-56% of wild mammals appear to be colonised by the organism [2, 389-392]. The prevalence rate in human-associated animals (such as chickens and cats) is estimated to be above 60% [392].

In the gut, *E. coli* reside in the mucous covering of the epithelial cells and is shed with the degraded mucus components into the intestinal lumen and subsequently excreted in faeces [393, 394]. Human faeces typically contain between $10^2$-$10^9$ colony-forming units (cfu) of *E. coli* per gram [390, 391, 395, 396], while an estimated $10^4$-$10^6$ cfu can be detected in the faeces of domestic animals [396]. Data on the quantity of *E. coli* in stools of wild animals is, however, lacking. As first observed by Escherich [397], *E. coli* is one of the first bacterial colonisers of the infant gut [398], achieving concentrations of up to $10^9$ cfu per gram of the stools of infants [390, 391]. Subsequently, anaerobic members of the microbiota expand and dominate the gut [399]. Given that *E. coli* is a facultative anaerobe, its ability to utilise oxygen probably helps create an anaerobic environment favouring the blooming of strict anaerobes [398]. As a gut microbiota member, *E. coli* produces vitamin K and mounts resistance against colonisation by pathogens [400, 401]. Thus, *E. coli* exists in a mutualistic relationship with the human host, although it is mostly described as a commensal [12, 402].

Given that most *E. coli* reside innocuously in the gut, an important ecological question that has been plaguing microbiologists is what makes *E. coli* an occasionally devastating pathogen [392, 402]? To address such questions requires an enhanced understanding of the ecology of the organism as a commensal. However, non-pathogenic *E. coli* have been traditionally underrepresented in ecological studies of this species [402]. More studies exploring the populations of resident or non-pathogenic *E. coli* within and between hosts and how these populations vary over time are needed to shed light on this evolutionary puzzle.

Humans are exposed to *E. coli* through multiple routes [403-412]:

- The consumption of contaminated food and water.
- Through fomites, for example, on bank coins and notes (particularly in the cracks of creased notes) and cell phones.
- Pets and domestic animals.
- The environment.

This high level of exposure is reflected in reports of more than one strain in normal stools [249, 395, 413-418].

Two plausible theories explain the fate of swallowed strains [395, 417, 418]:

1. **The displacement theory** suggests that newly ingested strains may fail to establish themselves, in which case they are voided out, or if they succeed in establishing themselves, will displace the 'resident' strains present.

2. **The dominant-minor strain** theory posits that freshly ingested strains do not replace the established strains within the gut, but co-exist as minority or transient strains, albeit in small numbers and may be detected from time to time in the stool along with the dominant strain.

Experimental studies with the Nissle 1917 strain have shown that not all strains are equal in their propensity to establish themselves following immigration [419]. There is evidence to suggest that repeated exposure may facilitate immigration and establishment of strains [189, 420]. The co-existence of multiple strains in a single host raises an essential question about the factors that govern residency in the gut, i.e., how incoming strains overcome the colonisation resistance posed by the existing *E. coli* population. The current body of evidence suggests the following:

1. **Freter's successful competition hypothesis**. Freter theorised that successful colonisation occurs due to successful competition for nutrients [421-423]. Accordingly, the gut microbiota composition is determined by several limiting substrates, which different members of the microbiota can utilise with variable efficacy. Conway and colleagues demonstrated this principle in *E. coli* strains (*E. coli* strains HS and Nissle 1917 vs *E. coli* O157:H7) using carbohydrate metabolism in a mouse model [424, 425]. This theory is in tandem with Gause's exclusion principle, which precludes two organisms' co-existence if they share the same limiting resource [426]. Iron competition appears to influence the colonisation of resident strains as demonstrated by studies which found strains lacking siderophore genes to have a reduced ability to establish themselves in mouse models, compared to wild-type strains. Conversely, resident strains in the human gut have been found to encode siderophores [427], signalling their potential contribution to successful colonisation. The mechanisms by which two organisms exclude each other from a particular niche might be through either direct (for example, through bacteriocin production or phage to damage or kill competitors) or indirect (for example through passive resource utilisation) competition [423, 428, 429].
2. **Efficient utilisers.** Some colonisers thrive because they utilise available nutrients much more efficiently than others who use the same nutrients. In murine studies, mice fed with *E. coli* MG1655, non-motile *flhD* mutants were found to persist in stools collected three days post-feeding and dominate the population (roughly 90%) by day 15. The mutants colonised better than the wild-type parent strain and grew in the caecal mucus faster than the wild type counterparts [430, 431]. Further analyses using high-throughput genomic approaches revealed that the *flhD* mutants possessed an enhanced ability to oxidise several carbon sources because the loss of FlhD conferred an increased expression of genes involved in carbon and energy metabolism. (The *flhDC*

operon encodes the FlhD4C2 regulatory complex, which has been shown to negatively regulate the genes involved in galactose transport and the citric acid cycle while positively regulating the genes involved in ribose transport [430, 432]).

3. **Restaurant hypothesis**. Leatham-Jensen *et al.* [433] have described how polysaccharide-degrading anaerobes break down polysaccharides into sugars, which they serve to *E. coli* cells within a shared biofilm [434]. This is an example of 'syntrophy'—an association where one organism feeds on the nutritional products of another [435]—as has been observed between *Bacteroides ovatus* and *B. thetaiotaomicron* [436]. The biofilms that feed these *E. coli* strains are referred to as "restaurants". Many of such restaurants are thought to exist, proposed to comprise a mix of different commensal strains, each serving different nutrients to the commensal *E. coli* strains residing therein [424].
4. **Different nutritional requirements.** This hypothesis suggests that pathogenic strains may utilise different nutrients than those of the commensal residents. For example, it has been shown that *E. coli* strain HS*, E. coli* Nissle 1917*, E. coli* MG1655 and *E. coli* EDL933 utilise unique metabolic niches in the mouse intestine [424, 425, 437]:
    - *E. coli* HS uses six out of the twelve sugars available in the mucous layer, namely, arabinose, galactose, gluconate, lactose, ribose and N-acetylglucosamine.
    - *E. coli* MG1655 uses five sugars, namely, fucose, arabinose, gluconate, N-acetylneuraminate and N-acetylglucosamine.
    - *E. coli* Nissle 1917, on the other hand, utilises the following seven sugars: fucose, galactose, arabinose, N-acetylglucosamine, gluconate, N-acetylneuraminate and mannose.
    - Lastly, *E. coli* EDL933 also uses a unique selection of carbon sources, namely, galactose, arabinose, hexuronates, N-acetylglucosamine, mannose, sucrose and ribose.
5. **Different biogeographical niches.** Pathogenic strains of *E. coli* may colonise different biogeographic regions of the gut from commensal strains. Scientists from the University of Oklahoma have shown that *E. coli* EDL933, a strain of enterohaemorrhagic *E. coli* O157:H7 colonises a different niche in the mouse intestine to human commensal strains [425, 438]. In the bovine host (where *E. coli* O157:H7 commonly colonizes the gut innocuously), studies in 12-month-old naturally colonized steer revealed a unique tropism of *E. coli* strain O157:H7 for the rectal mucosa adjoining the recto-anal junction [439].

6. **Hierarchical nutrient utilisation.** Here, closely related bacterial species may co-colonise the gut by hierarchical utilisation of similar nutrients, as has been demonstrated between two generalists, *B. ovatus* and *B. thetaiotaomicron* [436]. Thus, direct competition for substrates is excluded by each species utilising the available glycans with differing priorities (i.e., order and speed of consumption).

### 1.7.1 Exploiting colonisation resistance for therapy

Antibiotic treatment has significantly reduced mortality and morbidity associated with potentially life-threatening diseases and, thus, saved millions of lives [440]. However, broad-spectrum antibiotics rarely target pathogens alone and concomitantly result in deleterious effects on the commensal bacterial populations, resulting in an increased susceptibility to infections due to alteration of the host microbiota (a phenomenon termed as dysbiosis) [441-445]. This problem is compounded by the fact that many pathogens have become increasingly antimicrobial resistant. Consequently, the administration of live bacteria (probiotics) to restabilise the altered microbiota and restore the colonisation resistance conferred by the resident microbiota has increasingly been the focus of intense research [440, 441, 443, 446]. This heightened interest has been fuelled by the success of faecal transplantation in patients suffering from diarrhoea associated with *Clostridioides difficile* infection [447-449].

Probiotics are defined as living microbes of human origin which can colonise host sites such as the gut and oropharynx when ingested in adequate quantities and thus deliver benefits to the host [446, 450]. Probiotic strains are attractive as preventive measures against gut infections. Bacterial candidates that have been proposed to promote high-level colonisation resistance to infection include members of the lactate-producing genera, e.g., *Lactobacillus* [451] and *Bifidobacteria* [452]—although evidence of their effectiveness in health in humans, reducing infections or fostering longevity is sparse [440].

*E. coli* Nissle 1917 (also known as Mutaflor) is one of the most extensively researched probiotic strains globally [453]. Alfred Nissle observed that this strain ameliorates ulcerative colitis [454], following his isolation of the strain in 1917 from the stool of a soldier who did not suffer from diarrhoea like his comrades did during the First World War [455, 456]. The available data shows that Mutaflor possesses many traits that favour its suitability as a probiotic strain. A few examples of these include:

- The strain forms a biofilm well and in this way, outcompetes numerous pathogenic and non-pathogenic strains such as EPEC and ETEC [457].

- Complete genome sequence information revealed that Mutaflor contains several fitness traits such as adhesion factors and an array of iron uptake systems that enable it to outcompete other bacteria and block the adherence and invasion of pathogenic strains [458-460].
- The strain also boosts intestinal barrier function and protects against epithelial disruption by EPEC [461].
- *E. coli* Nissle 1917 induces the release of beta-defensin-2 (an antimicrobial peptide) from human epithelial cells, thus eliciting a broad antimicrobial response against both Gram-positive and Gram-negative bacteria, as well as fungi and viruses and, thus, inhibit invasion and colonisation of other bacteria [462].

Accumulating evidence suggests that probiotics comprised of a combination of strains are effective, owing to the symbiosis among strains—including those from different genera [463]. An example is the VSL#3 multi-strain probiotic consortium—comprising *Eubacterium faecium, Streptococcus thermophilus, Lactobacillus acidophilus*, *Bifidobacterium breve*, *Bifidobacterium infantis, Lactobacillus delbrueckii subspecies bulgaricus*, *Bifidobacterium longum*, *Lactobacillus plantarum* and *Lactobacillus casei*—which has been shown to treat ulcerative colitis effectively [463, 464]. Researchers working with *E. coli* EDL 933 found a combination of two strains—*E. coli* Nissle 1917 and *E. coli* HS, both promising probiotic candidates—can overlap nearly all the nutrient requirements of *E. coli* EDL 933 [465]. However, the same combination of commensals could not prevent colonisation by uropathogenic strain, *E. coli* CFT073 and the enteropathogenic strain, *E. coli* E2348/69. Different pathogenic strains may thus occupy distinct nutrient niches within the gut microbiome.

As interest in probiotics heightens, there is the need to understand the mechanics of how potential probiotic strains compete or co-exist with the commensal bacteria in the gut. We do not yet fully understand how probiotic strains interact with the usual residents of the gut. Previous attempts at unravelling the interactions between pathogens and commensals in the gut have been mainly carried out using well-characterised reference strains within in-vitro environments. It will be desirable to test these hypotheses with the 'real world' strains in the natural habitat (i.e., the gut).

## 1.8 Within host bacterial diversity

As discussed in Section 1.7, continuous exposure to *E. coli* via multiple routes contributes to the turnover of strains in the vertebrate gut, thus resulting in considerable diversity among the population of *E. coli* that exists within a single individual and between different hosts. Besides immigration events, the diversity of *E. coli* is enriched by within-host evolution events [137].

Members of a bacterial community undergo changes as they interact with each other and their hosts or environment. The sources of novel variation that account for within-host evolution may be:

1. Point mutations, involving the substitution, insertion or deletion of a single nucleotide. Point mutations represent the smallest evolution unit [466]. A within-host point mutation of approximately 1 per year per genome has been reported for *E. coli* [407].
2. Insertion and deletion events up to 1000 bp and the uptake or loss of plasmids and bacteriophages (mobile genetic elements) [466, 467]. For example, phylogenetic analysis of the O104:H4 strain that caused the large outbreak of gastroenteritis and haemolytic uraemic syndrome in Germany in 2011 revealed that the acquisition of a Shiga toxin 2-encoding prophage and an extended spectrum beta-lactamase CTX-M-15-encoding plasmid contributed to its emergence [468-471]. Similarly, the acquisition of virulence gene-encoding plasmids into many ancestral *Shigella* spp. was crucial to the evolution of *Shigella* as human pathogens [53].
3. Recombination and horizontal gene transfer (HGT) (discussed below).
4. Genomic rearrangements (discussed below).

HGT involves the uptake of extracellular DNA (transformation), cell-to-cell transfer of genetic material via surface appendages (conjugation) or viral import (transduction) from genetically distant relatives. The acquisition of genetic sequences from unrelated organisms via HGT results in faster diversification of the genome compared to point mutation alone [467]—consequently termed ‘evolution in quantum leaps’ [468]. Fragments of the chromosomal genome can be replaced with homologous sequences from another cell through homologous recombination, which plays a significant role in the evolution of the *E. coli* [469]. In particular, the presence of mixed infection facilitates homologous recombination by providing material for import in the chromosome—as has been exemplified in *H. pylori*, where evolution is noted to accelerate up to a 100-fold in the presence of mixed infections [470-472]. Thus, homologous recombination is a strong driver of within-host evolution. Alternatively, non-homologous sequences can be gained and incorporated into the genome—a phenomenon that is compensated by genome degradation [473].

Bacterial genomes can also undergo rearrangements during DNA recombination, replication or error-prone DNA repair [474, 475]. Genomic rearrangements alter chromosomes or large chromosomal regions. They involve the processes of deletion, duplication, insertion, inversion or translocation. Homologous recombination leads to the reassortment of genes between chromosomal pairs; however, the genome's arrangement remains unchanged. Other forms of recombination may result in rearrangements of genomic DNA. DNA rearrangements contribute to gene expression and function and may contribute to genetic diversity [475]. Other factors that shape within-host diversity include genetic drifts: random processes by which allelic frequencies change over time due to birth and death of individuals within the population [476, 477], as well as by both purifying and diversifying selection [137].

### 1.8.1 Hypermutators

Besides spontaneous mutations, genetic variation can arise due to the breakdown of DNA repair mechanisms such as the mismatch repair system, producing bacteria with decreased replication fidelity [478]. In particular, hypermutation is advantageous during infection when the ability to adapt quickly can facilitate evasion of host immune defences and antimicrobial therapy [478]. Hypermutator strains have been identified during an MRSA outbreak in a neonatal intensive care unit [479] and among *S. aureus* isolates from cows suffering from benign forms of mastitis [480]. Strikingly, in one study where cystic fibrosis patients with chronic respiratory problems were followed up over thirty-eight years, nearly half of them harboured hypermutator *Pseudomonas aeruginosa* isolates [481]. Of note, the accumulation of SNPs via hypermutation can hamper the analysis of transmission and result in inaccurate conclusions [478]. In particular, the use of SNP thresholds to identify and delineate transmission lines could potentially exclude hypermutator strains or the inclusion of such strains may bias results [478].

### 1.8.2 Investigating within-host bacterial diversity

To study within-host bacterial evolution in detail requires the application of whole-genome sequencing to:

- Multiple clinical samples taken from an individual host. These samples may be collected longitudinally or simultaneously, from a single body site or several sites.
- Multiple isolates or genomes derived from a single clinical sample.

In its simplest form, two genomes or more can be compared to each other by counting the number of positions where they differ. Within-host diversity is determined by comparing two or more genomes from the same host [482]. Comparing two genomes from different hosts lends insight into chains of transmission. These approaches have been successfully applied to several organisms beside *E. coli* [99, 100] [472, 483-490].

Genotyping multiple isolates per host can also facilitate the identification of multiply infected individuals, representing the carriage of bacterial sub-populations descended from a distinct founder strain. Multiple concurrent infections are clinically crucial for the evolution of pathogens, particularly in facilitating recombination between divergent strains as discussed above and hence, the emergence of novel genotypes [491]. Clinically, multiple infections are known to affect disease progression. For example, in HIV infection, the disease is reported to accelerate in the presence of dual infections [492]. Two plausible scenarios could explain the origin of multiple infections. In the first instance, two distinct strains could be transmitted simultaneously from an external source which harbours sufficient strain diversity. Alternatively, the two strains could have arisen from separate sources at different times.

### 1.8.3 Studies of the within-host diversity of *E. coli*

Studies of the genetic within-host diversity of *E. coli* in the human gut date back to well over a hundred years ago and have involved a wide variety of methodologies—encompassing both microbiologic and molecular techniques (Table 1.6). Earliest studies involved the use of serotyping and subsequently MLEE. Later studies employed MLST and WGS to characterise strains and infer within-host diversity and evolution of strains.

***Table 1.6:*** *Summary of studies investigating within-host diversity of E. coli since 1899.*

| Reference | Key findings |
|---|---|
| **Pre-PCR era (by serotyping and MLEE techniques)** | |
| [80] | In this pioneering study (reviewed in [79]), Smith analysed forty-eight isolates obtained from four separate cultures of the stools of one normal infant and reported a high degree of temporal antigenic stability. This study documents the first evidence of temporal changes in the population of *E. coli* in the human gut. |
| [78] | Totsuka (reviewed in [79] and [402]) self-collected and cultured his own stool once a week for twelve weeks and recovered up to 32 isolates per culture, yielding a total of 332 isolates. Upon agglutinating these against antisera which he had prepared against isolates obtained from the first two cultures, he provided evidence of gradual shifts from one antigenic type to the other over the course of time in the intestine of one individual. |
| [79] | Wallick and Stuart isolated 650 *E. coli* strains from one individual over the course of 14 months and described four distinct serotypes that persisted for a month at a time—including co-occurrence of dominant and minority strains. They also inferred household transmission based on antigenic identity (serotype). |
| [417, 418] | Sears, Brownlee and Uchiyama investigated changes in *E. coli* serotypes over time in five adults sampled over a period of three and thirty months. They were the first to coin the term "residents" and "transients" to describe strains of *E. coli* "which establish themselves firmly and continue to multiply over extended periods of time" and those that persisted for only a few weeks at most respectively. Also investigated for the first time, the production of bacteriocins by resident clones to ward off the invasion of other clones. |
| [81, 493] | Robinet confirmed the periodic fluctuations of antigenic types in six healthy individuals followed up over the course of six months and concluded that host antibody production does not explain the turnover of resident strains of *E. coli* in the gut. Robinet was involved with further work which investigated bacteriocin production by the resident clones by testing activity of strains collected in one month against those recovered the previous month. They concluded that the turnover of strains was lower when the resident strains produced colicins. |
| [415] | Shooter et al investigated the serotype dynamics of *E. coli* isolated from nine adults over a three-month period. They combined H- and O-antigen testing for the first time and observed that several strains with an identical O antigen in fact displayed different H antigens. |
| [96, 104] | The authors sampled *E. coli* from a single individual over the course of eleven months and using multilocus enzyme electrophoresis, provided seminal insight on the genetic structure of the resident and transient clones in the healthy adult gut. Importantly, they documented the loss and acquisition of plasmids in identical clones over time, as well as the gain and loss of novel clones over the study period as important contributors to the generation of *E. coli* diversity in the gut. |
| [494] | This study explored the dynamics of *E. coli* carriage in a group of individuals residing at British Antarctic Survey research base over the course of twenty-six weeks. Using MLEE, Tzabar *et al.* provided evidence of strain-sharing among the members of the research station—although several isolates were lost to viability, hampering the study conclusions on the clonal turnover over time. |
| **PCR era (by MLST and whole-genome sequencing-based phylogenomics)** | |
| [189, 420, 495] | These studies explored the diversity of *E. coli* among adults, children and companion animals living in the same household over time. Key findings include the detection of clones that appeared to persist over a period of four to forty-five weeks, as well as documenting evidence of strain sharing among members of a household and their pets. Shared strains were found to be of phylogroups D and B2, with an ST73 strain (phylogroup B2) causing UTI in a family dog. |
| [496] | Martinson et al studied eight healthy adults via biweekly sampling for a period of six months to two years, picking up to ninety-five colonies per sample. They observed that the resident clones often belonged to phylogroups A, B2 and F, including multiple resident clones persisting for more than a year in one individual. |
| [497] | This temporal survey of *E. coli* residency in fifty-four mountain brushtail possums sampled on four occasions over the course of a year found that the resident stains mostly belonged to phylogroup B2. The authors also concluded that *E. coli* was rapidly gained and lost among these non-human mammals, with just 36% of resident strains recovered at one timepoint recovered at the subsequent timepoint. |
| [498] | Stoesser and colleagues applied multiple colony sampling (sixteen colonies from each of eight faecal samples) followed by whole-genome sequencing to the investigation of the transmission of *E. coli* strains harbouring expanded-spectrum beta-lactamase genes collected in Cambodia. The authors reported substantial core- and accessory genome diversity, with a median of four STs recovered per individual. Remarkably, different clones from a single individual tended to share the same *bla*$_{CTX-M}$ variant and identical clones were found with different *bla*$_{CTX-M}$ variants. Significant accessory genome diversity was also observed within and between clones—highlight the utility of multiple colony sampling and whole-genome sampling in the analysis of *E. coli* within-host diversity. |
| [499] | Li *et al.* sampled more >100 colonies per caecum sample from nine mcr-1-positive broiler chickens from three provinces in China (a total of 962 *E. coli* isolates). A high rate of co-colonisation (three to nine STs per chicken) was observed, with several birds harbouring one to five Inc type plasmids encoding mrc-1—this high level of heterogeneity facilitating the transmission of mcr-1 among these chickens. The authors concluded that the "gut is a 'melting pot' for active horizontal transfer of the mcr-1 gene". |
| [500, 501] | Knudsen *et al.* collected *E. coli* (and other enterobacteria) from the stools of children with cystic fibrosis or cancer along with matched healthy controls over a nine-month period, picking up to five colonies per faecal sample. Children with cystic fibrosis or cancer received antibiotic treatment over the course of the study. Of the isolated *E. coli* (90% of children with cystic fibrosis, 93% of children with cancer and 94% of healthy controls), the prevalence of antibiotic resistant enterobacteria did not significantly differ between the children who received antibiotics and the healthy controls at the start and at the end of the study, suggesting that the level of antibiotic resistance they observed arose from the community.<br><br>As a follow-up study, seven isolates collected from three consecutive samples collected from one child with cystic fibrosis was characterised by WGS. Here, three distinct strains were detected, and the different strains found to harbour Inc1 plasmids encoding *bla*$_{CTX-M-1}$. However, the plasmids from the three different strains were found to differ by only a few SNPs and varied with limited regions, suggesting recombination events. This study documents horizontal transfer of *bla*$_{CTX-M-1}$-harbouring plasmids with a single individual. |
| [502] | Stegger *et al.* analysed twenty *E. coli* isolates from each of nine urine samples collected from nine women who presented with UTI at a general practice in Zealand, Denmark, from which a single clone was detected in eight out of the nine samples. The authors then selected a total of forty isolates belonging to the same clone: ten each from two urine samples and two rectal swabs isolates from two healthy individuals (collected in a previous study [503] and investigated the intra-clonal diversity of *E. coli* among the commensal and uropathogenic strains. A low intra-clonal diversity was observed for each clone, in both the commensal and pathogenic strains (0-2 non-synonymous SNPs). The authors reached an interesting conclusion, stating that "sampling of one colony would be enough for surveillance, outbreak investigations and clonal evolution", although there is overwhelming evidence to support the opposite. Evidence from their own study showed that among other clones, a variation in gene content of 2-15 genes was detected for all clones—which would have been unnoticeable had they isolated only a single colony from each sample. |

## 1.9 Study rationale and objectives

It has been said: "All cell biologists have at least two cells of interest: the one they are studying and *E. coli*" [504]. Yet, in the 135 years since *E. coli* was first described, most studies on the ecology of *E. coli* have been biased to pathogenic strains. Only a handful of studies focused on the diversity of *E. coli* populations in the healthy human gut, particularly in the post-PCR era. Even fewer studies have explored the within-host diversity of non-pathogenic *E. coli* in non-human vertebrates in the post-PCR era, leaving much to be learned in terms of the resident populations of *E. coli* in the healthy vertebrate gut. However, such studies are vital to our understanding of the evolution of *E. coli* in health and disease of both humans and animals.

An exploration of the commensal population of *E. coli* in non-human vertebrate hosts such as poultry and non-human primates is timely and relevant for several reasons:

1. Emerging infections are often linked with pathogens (such as *E. coli*) that inhabit both humans and non-human vertebrates and can cross the species barrier, potentially causing zoonotic as well as anthroponotic infections. Understanding the diversity within the commensal population would provide vital insights into the evolution of pathogens and antimicrobial resistance within this ecological context.
2. Understanding the within-host evolution of *E. coli* during health may inform new therapeutic approaches that exploit our understanding of the gut microbiome, for example, faecal microbiome transplantation, probiotic development and the control of foodborne diseases.

In particular, *E. coli* diversity studies based on whole-genome sequence data from sub-Saharan Africa are scarce. Very few studies have examined the population of *E. coli* among healthy individuals in this setting, particularly in healthy children. However, given the increased exposure to the environment, unsanitary conditions and proximity to animals especially in rural areas, such studies are likely to yield crucial insights into the dynamics of strain turnover and residency of *E. coli* in the gut, ultimately facilitating our understanding of infection by this organism. Furthermore, *E. coli* diversity in poultry, particularly backyard chickens, which are reared in most households in Sub-Saharan Africa or that in non-human primates which are human habituated in some areas within this setting and frequently come into contact with humans, remain largely unexplored. The recent advances in Illumina sequencing platforms twinned with the potential of long-read sequencing (Oxford Nanopore technology) provides a timely opportunity to apply these approaches to these lines of study.

### 1.9.1 Research questions

The critical question driving this thesis is: Given that pathogenic and drug-resistant *E. coli* strains have emerged as global challenges that fail to respect boundaries between species or countries, how do these pathogenic and or drug-resistant strains colonise the human or animal gut and cause disease in the face of competition and colonisation resistance from commensal (or probiotic) strains of *E. coli*? Moreover, how might we exploit a better understanding of microbial ecology to strengthen such ecology-driven competitive anti-pathogen effects to counter these threats? From this flow the following subsidiary questions:

1. How many kinds of *E. coli* does each human or animal carry in their gastrointestinal tract at any one time?
2. What is the population structure and phylogenomic diversity of *E. coli* in Gambian non-human primates?
3. What is the burden of AMR among the population of *E. coli* that reside in the gut of non-human primates, backyard chickens and guinea fowl and healthy children from the Gambia?

### 1.9.2 Aims and objectives

To address these gaps, my work aims to investigate the within-host intra- and inter-strain diversity among the population of *E. coli* isolates from the healthy vertebrate gut in:

a. Non-human primates across a range of habitats in the Gambia (Chapter Three).
b. Free-range poultry reared near humans in the Gambia (Chapter Four).
c. Healthy children in rural Gambia (Chapter Five).

# 2 CHAPTER TWO: METHODS

## 2.1 Microbiological processing

For the growth and isolation of *E. coli*, 0.1–0.2 g aliquots were taken from each stool sample into 1.5 ml microcentrifuge tubes under aseptic conditions. To each tube, 1 ml of physiological saline (0.85%) was added, and the saline-stool samples were vortexed for 2 min at 4200 rpm. The homogenised samples were taken through four ten-fold serial dilutions and a 100 µl aliquot from each dilution was spread on a plate of tryptone-bile-X-glucuronide (TBX) agar (Oxoid, Basingstoke, UK) using the cross-hatching method.

Inoculated TBX agar plates were incubated at 37°C for 18–24 h in air. Colony counts were performed for each serial dilution, counting translucent colonies with blue-green pigmentation and entire margins as *E. coli*. Up to five colonies from each sample were sub-cultured on MacConkey agar at 37°C for 18–24 h and then stored in 20% glycerol broth at -80°C. If a sample showed growth of fewer than five colonies, all the observed colonies were selected for the subsequent analysis. Previous studies have shown that sampling five colonies provides a 99.3% chance of recovering at least one of the dominant genotypes present in a single stool specimen [505, 506].

The isolates derived from non-human primate stools were designated by the primate species and the site from which they were sampled as follows: *Chlorocebus sabaeus*, 'Chlos'; *Papio papio*, 'Pap'; *Piliocolobus badius*, 'Prob'; Abuko Nature Reserve, 'AN'; Bijilo Forest Park, 'BP'; Kartong village, 'K'; Kiang West National Park, 'KW'; Makasutu Cultural Forest, 'M'; and River Gambia National Park, 'RG'. The colony number was then given after the primate species and site code; for example, "ChlosBP-25-1" represents the first *E. coli* isolate (colony 1) derived from a *Chlorocebus sabaeus* monkey (individual 25) from Bijilo Park. The isolates from chickens were designated "C1-C10", while those from guinea fowl were prefixed by "GF1-GF9", followed by the respective colony number (1 up to 5), while for the human stools, the individual isolates were designated by the study subject ID followed by the colony number ("1-5").

## 2.2 Genomic DNA extraction

A 96-well lysate method gratefully obtained from Professor Tom Connor's lab at Cardiff University, Wales (https://github.com/connor-lab) was adapted for the genomic DNA extraction as follows.

### 2.2.1 Overnight cultures

A single colony from each subculture was picked into 1 ml of Luria-Bertani broth and incubated overnight at 37°C in a 96-well deep-well plate. Following overnight incubation, the broth cultures were spun at 3500rpm for 2 min in a large centrifuge to pellet the bacteria growth. The culture supernatant was removed by placing a clean 1000 µl tip box over the plate and rapidly inverting the plate upside down. The plate was tapped gently to break surface tension in some wells. The plate was then pulse-spun in a large centrifuge to sediment the bacterial growth.

### 2.2.2 DNA extraction

Genomic DNA was extracted using an in-house 96-well plate lysate method. Briefly:

1. A lysing buffer (10.2 ml) was prepared using
    a. 10 ml of TE buffer
    b. 100 µl of lysozyme
    c. 10 µl of RNAse A
2. A 100 µl of the lysing buffer was added to each well and the sediment resuspended by careful pipetting.
3. A 100 µl of the resuspended bacterial growth was transferred to a new deep-well 96-well plate, sealed firmly using an adhesive seal and placed on a plate shaker at 37°C and 1600 rpm for 25 min.
4. While the plate was incubating, a lysing additive was freshly prepared as follows:
    a. 528 µl of TE buffer
    b. 600 µl 10% of SDS buffer
    c. 60 µl of Proteinase K
    d. 12 µl of RNAse A
5. The plate was removed from the plate shaker and 10 µl of the lysing additive added to each well.
6. The plate was sealed firmly with adhesive tape and placed on a plate shaker, this time at -65°C 1600 rpm for 15 min.
7. The plate was spun briefly in a large centrifuge and a 100 µl transferred from each well into a new lo-bind PCR 96 well plate.

### 2.2.3 Solid-Phase Reversible Immobilisation (SPRI) clean-ups

1. To each well, 50 µl of SPRI magnetic beads (Becter Coulter Inc., Brea, CA, USA) was added and carefully mixed by pipetting.

2. The SPRI beads-extractions mixture was incubated at room temperature for 5 min.
3. The plate was placed on a magnetic 96-well plate holder for 2-5 min (The plate was kept on the magnetic apparatus till step 8).
4. The liquid was removed and discarded and 100 ul of 80 % ethanol added to all wells, carefully running the liquid over the magnetic beads.
5. The ethanol was removed and discarded, and this washing step repeated two times.
6. The final ethanol was removed, and the plate allowed to dry for 2 min.
7. The plate was then taken off the magnetic plate holder.
8. To each well, 50 of 10mM Tris-Cl was added, mixed by pipetting to resuspend all the magnetic beads and incubated at room temperature for 5 min.
9. The plate was placed back on the magnetic 96-well plate holder and left to stand for 2 min.
10. Finally, the 50 µl genomic extraction was transferred from each. Well into a new lo-bind 96 well PCR plate and stored at -20 °C until further analysis.

### 2.2.4 Post-extraction quality assessment

The DNA was evaluated for protein and RNA contamination using $A_{260}/A_{280}$ and $A_{260}/A_{230}$ ratios on the NanoDrop 2000 Spectrophotometer (Fisher Scientific, Loughborough, UK). DNA concentrations were measured using the Qubit HS DNA assay (Invitrogen, MA, USA). DNA was stored at -20°C.

## 2.3 Illumina sequencing

Whole-genome sequencing was carried out for all the study isolates on the Illumina NextSeq 500 platform (Illumina, San Diego, CA). I used a modified Nextera XT DNA protocol for the library preparation as follows. The genomic DNA was normalised to 0.5 ng µl$^{-1}$ with 10 mM Tris-HCl. Next, 0.9 µl of Tagment DNA buffer (Illumina Catalogue No. 15027866) was mixed with 0.09 µl of Tagment DNA enzyme (Illumina Catalogue No. 15027865) and 2.01 µl of PCR-grade water in a master-mix. Next, 3 µl of the master-mix was added to a chilled 96-well plate. To this, 2 µl of normalised DNA (1 ng total) was added, pipette-mixed and the reaction heated to 55°C for 10 min on a PCR block. To each well, I added 11 µl of KAPA2G Robust PCR master-mix (Sigma Catalogue No. KK5005), comprising 4 µl KAPA2G buffer, 0.4 µl dNTPs, 0.08 µl polymerase and 6.52 µl PCR-grade water, contained in the kit per sample. Next, 2 µl each of P7 and P5 Nextera XT Index Kit v2 index primers (Illumina Catalogue numbers FC-131-2001 to 2004) were added to each well. Finally, the 5 µl of Tagmentation mix was added and mixed. The PCR was run as follows:

- 72°C for 3 min
- 95°C for 1 min
- 14 cycles of 95°C for 10 sec
- 55°C for 20 sec and 72°C for 3 min

Following the PCR, the libraries were quantified using the Quant-iT dsDNA Assay Kit, high sensitivity kit (Catalogue No. 10164582) and run on a FLUOstar Optima plate reader. After quantification, libraries were pooled in equal quantities. The final pool was double-SPRI size-selected between 0.5 and 0.7x bead volumes using KAPA Pure Beads (Roche Catalogue No. 07983298001). I then quantified the final pool on a Qubit 3.0 instrument (Invitrogen, MA, USA) and ran it on a high sensitivity D1000 ScreenTape (Agilent Catalogue No. 5067-5579) using the Agilent TapeStation 4200 to calculate the final library pool molarity. The pooled library was run at a final concentration of 1.8 pM on an Illumina NextSeq500 instrument using a mid-output flow cell (NSQ® 500 Mid Output KT v2 300 cycles; Illumina Catalogue No. FC-404-2003) following the Illumina recommended denaturation and loading parameters, which included a 1% PhiX spike (PhiX Control v3; Illumina Catalogue FC-110-3001). The data was uploaded to BaseSpace (http://www.basespace.illumina.com) and then converted to FASTQ files.

### 2.4 Oxford Nanopore sequencing

Eight novel strains (six derived from non-human primates and two from guinea fowl) and six isolates with interesting plasmid profiles were sequenced using the Oxford Nanopore technology. I used the rapid barcoding kit (Oxford Nanopore Catalogue No. SQK-RBK004) to prepare libraries according to the manufacturer's instructions. I used 400 ng DNA for library preparation and loaded 75 µl of the prepared library on an R9.4 MinION flow cell. The size of the DNA fragments was assessed using the Agilent 2200 TapeStation (Agilent Catalogue No. 5067-5579) before sequencing. The concentration of the final library pool was measured using the Qubit high-sensitivity DNA assay (Invitrogen, MA, USA).

### 2.5 Genome assembly and phylogenetic analysis

Sequences were analysed on the Cloud Infrastructure for Microbial Bioinformatics (CLIMB) [507]. This included concatenating paired-end short reads, quality checks with FastQC v0.11.7 [508], trimming low quality reads (median quality below a Phred score of ~30 and read lengths below 36bp) and Illumina adapters with Trimmomatic v0.39 [509] and assembly by Spades v3.13.2

[314]. The quality of the assemblies was assessed using QUAST v 5.0.0, de6973bb [510]. Draft bacterial genomes were annotated using Prokka v 1.13 [322]. Multi-locus sequence types (STs) were called from assemblies according to the Achtman scheme [511] using the mlst software (https://github.com/tseemann/mlst) to scan alleles in PubMLST (https://pubmlst.org/) [512]. Novel STs were assigned by EnteroBase. Snippy v4.3.2 (https://github.com/tseemann/snippy) was used for variant calling and core genome alignment, including references genome sequences representing the significant phylogroups of *E. coli* and *Escherichia fergusonii* as an outgroup (Table 2.1). Given that recombination is widespread in *E. coli* and tends to blur phylogenetic signals [511], I used Gubbins (Genealogies Unbiased By recomBinations In Nucleotide Sequences) [513] to detect and mask recombinant regions of the core-genome alignment. RAxML v 8.2.4 [329] was used for maximum-likelihood phylogenetic inference from this masked alignment based on a general time-reversible nucleotide substitution model with 1,000 bootstrap replicates. The phylogenetic tree was visualised using Mega v. 7.2 [514] and FigTree v1.4.3 (https://github.com/rambaut/figtree/) and annotated in RStudio v3.5.1 and annotated using Adobe Illustrator v 23.0.3 (Adobe Inc., San Jose, California). For visualisation, a single colony was chosen to represent replicate colonies of the same strain (ST) with identical virulence, plasmid and antimicrobial resistance profiles and a de-replicated phylogenetic tree reconstructed using the representative isolates. Pair-wise single nucleotide polymorphism (SNP) distances between genomes were computed from the core-gene alignment using snp-dists v0.6 (https://github.com/tseemann/snp-dists).

***Table 2.1:*** *Reference strains included in the phylogenetic analysis.*

| **Strain** | **Sequence Type** | **Phylogroup designation** | **GenBank assembly accession** |
|---|---|---|---|
| K-12 strain MG1655 | ST10 | A | NC_000913.3 |
| 536 | ST127 | B2 | GCA_000013305.1 |
| UMN026 | ST597 | D | GCA_000026325.2 |
| IAI39 | ST62 | F | GCA_000026345.1 |
| O157:H7 str. EDL933 | ST11 | E | GCA_000732965.1 |
| IAI1 | ST1128 | B1 | GCA_000026265.1 |
| IHE3034 | ST95 | B2 | GCA_000025745.1 |
| *Escherichia fergusonii* | ST5298 | Outroot species | GCA_000026225.1 |

## 2.6 Population structure analysis

Merged short reads were uploaded to EnteroBase [139] where I used the Hierarchical Clustering (HierCC) algorithm to assign my genomes from non-human primates to HC1100 clusters, which in *E. coli* correspond roughly to the clonal complexes seen in seven-allele MLST. I reconstructed neighbour-joining phylogenetic trees using NINJA [346] —a hierarchical clustering algorithm for inferring phylogenies that is capable of scaling to inputs larger than 100,000 sequences [346], based on clustering at HC1100 to display the population sub-clusters at this level as an indicator of the genomic diversity within my study population and to infer the evolutionary relationship among my strains and others in the public domain.

I used GrapeTree [333] to visualise and annotate phylogenetic trees. Further annotation of the phylogenetic trees was carried out using Adobe Illustrator v 23.0.3 (Adobe Inc., San Jose, California).

### 2.6.1 Non-human primate strains

Entries within EnteroBase were interrogated by host and location to isolate *E. coli* strains derived from non-human primates from the rest of the world. All such strains were stratified by continent and the prevalence of STs and distribution of phylogroups determined. Also, a neighbour-joining phylogenetic tree was reconstructed using these strains from the rest of the world to infer the evolutionary relationships among them and assess the patterns of distribution of phylogroups and sequence types.

### 2.6.2 Backyard poultry strains

In order to compare the strain distribution that was observed among my study isolates with what pertains in poultry *E. coli* isolates from elsewhere, I further retrieved genomic assemblies from all publicly available poultry *E. coli* isolates, stratified by their source continent and reconstructed NINJA neighbour-joining trees depicting the prevalence of STs per continent.

### 2.6.3 Human *E. coli* strains

For the human strains, I interrogated the HC1100 clusters that encompassed my study isolates and Gambian pathogenic isolates recovered from diarrhoeal cases and commensal *E. coli* strains recovered from the Global Enteric Multicenter Study (GEMS) [515, 516]. GEMS is described in more detail in Chapter Four. For the clusters that encompassed commensal and pathogenic strains belonging to the same ST (HC1100_200 cluster, comprising pathogenic isolates from GEMS cases 100415, 102106 and 102098 and the resident ST38 strain recovered from my study subject 18), I reconstructed both neighbour-joining and SNP phylogenetic trees to display the genetic relationships among these strains. I visualised the accessory genomes for the overlapping STs mentioned above to determine genes associated with phages, virulence factors and AMR. The resulting phylogenetic trees were annotated in Adobe Illustrator v 23.0.3 (Adobe Inc., San Jose, California).

## 2.7 Analysis of accessory gene contents

ARIBA v2.12.1 [517] was used to scan the short reads against the Virulence Factors Database [518] (VFDB-core) (virulence-associated genes), ResFinder (AMR) [519] and PlasmidFinder (plasmid-associated genes) [520] databases (both ResFinder and PlasmidFinder databases downloaded 29 October 2018). Percentage identity of ≥ 90% and coverage of ≥ 70% of the respective gene length were taken as a positive result. The VFDB-core, ResFinder and PlasmidFinder databases were downloaded on 29 October 2018. As a quality check, the results were confirmed by running ABRicate v0.9.8 (https://github.com/tseemann/abricate) (databases updated 12 October 2020) using the assembled contigs. Briefly, ABRicate v0.9.8 (https://github.com/tseemann/abricate) predicts virulence factors, acquired antimicrobial resistance (AMR) genes and plasmid replicons by scanning the contigs against the VFDB, ResFinder and PlasmidFinder databases respectively, using an identity threshold of ≥ 90% and a coverage of ≥ 70%. A heat map of detected virulence- and AMR-associated genes was plotted on the phylogenetic tree using ggtree and phangorn in RStudio v 3.5.1.

### 2.7.1 Non-human primate population

I searched EnteroBase for all *E. coli* strains isolated from humans in the Gambia (n=128), downloaded the genomes and screened them for resistance genes using ABRicate v 0.9.8. Also, all assembled genomes for isolates that clustered with my colibactin-encoding ST73, ST127 and ST681 isolates were downloaded and screened for the colibactin operon using ABRicate's VFDB database (accessed 28 July 2019). Assemblies predicted to contain colibactin genes were aligned against the colibactin-encoding *Escherichia coli* IHE3034 reference genome (NCBI Accession: GCA_000025745.1) using minimap2 2.13-r850. BAM files were visualised in Artemis Release 17.0.1 [521] to confirm the presence of the *pks* genomic island, which encodes the colibactin operon.

### 2.7.2 Poultry *E. coli* strains from the rest of the world

I determined the prevalence of AMR genes among poultry *E. coli* isolates from the rest of the world, for comparison with what I found in isolates from this study. To do this, I interrogated the downloaded continent-stratified genomes as above using ABRicate v0.9.8

(https://github.com/tseemann/abricate) to predict AMR-associated genes by scanning against the ResFinder database (accessed 28 July 2019), based on a percentage identity threshold of ≥ 90% and a coverage of ≥ 70%.

## 2.8 Hybrid assembly and analysis of plasmids and phages

Base-called FASTQ files were concatenated into a single file and demultiplexed into individual FASTQ files based on barcodes, using the qcat python command-line tool v 1.1.0 (https://github.com/nanoporetech/qcat). Hybrid assemblies of the Illumina and Nanopore reads were created with Unicycler [312]. The quality and completion of the hybrid assemblies were assessed with QUAST v 5.0.0, de6973bb and CheckM [510, 522]. Hybrid assemblies were interrogated using ABRicate PlasmidFinder [520] and annotated using Prokka [322]. Plasmid sequences were visualised in Artemis using coordinates from ABRicate. Prophage identification was carried out using the phage search tool, PHASTER [523].

## 2.9 Antimicrobial susceptibility testing

I determined the minimum inhibitory concentrations of amikacin, trimethoprim, sulfamethoxazole, ciprofloxacin, cefotaxime and tetracycline for the isolates from non-human primates and backyard chickens and guinea fowl using agar dilution [524]. Briefly:

- Bacteria from four to five morphologically similar colonies obtained from fresh, pure cultures were taken (by lightly touching the top of the colonies using a sterile loop) into 2.5ml of sterile saline solution to prepare a bacterial suspension with a turbidity equivalent to 0.5 McFarland’s standard ($10^8$ cfu $ml^{-1}$).
- The suspension was mixed briefly using a vortex mixer.
- In the event of the suspension’s turbidity being too high or too low, the turbidity was adjusted to match that of the 0.5 McFarland by adding sterile saline or more bacterial cells respectively.
- 200 µl of each sample suspension was transferred into a well of a 96-well plate
- Stock solutions of 1000 mg $l^{-1}$ were initially prepared, from which the working solutions were made.
- Two-fold serial dilutions of each antibiotic were performed in molten Mueller-Hinton agar (Oxoid, Basingstoke, UK), from 32mg/L to 0.03 mg $l^{-1}$ (512 mg $l^{-1}$ to 0.03 mg $l^{-1}$ for sulfamethoxazole), using *E. coli* NCTC 10418 as control.

- Spot inoculations were performed on the agar plates using an automated inoculator available in the Quadram Institute Bioscience shared microbiology laboratory.
- MICs were performed in duplicate and interpreted using breakpoint tables from the European Committee on Antimicrobial Susceptibility Testing v. 9.0, 2019 (http://www.eucast.org), Where EUCAST cut-off values were not available, the recommended cut-off values from the Clinical Laboratory Standards Institute (https://www.clsi.org) were used.

## 2.10 Statistical analysis

Fisher's exact tests were carried out to assess possible associations between the sampling site or non-human primate species and the phylogroups of *E. coli* that were observed using STATA version 14.2. I based my calculations on the assumption of independence across the observed phylogroups, i.e., the finding of one phylogroup does not predict or preclude the occurrence of another. Before the association tests, replicate phylogroups arising from copies of the same ST from a single individual were dropped from the analysis. For the human *E. coli* population, Fisher's exact tests were computed between the detected virulence factors and the observed phylogroups in RStudio v3.5.1.

I generated contingency tables to display the correlation between the phenotypic results and the detected resistance genes among the study isolates and calculated the percentage concordance between the genotypic and phenotypic resistances. Also, the co-occurrence of AMR genes among study isolates was calculated by transforming the binary AMR gene content matrix and visualising this as a heat map using the pheatmap package v 1.0.12 (https://CRAN.R-project.org/package=pheatmap) in RStudio v3.5.1.

## 2.11 Preparation of microbiologic media

### 2.11.1 Tryptone-Bile X-Glucuronide medium (TBX)

TBX media were prepared according to the manufacturer's instructions as follows:

1. 36.6 g of the dehydrated media powder was dissolved in 1 L of distilled water with stirring at room temperature
2. The medium was sterilised by autoclaving at 121°C for 15 min
3. The sterilised medium was allowed to cool to 45°C – 50°C in a water bath
4. 20 ml of the medium was poured into each sterile Petri plate in an air flow cabinet and allow to solidify at room temperature

5. The plates were labelled with the batch of media, date of preparation and expiry dates and stored at 4°C – 8°C
6. The plates were dried in the hot air oven at 50°C for 10 min before inoculation to prevent bacterial swarming

### 2.11.2 MacConkey agar

According to the manufacturer's instructions, MacConkey agar plates were prepared as follows:

1. 52 g of the dehydrated medium powder was suspended in 1 L of distilled water with stirring at room temperature.
2. The solution was brought to the boil to dissolve completely.
3. The medium was sterilised by autoclaving at 12°C for 15 min.
4. The sterilised medium was allowed to cool to 45°C – 50°C in a water bath.
5. 20 ml of the medium was poured into each sterile Petri plate in an airflow cabinet and allow to solidify at room temperature.
6. The plates were labelled and packed in clear plastic bags and stored at 4°C – 8°C until further use.
7. The surface of the gel was dried at in the hot air oven at 50°C for 10 min before inoculation to prevent bacterial swarming.

### 2.11.3 Mueller Hinton agar

Mueller-Hinton agar was prepared as follows:

1. 38 g of the dehydrated powder was added to 1 L of distilled water.
2. The solution was brought to the boil to dissolve completely and sterilised at 121°C for 15 min.
3. The medium was then sterilised as for TBX and MacConkey above.
4. As the medium was cooling, twelve sterile Erlenmeyer flasks were labelled with the final antibiotic concentration and 100 ml of the sterilised medium poured into each container when it had cooled sufficiently (approximately 50°C).
5. The square plates were also pre-labelled with the respective antibiotic concentrations.
6. 25 ml of the medium was dispensed into each of 120 mm square plates, after adding the appropriate amounts of antibiotic solution (Two-fold serial dilutions of each antibiotic from 32 mg/L to 0.03 mg $l^{-1}$ for amikacin, trimethoprim, cefotaxime, ciprofloxacin and tetracycline (512 mg $l^{-1}$ to 0.03 mg $l^{-1}$ for sulfamethoxazole) into the respective containers.

7. The plates were stored at 4°C – 8°C until further use.

### 2.11.4 Skim-milk-Tryptone-Glucose-Glycerol broth

Skim-milk-Tryptone-Glucose-Glycerol broth was prepared as follows:

- 3 g Oxoid tryptone soya broth
- 0.5 g Glucose
- 2.0 g Oxoid skim milk powder
- 10 ml Glycerol
- 100 ml Distilled water

1 ml of the prepared medium was dispensed into each of 1.8 ml Nunc tubes and autoclaved for 10 minutes at 121°C. The sterilised medium was allowed to cool to room temperature and stored at -20°C or refrigerated (4-8°C) until use. Prior to use, the STGG suspensions were rigorously vortexed at 4200 rpm for a minimum of 10-15 s to resuspend the pellet.

### 2.11.5 Glycerol Broth (16%)

1. First, nutrient broth medium (Oxoid, Basingstoke, UK) was prepared as described above.
2. The glycerol broth was then made up as follows:
    a. 84 ml of nutrient broth
    b. 16 ml of glycerol
    c. This makes up 100 ml of glycerol broth
3. The solution was mixed by inverting for a few times and then distributed in 15 ml amounts in universal bottles.
4. The broth was sterilised by autoclave at 115°C for 20 minutes.
5. The cooled medium was stored in the refrigerator at 4°C – 8°C.

### 2.11.6 Normal saline (0.85%)

1. 4.25 g (1 tablet, Oxoid product code BR053G) was dissolved in 500 ml of distilled water and allowed to dissolve by stirring at room temperature.
2. The solution was sterilised by autoclave at 121°C for 15minutes.
3. The sterilised solution was allowed to cool and stored at room temperature or in the refrigerator at 4°C – 8°C until further use.

# 3 CHAPTER THREE: GENOMIC DIVERSITY OF *ESCHERICHIA COLI* ISOLATES FROM NON-HUMAN PRIMATES IN THE GAMBIA

## 3.1 Introduction

As previously discussed (Chapter One), *Escherichia coli* is a highly versatile species, capable of adapting to a wide range of ecological niches and colonising a diverse range of hosts [46, 525]. In non-human primates [2], data from captive animals suggest that gut isolates are dominated by phylogroups B1 and A, which, in humans, encompass commensals as well as strains associated with intestinal pathology [116, 526]. *E. coli* strains encoding colibactin, or cytotoxic necrotising factor 1 have been isolated from healthy laboratory rhesus macaques [527], while enteropathogenic *E. coli* strains can—in the laboratory—cause colitis in marmosets [528], rhesus macaques infected with simian immunodeficiency virus [529] and cotton-top tamarins [530].

There are two potential explanations for the occurrence of *E. coli* in humans and non-human primates. Some bacterial lineages may have been passed on through vertical transmission within the same host species for long periods, perhaps even arising from ancestral bacteria that colonised the guts of the most recent common ancestors of humans and non-human primate species [531, 532]. In such a scenario, isolates from non-human primates would be expected to be novel and distinct from the diversity seen in humans [532]. However, there is also clearly potential for horizontal transfer of strains from one host species to another [533].

The exchange of bacteria between humans and human-habituated animals, particularly non-human primates, is of interest in light of the fragmentation of natural habitats globally [534-536]. Wild non-human primates in the Gambia are frequently exposed to humans through tourism, deforestation and urbanisation [537]. In Uganda, PCR-based studies have suggested transmission of *E. coli* between humans, non-human primates and livestock [538, 539]. These studies are complicated by the low resolution of PCR-based methods, nonetheless, their findings highlight the possibility that wild non-human primates may constitute a reservoir for the zoonotic spread of *E. coli* strains associated with virulence and antimicrobial resistance to humans. Alternatively, humans might provide a reservoir of strains with the potential for anthroponotic spread to animals—or transmission might occur in both directions [296].

We do not know how many different lineages can co-exist within the same non-human primate host. Such information may help us contextualise the potential risks associated with transmission of bacterial strains between humans and non-human primates. In humans, up to eleven serotypes could be sampled from picking eleven colonies from individual stool samples [249, 413, 414].

To address these issues, I have exploited whole-genome sequencing to explore the population structure and phylogenomic diversity of *E. coli* in wild non-human primates from rural and urban Gambia.

## 3.2 Materials and methods

### 3.2.1 Study population and sample collection

In June 2017, wild non-human primates were sampled from six sampling sites in the Gambia: Abuko Nature Reserve (riparian forest), Bijilo Forest Park (coastal fenced woodland), Kartong village (mangrove swamp), Kiang West National park (dry-broad-leaf forest), Makasutu Cultural Forest (ecotourism woodland) and River Gambia National park (riparian forest) (Figure 3.1). The sampling was throughout the range of the primates in the country (all four of the diurnal non-human primate species indigenous to the Gambia), where primates overlap with human communities to varying degrees.

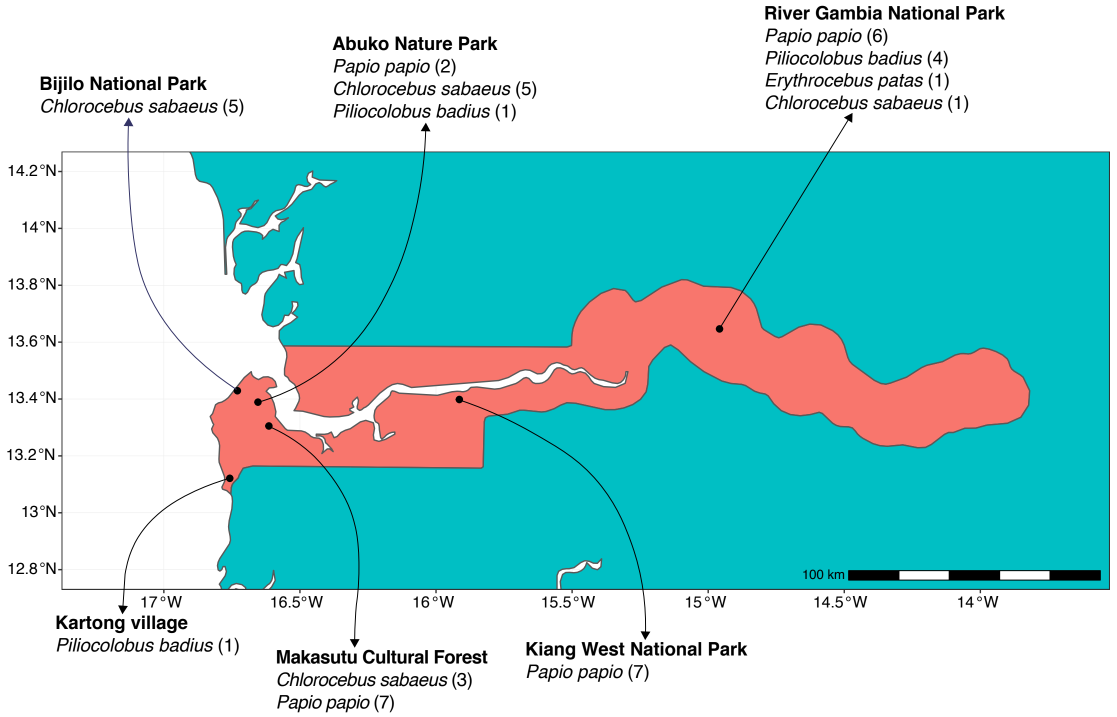


***Figure 3.1:*** *Study sites and distribution of study subjects per sampling site.*

Monkeys in Abuko and Bijilo are frequently hand-fed by visiting tourists, despite prohibiting guidelines [537].

Troops of monkeys were observed and followed. A single freshly passed formed stool specimen was collected from 43 visibly healthy individuals (38 adults, 5 juveniles; 24 females, 11 males, 8 of undetermined sex), drawn from four species: *Erythrocebus patas* (patas monkey), *Papio papio* (Guinea baboon), *Chlorocebus sabaeus* (green monkey) and *Piliocolobus badius* (Western colobus monkey). Stool samples were immediately placed into sterile falcon tubes, taking care to collect portions of stool material that had not touched the ground, then placed on dry ice and stored at -80°C within 6 h. The sample processing flow is summarised in Figure 3.2.

### 3.2.2 Chapter-specific processes and overall sample processing workflow

A detailed explanation of the methods is given in the methods chapter (Chapter Two), summarised in the flowchart presented in Figure 3.2 below.

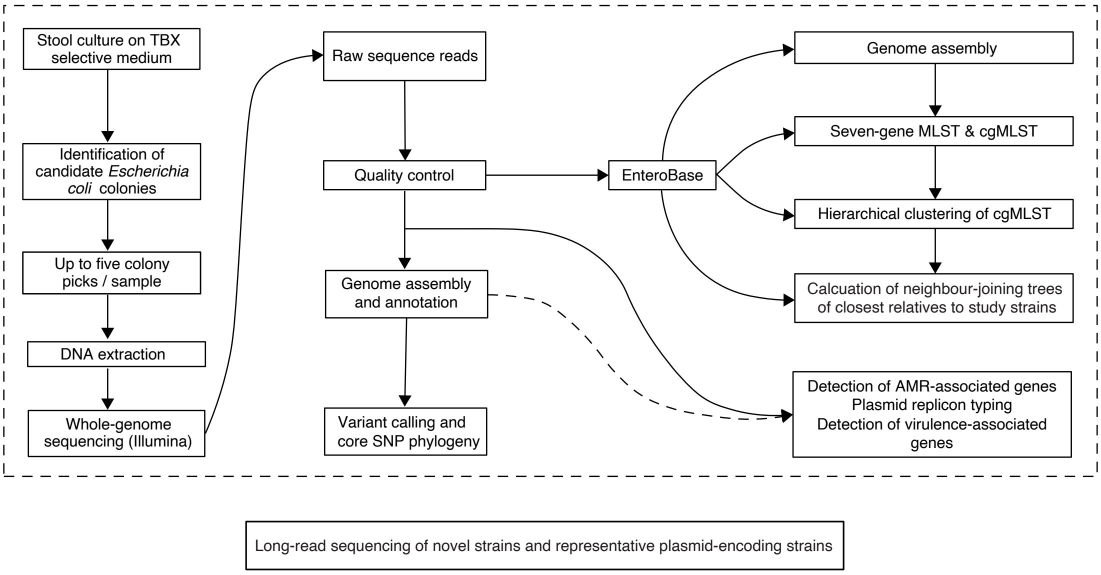

Stool culture on TBX selective medium
Identification of candidate *Escherichia coli* colonies
Up to five colony picks / sample
DNA extraction
Whole-genome sequencing (Illumina)
Raw sequence reads
Quality control
Genome assembly and annotation
Variant calling and core SNP phylogeny
EnteroBase
Genome assembly
Seven-gene MLST & cgMLST
Hierarchical clustering of cgMLST
Calcuation of neighbour-joining trees of closest relatives to study strains
Detection of AMR-associated genes
Plasmid replicon typing
Detection of virulence-associated genes
Long-read sequencing of novel strains and representative plasmid-encoding strains

***Figure 3.2:*** *A flowchart summarising the study sample microbiological processing and bioinformatics analysis.*

## 3.3 Results

### 3.3.1 Study population

Twenty-four of 43 samples (56%) showed growth indicative of *E. coli*, yielding a total of 106 colonies. After genome sequencing, five isolates (PapRG-04, (n=1); PapRG-03 (n=1); ChlosRG-12 (n=1); ChlosAN-13 (n=1); ProbAN-19 (n=1) were excluded due to low depth of coverage (<20x), leaving 101 genomes for subsequent analysis (Table 3.1).

***Table 3.1:*** *Metadata for the 101 E. coli isolates analysed in this chapter.*

| Name | Source | Individual sampling number | Colony-pick | Sampling site | ST |
|---|---|---|---|---|---|
| PapRG-03-1 | *Papio papio* | 3 | 1 | River Gambia national park | 336 |
| PapRG-03-2 | *Papio papio* | 3 | 2 | River Gambia national park | 336 |
| PapRG-03-3 | *Papio papio* | 3 | 3 | River Gambia national park | 336 |
| PapRG-03-4 | *Papio papio* | 3 | 4 | River Gambia national park | 336 |
| PapRG-03-5 | *Papio papio* | 3 | 5 | River Gambia national park | 336 |
| PapRG-04-1 | *Papio papio* | 4 | 1 | River Gambia national park | 1665 |
| PapRG-04-2 | *Papio papio* | 4 | 2 | River Gambia national park | 1204 |
| PapRG-04-4 | *Papio papio* | 4 | 3 | Makasutu cultural forest | 8826 |
| PapRG-04-5 | *Papio papio* | 4 | 4 | Makasutu cultural forest | 1204 |
| PapRG-05-2 | *Papio papio* | 5 | 1 | Makasutu cultural forest | 1431 |
| PapRG-05-3 | *Papio papio* | 5 | 2 | Makasutu cultural forest | 99 |
| PapRG-05-4 | *Papio papio* | 5 | 3 | Makasutu cultural forest | 6316 |
| PapRG-05-5 | *Papio papio* | 5 | 4 | Makasutu cultural forest | 1431 |
| PapRG-06-1 | *Papio papio* | 6 | 1 | Makasutu cultural forest | 4080 |
| PapRG-06-2 | *Papio papio* | 6 | 2 | Makasutu cultural forest | 2521 |
| PapRG-06-3 | *Papio papio* | 6 | 3 | Makasutu cultural forest | 8827 |
| PapRG-06-4 | *Papio papio* | 6 | 4 | Makasutu cultural forest | 1204 |
| PapRG-06-5 | *Papio papio* | 6 | 5 | River Gambia national park | 8525 |
| ProbRG-07-1 | *Piliocolobus badius* | 7 | 1 | River Gambia national park | 73 |
| ProbRG-07-2 | *Piliocolobus badius* | 7 | 2 | River Gambia national park | 73 |
| ProbRG-07-3 | *Piliocolobus badius* | 7 | 3 | River Gambia national park | 73 |
| ProbRG-07-4 | *Piliocolobus badius* | 7 | 4 | River Gambia national park | 73 |
| ProbRG-07-5 | *Piliocolobus badius* | 7 | 5 | River Gambia national park | 73 |
| ChlosRG-12-1 | *Chlorocebus sabaeus* | 12 | 1 | River Gambia national park | 8824 |
| ChlosRG-12-2 | *Chlorocebus sabaeus* | 12 | 2 | River Gambia national park | 196 |
| ChlosRG-12-3 | *Chlorocebus sabaeus* | 12 | 3 | River Gambia national park | 196 |
| ChlosRG-12-5 | *Chlorocebus sabaeus* | 12 | 4 | River Gambia national park | 40 |
| ChlosAN-13-1 | *Chlorocebus sabaeus* | 13 | 1 | Abuko Nature Reserve | 8526 |
| ChlosAN-13-2 | *Chlorocebus sabaeus* | 13 | 2 | Abuko Nature Reserve | 8550 |
| ChlosAN-13-4 | *Chlorocebus sabaeus* | 13 | 3 | Abuko Nature Reserve | 1973 |
| ChlosAN-13-5 | *Chlorocebus sabaeus* | 13 | 4 | Abuko Nature Reserve | 1973 |
| PapAN-14-1 | *Papio papio* | 14 | 1 | Abuko Nature Reserve | 2076 |
| PapAN-14-2 | *Papio papio* | 14 | 2 | Abuko Nature Reserve | 939 |
| PapAN-14-3 | *Papio papio* | 14 | 3 | Abuko Nature Reserve | 226 |
| PapAN-14-4 | *Papio papio* | 14 | 4 | Abuko Nature Reserve | 226 |
| PapAN-14-5 | *Papio papio* | 14 | 5 | Abuko Nature Reserve | 226 |
| PapAN-15-1 | *Papio papio* | 15 | 1 | Abuko Nature Reserve | 226 |
| PapAN-15-2 | *Papio papio* | 15 | 2 | Abuko Nature Reserve | 5073 |
| PapAN-15-3 | *Papio papio* | 15 | 3 | Abuko Nature Reserve | 226 |
| PapAN-15-4 | *Papio papio* | 15 | 4 | Abuko Nature Reserve | 126 |
| PapAN-15-5 | *Papio papio* | 15 | 5 | Abuko Nature Reserve | 8823 |
| ChlosAN-17-1 | *Chlorocebus sabaeus* | 17 | 1 | Abuko Nature Reserve | 681 |
| ChlosAN-17-2 | *Chlorocebus sabaeus* | 17 | 2 | Abuko Nature Reserve | 362 |
| ChlosAN-17-3 | *Chlorocebus sabaeus* | 17 | 3 | Abuko Nature Reserve | 681 |
| ChlosAN-17-4 | *Chlorocebus sabaeus* | 17 | 4 | Abuko Nature Reserve | 681 |
| ChlosAN-18-1 | *Chlorocebus sabaeus* | 18 | 1 | Abuko Nature Reserve | 681 |
| ChlosAN-18-2 | *Chlorocebus sabaeus* | 18 | 2 | Abuko Nature Reserve | 681 |
| ChlosAN-18-3 | *Chlorocebus sabaeus* | 18 | 3 | Abuko Nature Reserve | 681 |
| ChlosAN-18-4 | *Chlorocebus sabaeus* | 18 | 4 | Abuko Nature Reserve | 681 |
| ChlosAN-18-5 | *Chlorocebus sabaeus* | 18 | 5 | Abuko Nature Reserve | 349 |

***Table 3.2:*** *Metadata for the 101 E. coli isolates analysed in this chapter (continued).*

| Name | Source | Individual sampling number | Colony-pick | Sampling site | ST |
|---|---|---|---|---|---|
| ProbAN-19-2 | *Piliocolobus badius* | 19 | 1 | Abuko Nature Reserve | 8825 |
| ChlosBP-21-1 | *Chlorocebus sabaeus* | 21 | 1 | Bijilo forest park | 677 |
| ChlosBP-21-2 | *Chlorocebus sabaeus* | 21 | 2 | Bijilo forest park | 677 |
| ChlosBP-21-3 | *Chlorocebus sabaeus* | 21 | 3 | Bijilo forest park | 677 |
| ChlosBP-21-4 | *Chlorocebus sabaeus* | 21 | 4 | Bijilo forest park | 677 |
| ChlosBP-21-5 | *Chlorocebus sabaeus* | 21 | 5 | Bijilo forest park | 677 |
| ChlosBP-23-1 | *Chlorocebus sabaeus* | 23 | 1 | Bijilo forest park | 8527 |
| ChlosBP-23-2 | *Chlorocebus sabaeus* | 23 | 2 | Bijilo forest park | 8527 |
| ChlosBP-23-3 | *Chlorocebus sabaeus* | 23 | 3 | Bijilo forest park | 3306 |
| ChlosBP-24-1 | *Chlorocebus sabaeus* | 24 | 1 | Bijilo forest park | 73 |
| ChlosBP-24-2 | *Chlorocebus sabaeus* | 24 | 2 | Bijilo forest park | 73 |
| ChlosBP-24-3 | *Chlorocebus sabaeus* | 24 | 3 | Bijilo forest park | 73 |
| ChlosBP-24-4 | *Chlorocebus sabaeus* | 24 | 4 | Bijilo forest park | 73 |
| ChlosBP-24-5 | *Chlorocebus sabaeus* | 24 | 5 | Bijilo forest park | 73 |
| ChlosBP-25-1 | *Chlorocebus sabaeus* | 25 | 1 | Bijilo forest park | 73 |
| ChlosBP-25-2 | *Chlorocebus sabaeus* | 25 | 2 | Bijilo forest park | 73 |
| ChlosBP-25-3 | *Chlorocebus sabaeus* | 25 | 3 | Bijilo forest park | 73 |
| ChlosBP-25-4 | *Chlorocebus sabaeus* | 25 | 4 | Bijilo forest park | 73 |
| ChlosBP-25-5 | *Chlorocebus sabaeus* | 25 | 5 | Bijilo forest park | 73 |
| ChlosM-29-1 | *Chlorocebus sabaeus* | 29 | 1 | Makasutu cultural forest | 1873 |
| ChlosM-29-2 | *Chlorocebus sabaeus* | 29 | 2 | Makasutu cultural forest | 1873 |
| PapM-31-1 | *Papio papio* | 31 | 1 | Makasutu cultural forest | 2800 |
| PapM-31-2 | *Papio papio* | 31 | 2 | Makasutu cultural forest | 135 |
| PapM-31-3 | *Papio papio* | 31 | 3 | Makasutu cultural forest | 5780 |
| PapM-31-4 | *Papio papio* | 31 | 4 | Makasutu cultural forest | 1727 |
| PapM-31-5 | *Papio papio* | 31 | 5 | Makasutu cultural forest | 5780 |
| PapM-32-1 | *Papio papio* | 32 | 2 | Makasutu cultural forest | 8532 |
| PapM-32-2 | *Papio papio* | 32 | 3 | Makasutu cultural forest | 212 |
| PapM-32-3 | *Papio papio* | 32 | 4 | Makasutu cultural forest | 212 |
| PapM-32-4 | *Papio papio* | 32 | 5 | Makasutu cultural forest | 212 |
| PapM-32-5 | *Papio papio* | 32 | 6 | Makasutu cultural forest | 212 |
| PapM-33-1 | *Papio papio* | 33 | 1 | Makasutu cultural forest | 8533 |
| PapM-33-2 | *Papio papio* | 33 | 2 | Makasutu cultural forest | 8533 |
| PapM-33-3 | *Papio papio* | 33 | 3 | Makasutu cultural forest | 8533 |
| PapM-33-4 | *Papio papio* | 33 | 4 | Makasutu cultural forest | 38 |
| PapM-33-5 | *Papio papio* | 33 | 5 | Makasutu cultural forest | 8533 |
| PapM-34-1 | *Papio papio* | 34 | 1 | Makasutu cultural forest | 676 |
| PapM-34-2 | *Papio papio* | 34 | 2 | Makasutu cultural forest | 676 |
| PapM-34-3 | *Papio papio* | 34 | 3 | Makasutu cultural forest | 676 |
| PapM-34-4 | *Papio papio* | 34 | 4 | Makasutu cultural forest | 676 |
| PapM-36-1 | *Papio papio* | 36 | 1 | Makasutu cultural forest | 8535 |
| PapM-36-2 | *Papio papio* | 36 | 2 | Makasutu cultural forest | 8535 |
| PapKW-44-1 | *Papio papio* | 44 | 1 | Kiang West national park | 442 |
| PapKW-44-2 | *Papio papio* | 44 | 2 | Kiang West national park | 442 |
| PapKW-44-3 | *Papio papio* | 44 | 3 | Kiang West national park | 442 |
| PapKW-44-4 | *Papio papio* | 44 | 4 | Kiang West national park | 442 |
| ProbK-45-1 | *Piliocolobus badius* | 45 | 1 | Kartong village | 127 |
| ProbK-45-2 | *Piliocolobus badius* | 45 | 2 | Kartong village | 127 |
| ProbK-45-3 | *Piliocolobus badius* | 45 | 3 | Kartong village | 127 |
| ProbK-45-4 | *Piliocolobus badius* | 45 | 4 | Kartong village | 127 |
| ProbK-45-5 | *Piliocolobus badius* | 45 | 5 | Kartong village | 127 |

### 3.3.2 Distribution of sequence types and phylogroups

I recovered 43 seven-allele sequence types (ten of them novel), spanning five of the eight known phylogroups of *E. coli* and comprising 38 core-genome MLST complexes (Figure 3.3 and Figure 3.4). The majority of strains belonged to phylogroup B2 (42/101, 42%)—which encompasses strains that cause extraintestinal infections in humans (ExPEC strains) [109, 540-542]—followed by B1 (35/101, 35%), A and D (8/101, 8% each), E (7/101, 7%) and cryptic clade I (1/101, 1%). Among the study isolates, I found several STs associated with extraintestinal infections and/or AMR in humans: ST73, ST681, ST127, ST226, ST336, ST349 [543-548]. There was no significant association between the primate species and the prevalence of phylogroups (p=0.17), nor between the sampling sites and phylogroups (p=0.44).

### 3.3.3 Prevalence of virulence factors

I detected a total of 151 virulence factors among the study isolates and an additional 31 from the reference genomes. The following virulence factors were largely conserved across most of the study isolates: the enterobactin-associated cluster of genes (*fepA-D, G* and *entA-F, S*), type I fimbriae (*fimA-I*) and fimbria-associated genes (*yagV-Z*). However, iron acquisition genes (*chuA, S-Y)* appeared to be more prevalent in strains belonging to phylogroups B2, D and E. In general, I detected a higher prevalence of virulence genes in strains belonging to phylogroup B2, compared to those from phylogroups A, B1, D and E (Figure 3.4). These included additional siderophore-encoding genes (*ybt, fyu* and *irp*), capsular antigens (*kpsM1/D*), salmochelin (*iroN/C/B/D/E*), P, S and F1C fimbriae genes (*papC, D, I-K,* X, *focI/C/F* and *sfaY/B* respectively) and the adherence factor protein gene (*fde*C)—representing colonisation and fitness factors associated with extraintestinal disease in humans.

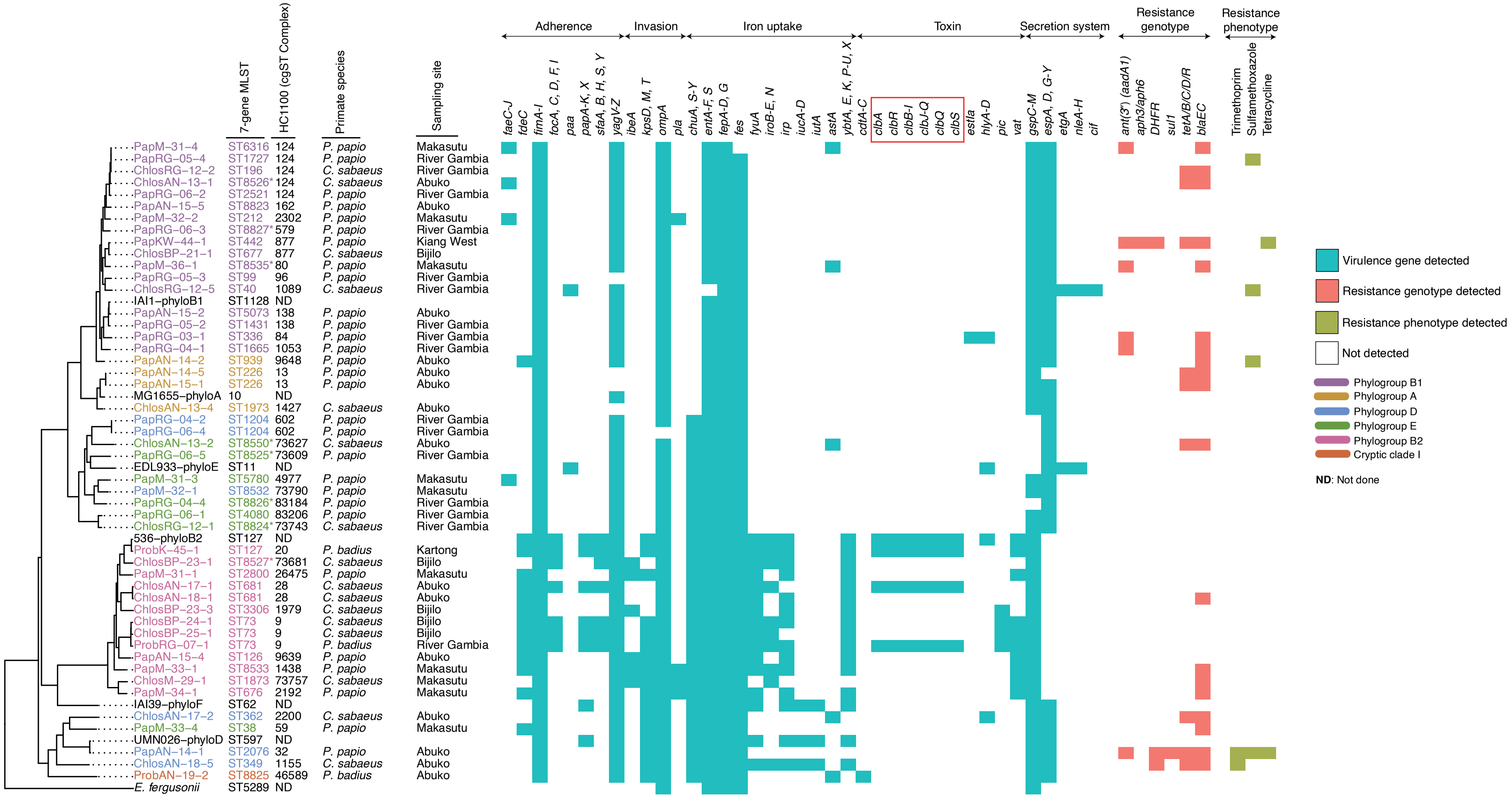
A subset of the B2 strains (13/42, 31%), belonging to STs 73, 681 and 127, carried the *pks* genomic island, which encodes the DNA alkylating genotoxin, colibactin (Figure 3.3, red box). Colibactin-encoding *E. coli* frequently cause colorectal cancer, urosepsis, bacteraemia and prostatitis and are highly associated with other virulence factors such as siderophores and toxins [549-551]. Also, all the ST73 B2 strains carried genes encoding the Serin protease autotransporter (*pic*) and 79% (33/42) of the B2 strains possessed the vacuolating autotransporter (*vat)* toxins.

Leaving aside the B2 strains, I also detected few toxins associated with intestinal and extraintestinal disease in humans among strains from other phylogroups as follows. The heat-stable enterotoxin 1 (*astA)* occurred in five isolates overall (two phylogroup B1 and one each of phylogroups E, D and the *Escherichia* cladel). In addition, the haemolyin genes (*hlyB-D)* were detected in a single Guinea baboon (PapRG-03, phylogroup B1). Also, the invasion of brain endothelium gene (*ibe*A)—responsible for neonatal meningitis in humans—was observed in six Guinea baboon isolates derived from PapM33 and PapM-34 and one Green monkey (ChlosM-29); all belonging to phylogroup B2.

### 3.3.4 Within-host genomic diversity

Thirteen individuals were colonised by two or more STs and nine by two or more phylogroups (Table 3.2). Five colony picks from a single Guinea baboon (PapRG-06) yielded five distinct STs, two of which are novel. Two green monkeys sampled from Bijilo (ChlosBP-24 and ChlosBP-25) shared an identical ST73 genotype, while two Guinea baboons from Abuko shared an ST226 strain—documenting transmission between monkeys of the same species.

In seventeen monkeys, I observed a cloud of closely related genotypes (separated by 0-5 SNPs, Table 3.3) from each strain, suggesting evolution within the host after acqusition of the strain. However, in two individuals, pair-wise SNP distances between genotypes from the same ST were susbtantial enough (25 SNPs and 79 SNPs) to suggest possible multiple acquisitions of each strain (Table 3.4). Reeves *et al.* [407] estimated a mutation rate of 1.1 per genome per year from characterising fourteen ST73 strains isolated from a single family over three years. Based on this data and with the assumption that equal rates of mutation occurred in both genomes, we can infer about 11-35 years of divergence for these strains. Thus, these strains may represent within-host diversity and persistence in the two hosts, going by the lifespan of a green monkey in the wild (averaged at seventeen years). However, it is equally plausible that these SNPS may have accumulated via within-strain recombination.



*The tree was reconstructed based on non-repetitive core SNPs calculated against the E. coli K-12 reference strain (NCBI accession: NC_000913.3), using RAxML with 1000 bootstrap replicates. E. coli MG1655 was used as the reference and E. fergusonii as the outroot species. Recombinant regions were removed using Gubbins (Reference 336). The tip labels indicate the sample IDs, with the respective in silico Achtman sequence types (STs) and HC1100 (cgST complexes) indicated next to the tip labels, along with the species of primate and the sampling site. Both the sample IDs and the STs (Achtman) are colour-coded to indicate the various phylogroups as indicated. Novel STs (Achtman) are indicated by an asterisk (*). Escherichia fergusonii and the E. coli reference genomes representing the major E. coli phylogroups are in black. Co-colonising seven-allele (Achtman) sequence types (STs) in single individuals are shown by the prefix of the strain names depicting the colony as 1, 2 up to 5. I do not show multiple colonies of the same Achtman ST recovered from a single individual. In such cases, only one representative is shown. Virulence genes are grouped according to their function, with genes encoding the colibactin genotoxin highlighted with a red box. The full names of virulence factors are provided in File S7 of Reference 605 (Foster-Nyarko et al., 2020).*

***Figure 3.3:*** *A plot showing the maximum-likelihood phylogeny of the study isolates overlaid with the prevalence of potential virulence genes among study isolates.*

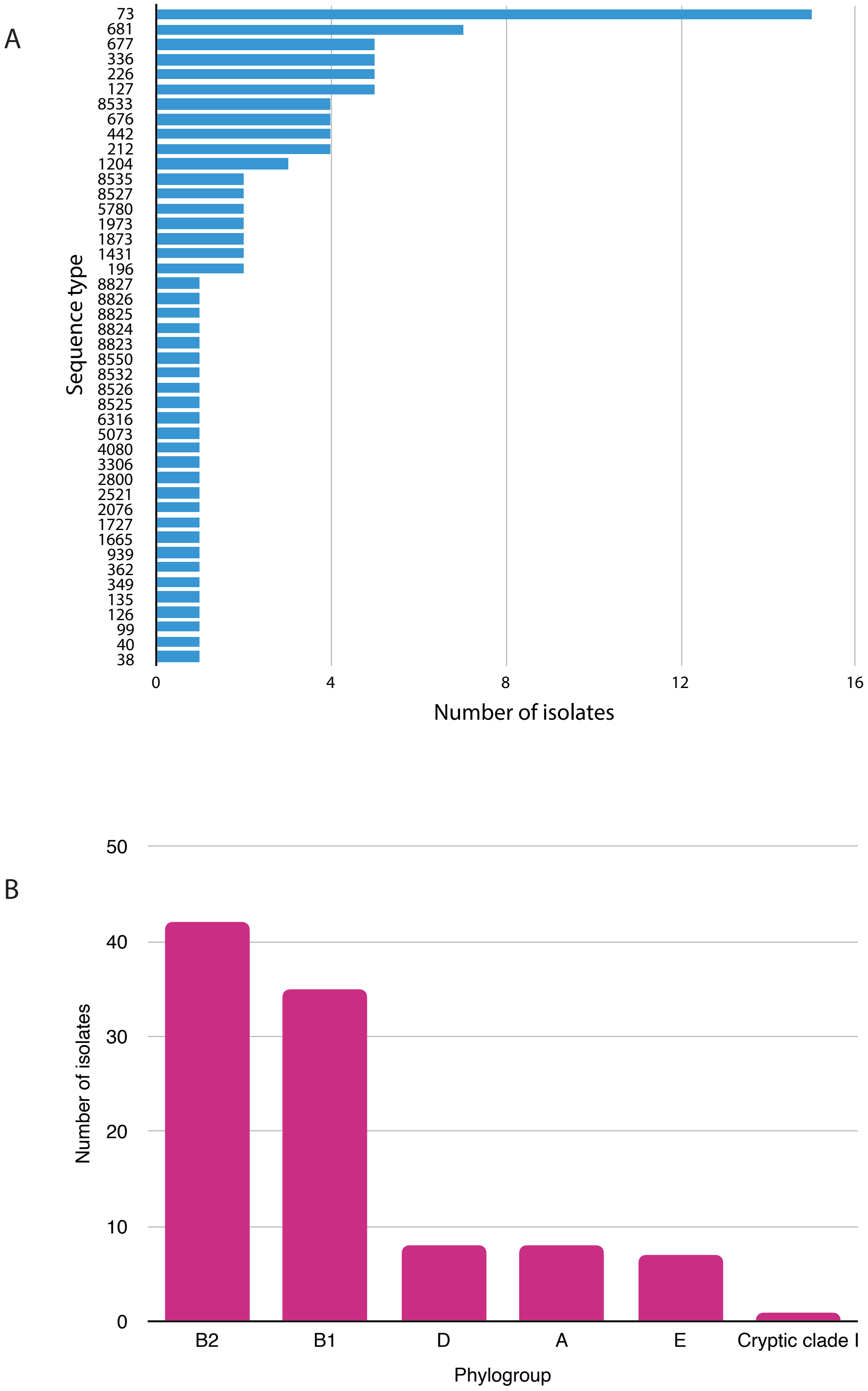


***Figure 3.4:*** *Distribution of sequence types (STs) (A) and phylotypes (B) among the study isolates.*

***Table 3.3**: Characteristics of the primate population, showing the number of recovered sequence types per individual.*

| Sample ID | Gender | Age | Colony picks | Recovered STs (colonies per ST) |
|---|---|---|---|---|
| PapRG-03 | F | Adult | 5 | ST336 (n=5) |
| PapRG-04 | M | Adult | 4 | ST1204 (n=2), ST8826* (n=1), ST1665 (n=1) |
| PapRG-05 | U | Juvenile | 4 | ST1431 (n=2), ST99 (n=1), ST6316 (n=1) |
| PapRG-06 | M | Adult | 5 | ST8827* (n=1), ST8525*, ST4080 (n=1), ST2521 (n=1), ST1204 (n=1) |
| ProbRG-07 | F | Adult | 5 | ST73 (n=5) |
| ChlosRG-12 | M | Adult | 4 | ST8824 (n=1), ST196 (n=2), ST40 (n=1) |
| ChlosAN-13 | F | Adult | 4 | ST8550* (n=1), ST8526* (n=1), ST1973 (n=2) |
| PapAN-14 | F | Adult | 5 | ST226 (n=3), ST2076 (n=1), ST939 (n=1) |
| PapAN-15 | F | Adult | 5 | ST8823 (n=1), ST5073 (n=1), ST226 (n=2), ST126 (n=1) |
| ChlosAN-17 | F | Juvenile | 4 | ST362 (n=1), ST681 (n=3) |
| ChlosAN-18 | F | Juvenile | 5 | ST681 (n=4), ST349 (n=1) |
| ProbAN-19 | F | Adult | 1 | ST8825* (n=1) |
| ChlosBP-21 | F | Adult | 5 | ST677 (n=4) |
| ChlosBP-23 | F | Adult | 3 | ST8527* (n=2), ST3306 (n=1) |
| ChlosBP-24 | M | Adult | 5 | ST73 (n=5) |
| ChlosBP-25 | U | Adult | 5 | ST3 (n=5) |
| ChlosM-29 | U | Adult | 2 | ST1873 (n=2) |
| PapM-31 | F | Adult | 5 | ST2800 (n=1), ST1727 (n=1), ST5780 (n=2), ST135 (1) |
| PapM-32 | F | Adult | 5 | ST8532* (n=1), ST212 (n=4) |
| PapM-33 | M | Adult | 5 | ST8533* (n=4), ST38 (n=1) |
| PapM-34 | M | Adult | 4 | ST676 (n=4) |
| PapM-36 | F | Adult | 2 | ST8535* (n=2) |
| PapKW-44 | U | Adult | 4 | ST442 (n=4) |
| ProbK-45 | F | Adult | 5 | ST127 (n=5) |
| Total | | | 101 | |

Novel sequence types are designated by an asterisk (*).

***Table 3.4:*** *Within-host single nucleotide polymorphism diversity between multiple genomes of the same ST recovered from the same monkey.*

| Sample ID | STs (colonies per ST) | Pair-wise SNP distances between multiple colonies of the same ST | Comment(s) |
|---|---|---|---|
| PapRG-03 | 336 (n=5) | 0-2 | |
| PapRG-04 | 1204 (n=2) | 4 | |
| PapRG-05 | 1431 (n=2) | 0 | |
| ProbRG-07 | 73 (n=5) | 0-1 | |
| ChlosRG-12 | 196 (n=2) | 25 | |
| PapAN-14 | 226 (n=3) | 1 | |
| PapAN-15 | 226 (n=2) | 1 | |
| ChlosAN-17 | 681 (n=3) | 0-3 | |
| ChlosAN-18 | 681 (n=4) | 0 | |
| ChlosBP-21 | 677 (n=4) | 5 | |
| ChlosBP-23 | 8527 (n=2) | 0 | |
| ChlosBP-24 | 73 (n=5) | 0-5 | |
| ChlosBP-25 | 73 (n=5) | 0-79 | Please see Table 3.4 |
| PapM-32 | 212 (n=4) | 0 | |
| PapM-33 | 8533 (n=4) | 0-4 | |
| PapM-34 | 676 (n=4) | 0-1 | |
| PapM-36 | 8535 (n=2) | 0-1 | |
| PapKW-44 | 442 (n=4) | 1-2 | |
| ProbK-45 | 127 (n=5) | 0-4 | |

*Red box highlights two monkeys in whom pair-wise SNP comparisons suggested multiple infection events (See Table 3.4).*

***Table 3.5:*** *Within-host diversity in green monkey 25 (ChlosBP-25).*

| Sample ID | Clone designation | | |
|---|---|---|---|
| ChlosBP-25-1 | 1 | | |
| ChlosBP-25-2 | 2 | | |
| ChlosBP-25-3 | 2 | | |
| ChlosBP-25-4 | 2 | | |
| ChlosBP-25-5 | 3 | | |
| **Pair-wise SNP distances between clones** | | | |
| | Clone 1 | Clone 2 | Clone 3 |
| Clone 1 | 0 | 12 | 79 |
| Clone 2 | 12 | 0 | 67 |
| Clone 3 | 79 | 67 | 0 |

### 3.3.5 Population structure of simian *E. coli* isolates

I identified the closest neighbours to all the recovered strains from my study (Table 3.5). My results suggest, in some cases, recent interactions between humans or livestock and non-human primates. However, I also found a diversity of strains specific to the non-human primate niche. Hierarchical clustering analysis revealed that simian isolates from ST442 and ST349 (Achtman)—sequence types that are associated with virulence and AMR in humans [544, 552]—were closely related to human clinical isolates, with differences of 50 alleles and seven alleles in the core-genome MLST scheme respectively (Figures 3.5 and 3.6). Similarly, I found evidence of recent interaction between simian ST939 isolates and strains from livestock (Figure 3.7)—with 40 cgMLST alleles (<40SNPs) separating the two genomes, representing less than eighteen years of divergence. Conversely, simian ST73, ST127 and ST681 isolates were genetically distinct from human isolates from these sequence types (Figures 3.8-3.10). The multi-drug resistant isolate PapAN-14-1 from ST2076 was, however, closely related to an environmental isolate recovered from water (Figure 3.11).

Five isolates were >1000 alleles away in the core-genome MLST scheme from anything in EnteroBase (Figures 3.12 & 3.13). Four of these were assigned to novel sequence types in the seven-allele scheme (Achtman) (ST8550, ST8525, ST8532, ST8826), while one belonged to ST1873, which has only two other representatives in EnteroBase: one from a species of wild bird from Australia (*Sericornis frontalis*); the other from water. In addition, ST8550, ST8525, ST8532, ST8826 belonged to novel HierCC 1100 groups (cgST complexes), indicating that they were unrelated to any other publicly available *E. coli* genomes.

In addition to my study isolates, I retrieved 94 *E. coli* genomes sourced from non-human primates from the rest of the world within EnteroBase: the US (83), Uganda (6), Kenya (4), Mexico (1). A total of 52 STs were found among primates from other parts of the world (Figure 3.14), four of which were also found among my study isolates (ST73, ST127, ST681 and ST939). Similar to what I observed among my monkey isolates, the most common ST among primates from the rest of the world was ST73 (11%). Also, most of the non-Gambian primate isolates belonged to phylogroup B2 (41%) and B1 (21%), consistent with what I found in my study population (Figure 3.14). Hierarchical clustering based on cgMLST types revealed clustering patterns that were largely consistent with the phylotype designations to which the primate isolates belonged. No discernible segregation of primate isolates based on geography was observed.

### 3.3.6 Prevalence of AMR-associated genes

I observed a modest level of genotypic antimicrobial resistance in my study population (Figure 3.3). The AMR-associated genotypes I found among the monkey isolates included *bla*-$_{EC}$ (beta-lactamase, penicillinase-type), *ant(3") (aadA1)* (streptomycin and spectinomycin), *aph3/aph6* ( neomycin and kanamycin), *DHFR* (trimethoprim), *sul1* (sulphonamides) and *tetA/B/C/D/R* (tetracyclines). A total of twenty-two isolates encoded resistance genes to a single antibiotic agent; twenty-two to two antibiotic classes and three isolates encoded resistance to three or more antibiotic classes. Pair-wise co-occurrence of AMR-associated genes in the same genome was sparse. The most common gene network was *bla-$_{EC}$-tetA/B/C/D/R* (12%), followed by *bla-$_{EC}$-ant(3") (aadA1)* (5%), *DFRA-tetA/B/C/D/R* (3%), then *ant(3")(aadA1)-DHFR* (2%).

Phenotypic resistance to single agents was confirmed in ten isolates: to trimethoprim in a single isolate, to sulfamethoxazole in four unrelated isolates and to tetracycline in four closely related isolates from a single animal. A single ST2076 (Achtman) isolate (PapAN-14-1) belonging to the ST349 lineage was phenotypically resistant to trimethoprim, sulfamethoxazole and tetracycline. The associated resistance genes were harboured on an IncFIB plasmid.

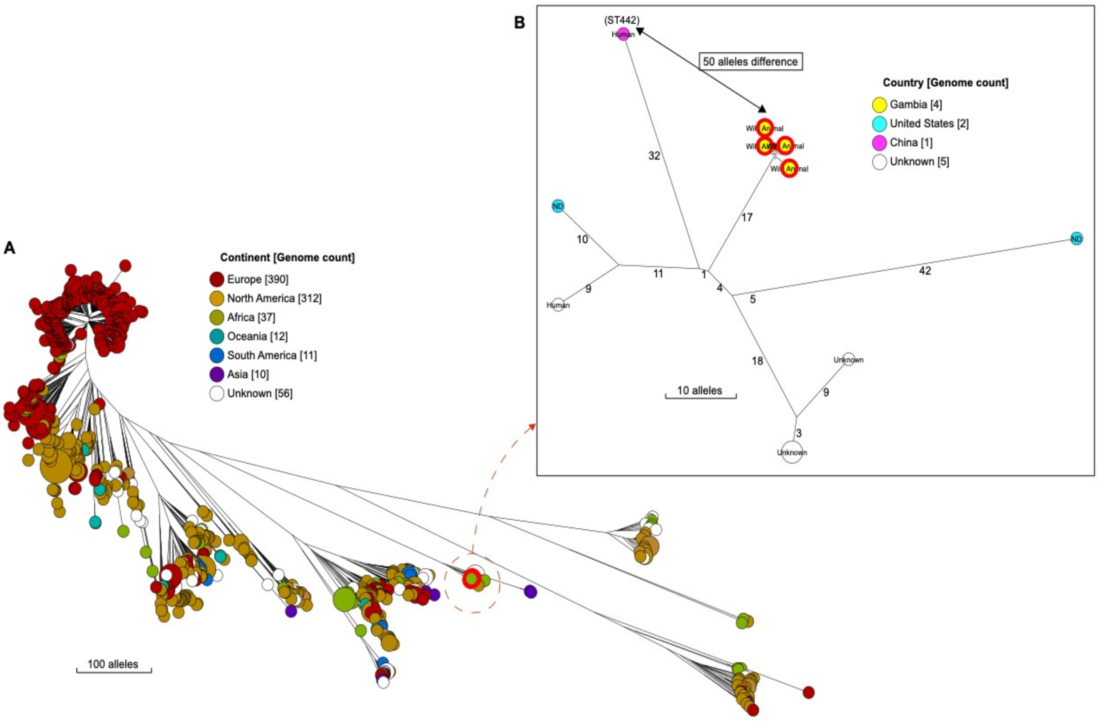


***Figure 3.5:*** *Population structure of ST442.*
A NINJA neighbour-joining tree showing the phylogenetic relationship between Achtman ST442 strains from this study and all other publicly available genomes that fell within the same HC1100 cluster (cgST complex). The locations of the isolates are displayed, with the genome count displayed in parenthesis. Branch lengths display the allelic distances separating genomes. Gambian strains from this study are highlighted in red. The sub-tree (B) shows the closest relatives to the study strains, with the allelic distance separating them displayed with the arrow.

***Table 3.6:*** *Genomic relationship between study isolates and publicly available E. coli genomes.*

| 7-allele ST | HC100 subgroups | Non-human primate host | Closest neighbours' source | Neighbours' country of isolation | Allelic distance |
|---|---|---|---|---|---|
| 349 | - | *Chlorocebus sabaeus* 18 | Human (bloodstream infection) | Canada | 7 |
| 2076 | - | *Papio papio* 14 | Environment (water) | Unknown | 25 |
| 939 | - | *Papio papio* 14 | Livestock | US | 40 |
| 442 | - | *Papio papio* 44 | Human | China | 50 |
| 2800 | - | *Papio papio* 31 | Unknown | Vietnam | 59 |
| 1973 | - | *Chlorocebus sabaeus* 13 | Unknown | Unknown | 64 |
| 8533 | - | *Papio papio* 33 | Environment (water) | Unknown | 69 |
| 6316 | - | *Papio papio* 05 | Human | Kenya | 97 |
| 1727 | - | *Papio papio* 34 | Human | Kenya | 98 |
| 676 | - | *Papio papio* 34 | Human (bloodstream infection) | UK | 98 |
| 8823 | - | *Papio papio* 15 | Rodent (guinea pig) | Kenya | 101 |
| 1431 | - | *Papio papio* 05 | Human | US | 109 |
| 5073 | - | *Papio papio* 15 | Human | US | 112 |
| 226 | 73641 | *Papio papio* 14 | Human | Tanzania | 112 |
| 8827 | - | *Papio papio* 06 | Human | Unknown | 122 |
| 1204 | 83197 | *Papio papio* 04 | Livestock | Japan | 127 |
| 1204 | 83197 | *Papio papio* 04 | Livestock | Japan | 130 |
| 677 | - | *Chlorocebus sabaeus* 21 | Human | US | 132 |
| 40 | - | *Chlorocebus sabaeus* 12 | Human | UK | 137 |
| 1204 | 83164 | *Papio papio* 06 | Livestock | Japan | 173 |
| 99 | - | *Papio papio* 05 | Human | UK | 180 |
| 362 | - | *Chlorocebus sabaeus* 17 | Food | Kenya | 180 |
| 8825 | - | *Piliocolobus badius* 19 | Human | France | 189 |
| 336 | - | *Papio papio* 03 | Poultry | Kenya | 189 |
| 73 | - | *Chlorocebus sabaeus* 24 | Human | Sweden | 189 |
| 196 | - | *Chlorocebus sabaeus* 12 | Human | Sweden | 197 |
| 2521 | - | *Papio papio* 06 | Livestock | US | 201 |
| 127 |  | *Pioliocolobus badius* 45 | Companion animal | US | 229 |
| 681 |  | *ChlosAN* 17 | Human | Norway | 251 |
| 38 | - | *Papio papio* 33 | human | UK | 265 |
| 135 | - | *Papio papio* 31 | Poultry | US | 281 |
| 8824 | - | *Chlorocebus sabaeus* 12 | Environmental* | US | 296 |
| 226 | 100039 | *Papio papio* 14 | Human | Sri Lanka | 318 |
| 8527 | - | *Chlorocebus sabaeus* 23 | Human | Kenya | 323 |
| 8535 | - | *Papio papio* 36 | Environmental (soil) | US | 368 |
| 1665 | - | *Papio papio* 04 | Livestock | UK | 371 |
| 4080 | - | *Papio papio* 06 | Human | Denmark | 507 |
| 8526 | - | *Chlorocebus sabaeus* 13 | Livestock | US | 708 |
| 8532 | - | *Papio papio* 32 | Non-human primate | Gambia (PapM-31-3) | 1102 |
| 8826 | - | *Papio papio* 04 | Livestock | Mozambique | 1255 |
| 8525 | - | *Papio papio* 06 | Livestock/companion animal | Switzerland | 1659 |
| 1873 | - | *Chorocebus sabaeus* 29 | Environment | US | 1685 |
| 8550 | - | *Chlorocebus sabaeus* 13 | Unknown | Unknown | 2006 |

*Source details unknown.

Isolates from humans were recovered from stools, except where indicated otherwise.

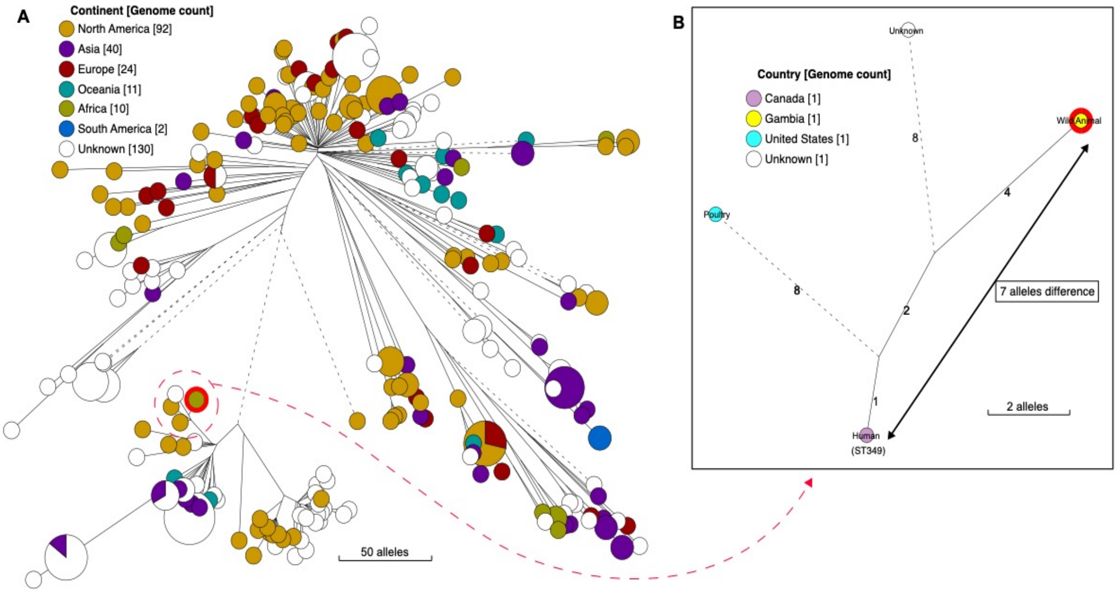
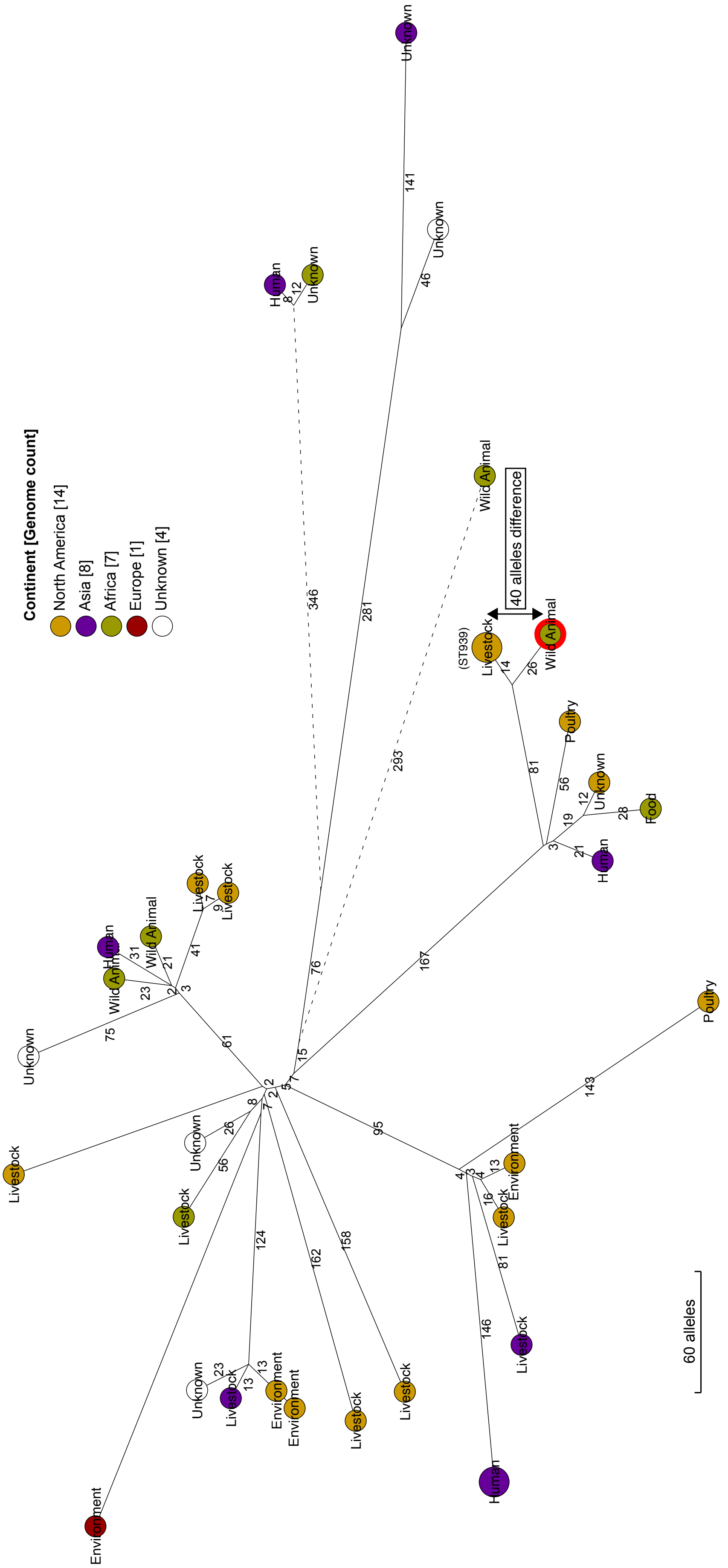


***Figure 3.6:*** *Population structure of ST349.*

A NINJA neighbour-joining tree showing the phylogenetic relationship between the ST349 (Achtman) strain from this study and all other publicly available genomes within the same HC1100 cluster (cgST complex). The legend shows the locations of the isolates, with genome counts displayed in parenthesis. Gambian strains are highlighted in red. The study ST349 strain is separated from a clinical ST349 strain by only seven alleles (<7 SNPs), as depicted in the subtree **(B)**. Long branches are shortened (indicated by dotted lines).



*A phylogenetic neighbour-joining tree reconstructed with the study ST939 (Achtman) strain and all publicly available genomes that fell within the same HC1100 cluster (cgST complex). The legend shows the locations of the isolates, with red highlights around the nodes indicating the Gambian strains. The allelic distance between the study strain and its nearest relative, a bovine ST939 strain, has been given, depicted by the arrow. Dotted lines indicate shortened long branches.*

***Figure 3.7:*** *Population structure of ST939.*

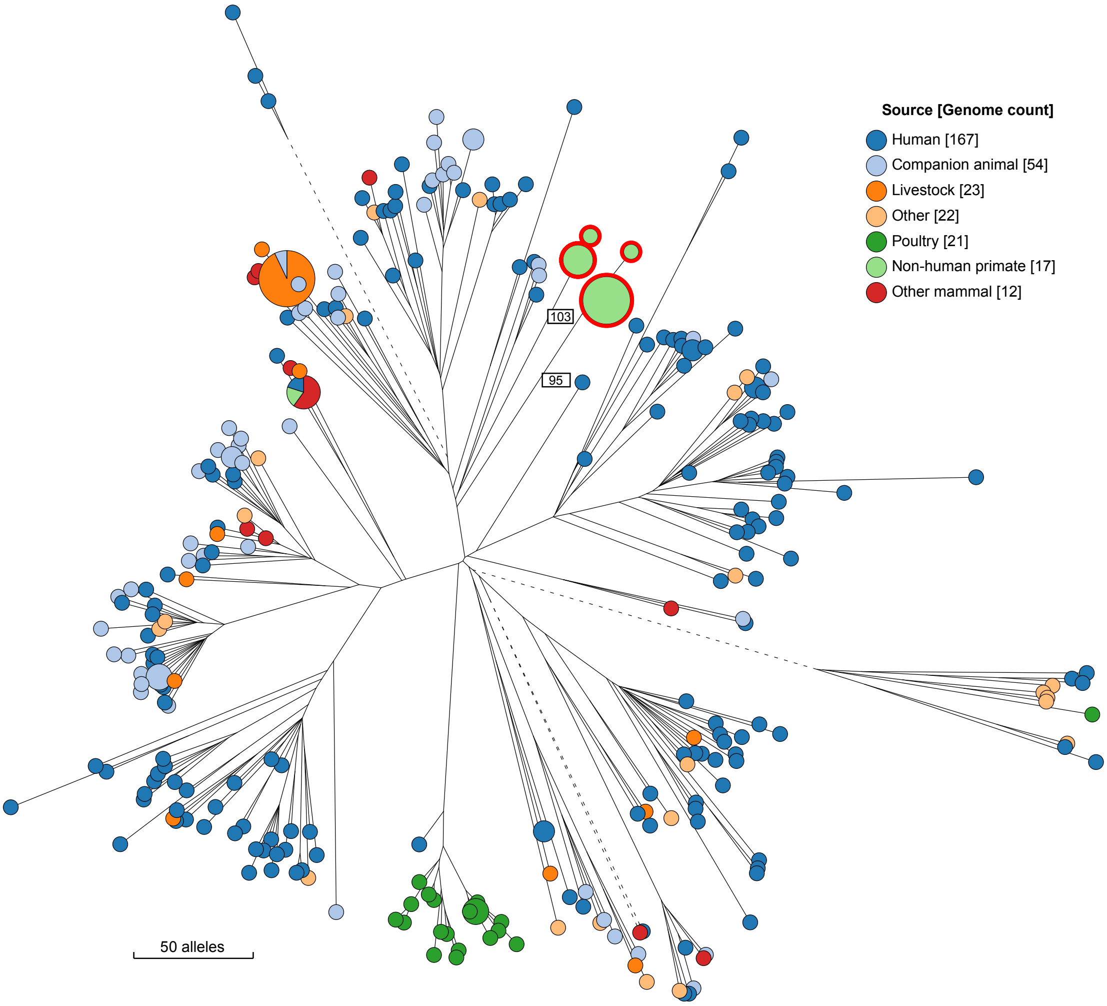


***Figure 3.8:*** *Population structure of ST73.*

NINJA neighbour-joining tree reconstructed with Achtman ST73 colibactin+ strains from this study and all other publicly available ST73 (Achtman) strains that fell within the same HC1100 cluster (cgST complex) in EnteroBase (Reference 139). The sources of the isolates are displayed, with Gambian strains highlighted in red. The Gambian non-human primate strains are on separate long branches, although nested within clades populated by human strains from other countries, suggestive of probably an ancient transmission between the two hosts. The branch lengths for the Gambian strains are displayed. Dotted lines represent long branches which have been shortened.

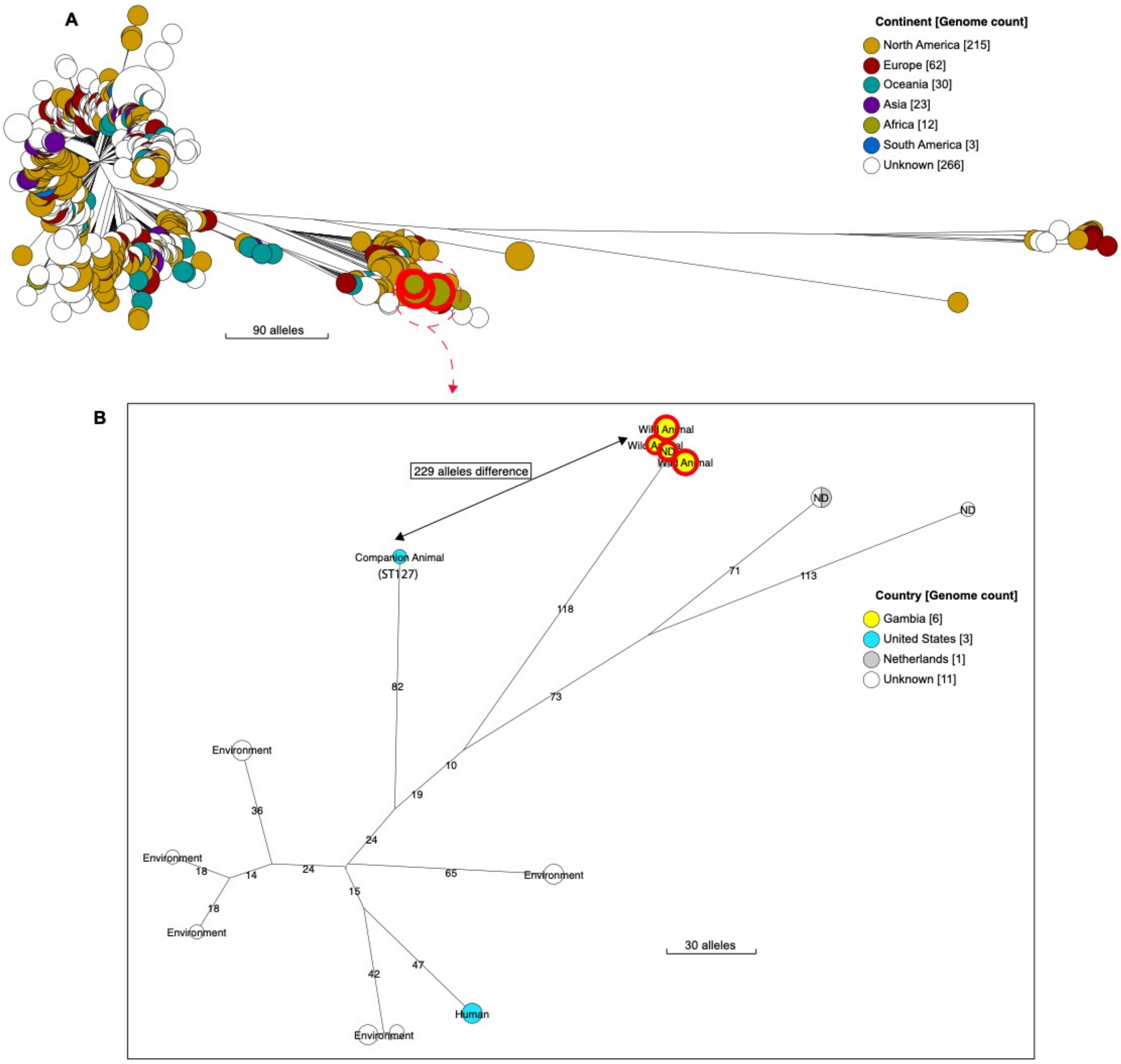


***Figure 3.9:*** *Population structure of ST127.*

A NINJA neighbour-joining tree showing the phylogenetic relationship between ST127 strains from this study and other publicly available strains that occur within the same HC1100 cluster (cgST complex). The sources of the isolates are displayed in the legends, with Gambian strains highlighted in red. Branch lengths display the allelic distances separating genomes. The sub-tree **(B)** shows the closest relatives to the study strains, with the allelic distances separating them displayed with the arrow.

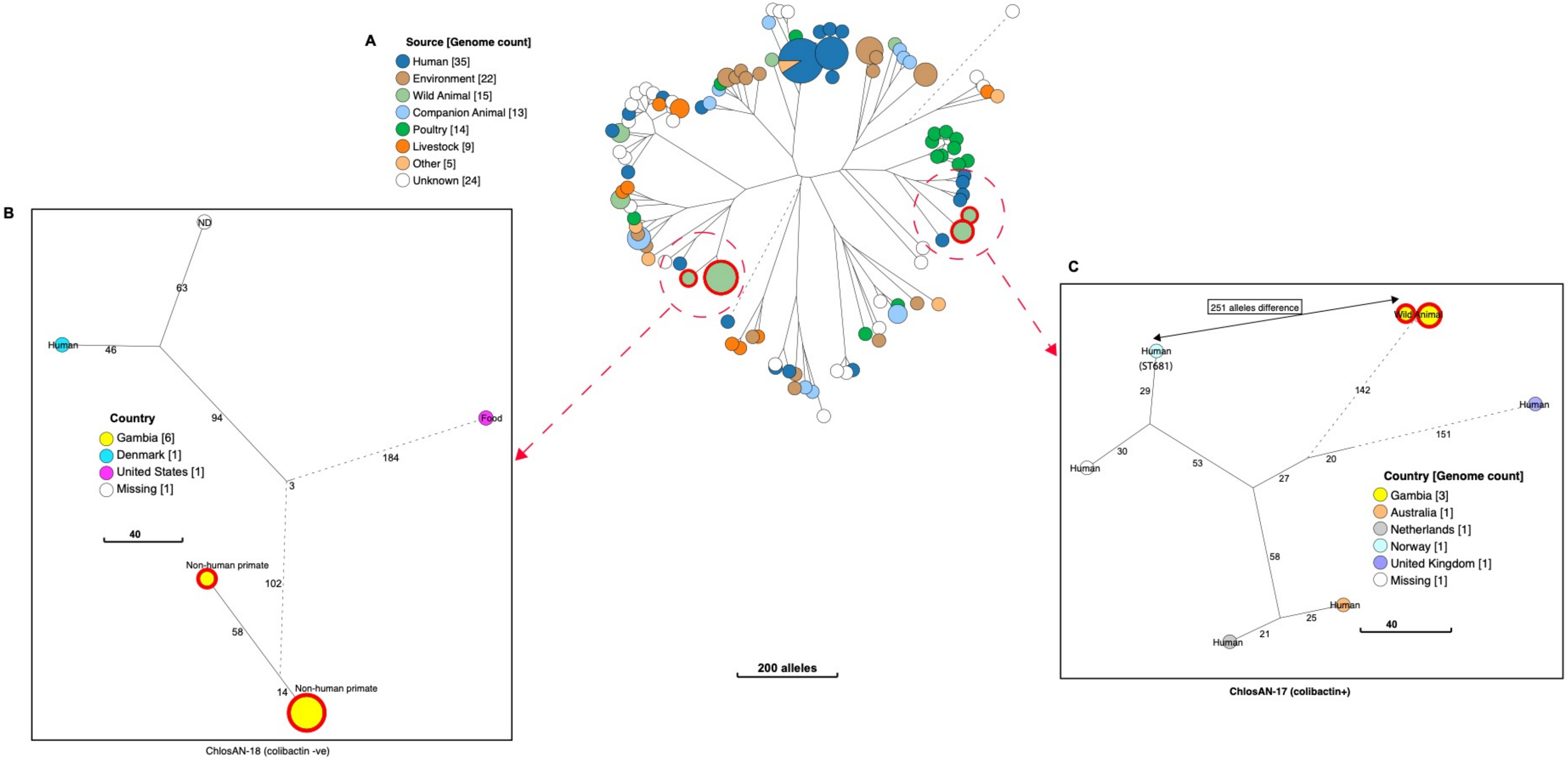


***Figure 3.10:*** *Population structure of ST681.*

*A Ninja neighbour-joining tree showing the phylogenetic relationship between ST681 strains from this study and other publicly available strains that fell within the same HC1100 cluster (cgST complex). The study strains fell into two separate HC100 clusters, which are depicted in the two subtrees (B and C). The closest neighbours to both HC100 clusters are displayed, with the branch labels indicating the allelic distances between strains. The locations of the isolates are displayed for each tree, with Gambian strains highlighted in red. Dotted lines represent long branches which have been shortened.*

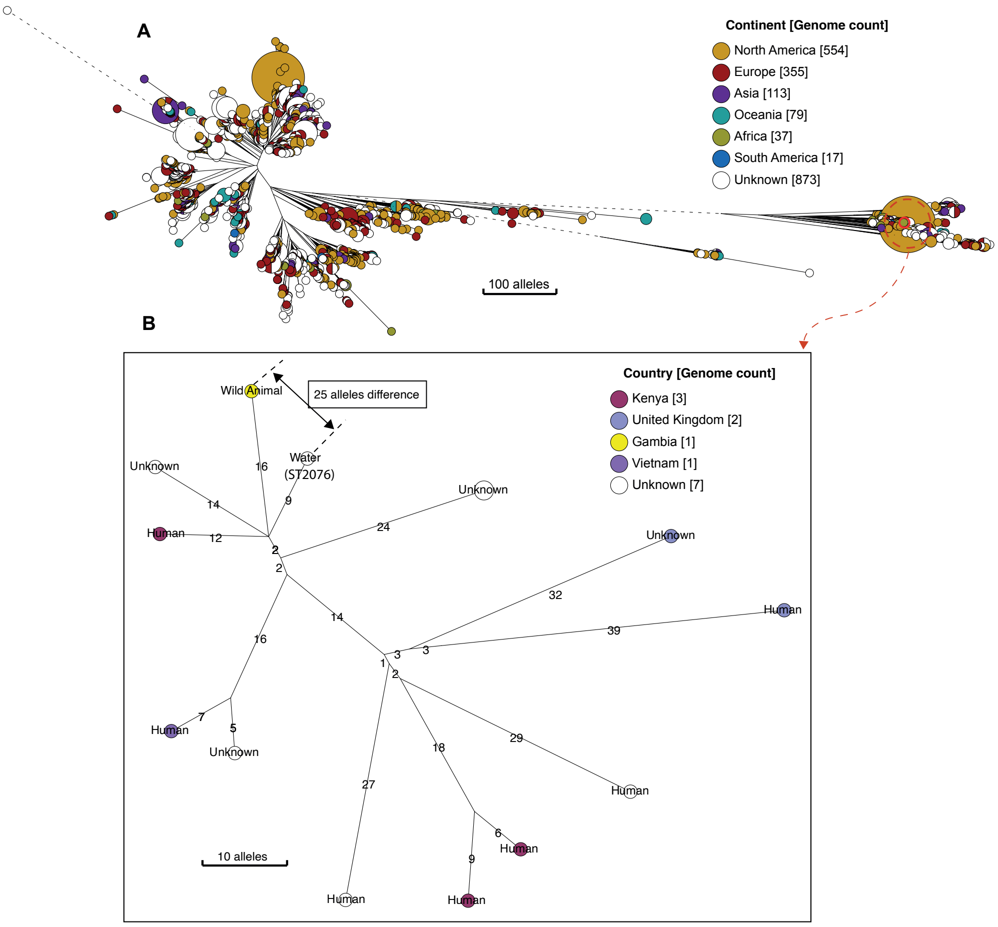
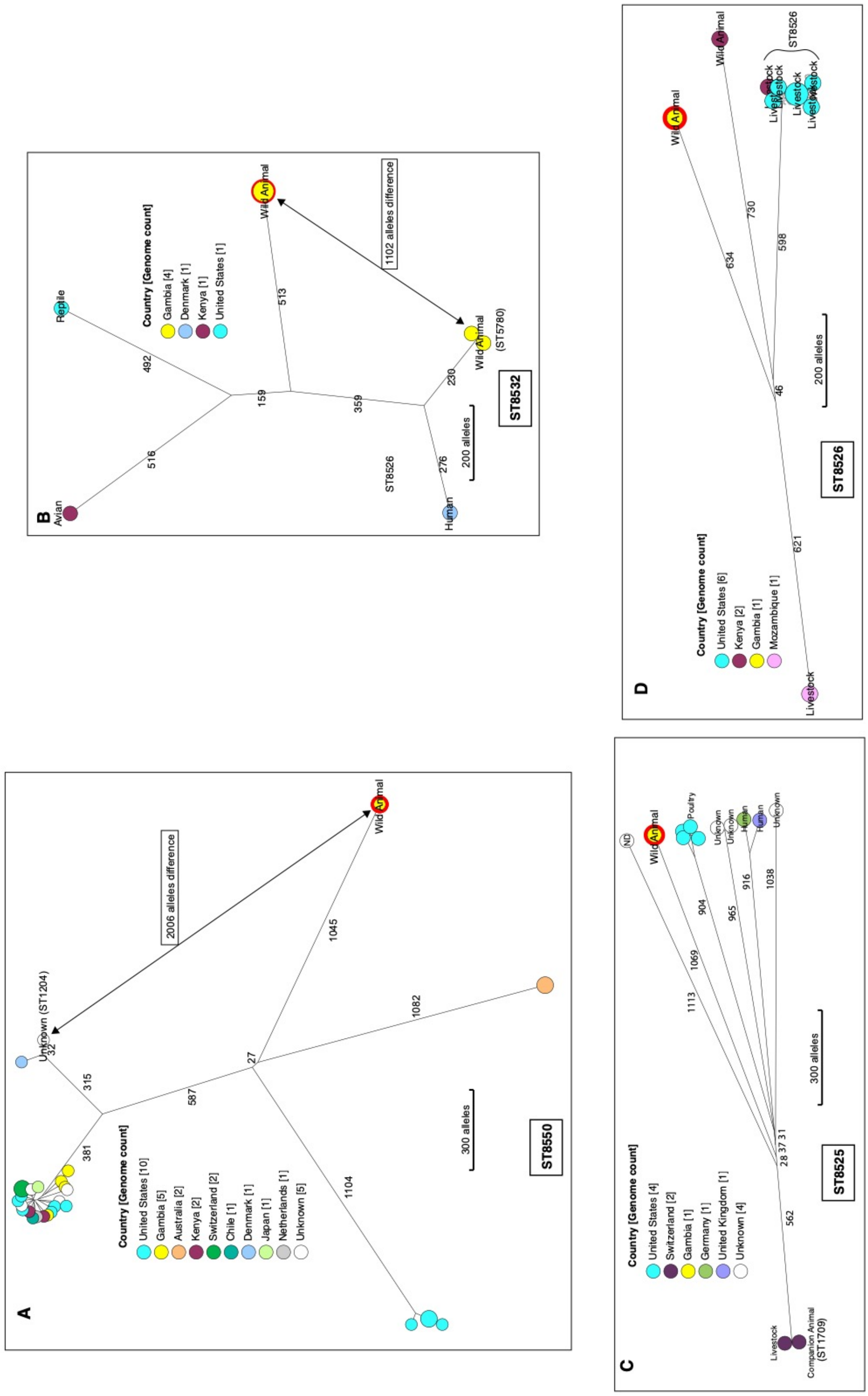


***Figure 3.11:*** *Population structure of ST2076.*

A phylogenetic tree showing the phylogenetic relationship between ST2076 strain (an MDR strain) and all other publicly available genomes that fell within the same HC1100 cluster (cgST complex). The legend shows the locations of the isolates, Gambian strains are highlighted in red. The subtree **(B)** shows the allelic distance between the study strain and its nearest relative, an ST2076 isolate recovered from water. Dotted lines indicate shortened long branches.



*Ninja phylogenetic trees showing the closest neighbours to simian isolates belonging to novel sequence types (Achtman) ST8550 (**A**), ST8532 (**B**) and ST8525 (**C**), ST8826 (**D**). The allelic distances between these study isolates and their closest neighbours are >1100 alleles, and the closest neighbours belong to seven-allele STs which share less than five out of the seven MLST loci. Each genome (ST8550, ST8532, ST8525) belongs to a unique cgST complex (novel groups at HierCC 1100), indicative of novel diversity within the non-human primate niche.*

***Figure 3.12:*** *NINJA phylogenetic trees depicting the closest neighbours to my study strains ST8550, ST8532 and ST8525.*

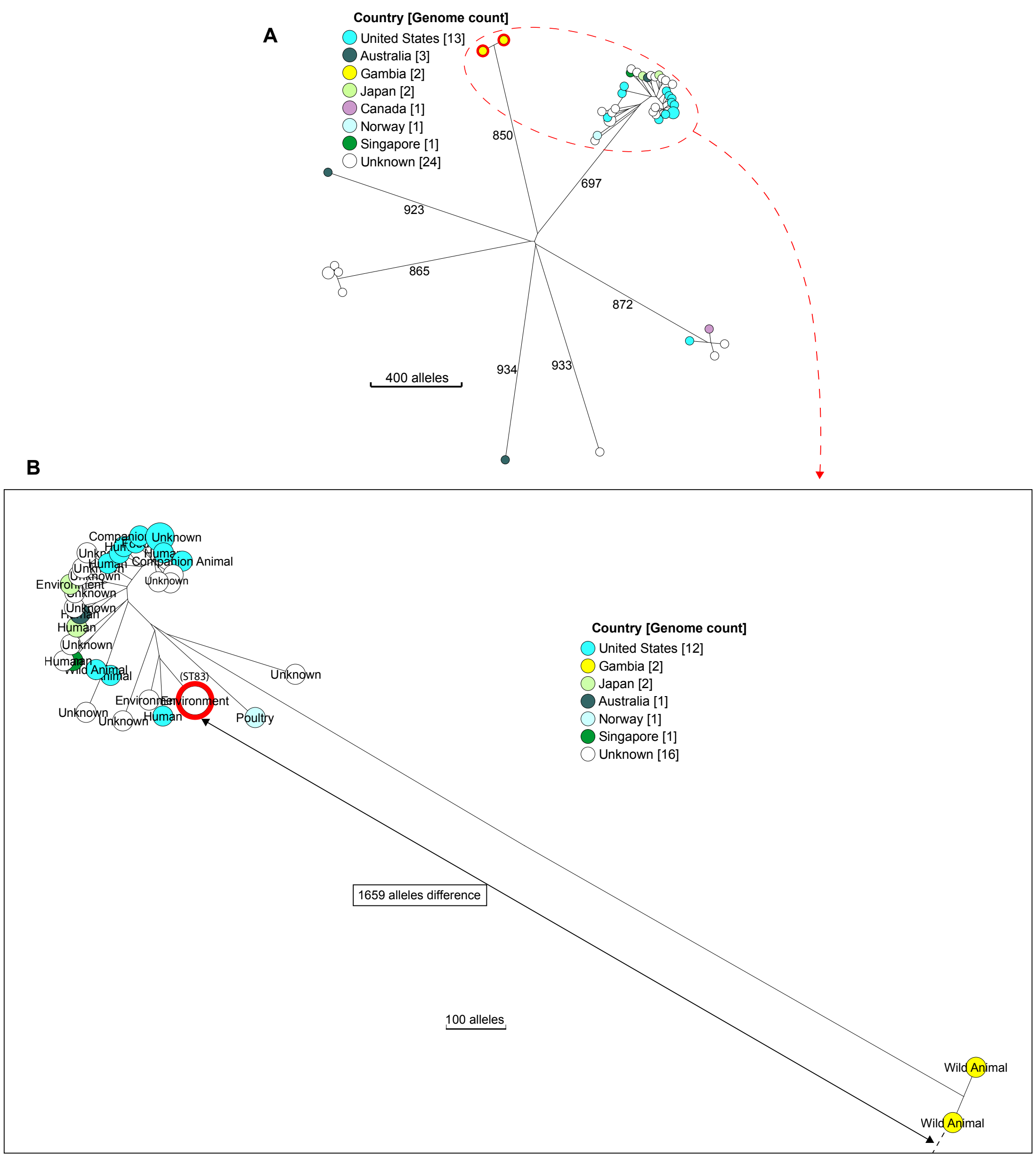


***Figure 3.13:*** *Population structure of ST1873.*

A NINJA phylogenetic tree showing the closest neighbours of simian ST1873 strain—an environmental (soil) isolate belonging to ST83, separated from the study strain by 1659 alleles. The legends of both the main tree and the subtree show the locations of the isolates Gambian strains are highlighted in red. In the subtree **(B)**, the closest neighbour to the simian ST1873 strain is also highlighted in red.

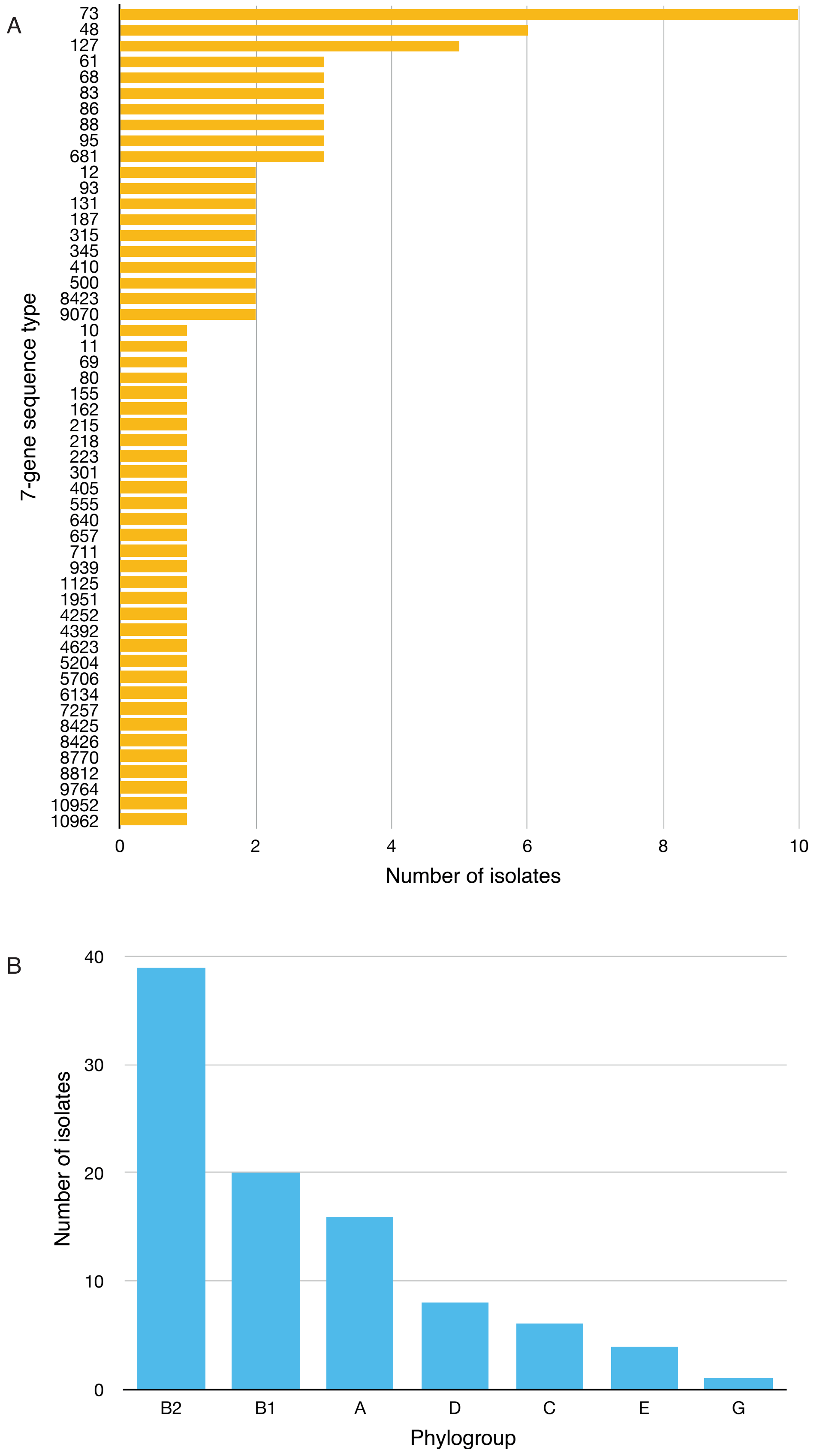


***Figure 3.14:*** *Distribution of sequence types (A) and phylogroups (B) among publicly available non-human primate E. coli isolates from the rest of the world.*

The balance of the study isolates were completely susceptible to the antibiotics tested. The genotypic resistance predictions were largely concordant with the results of phenotypic testing (range 90-99%, Table 3.6-3.8).

***Table 3.7:*** *Contingency and descriptive statistics describing the concordance between the genotypic resistance predictions and the results of phenotypic MIC tests.*

| *DHFR* | | Phenotype | |
|---|---|---|---|
| | | Negative | Positive |
| Genotype | Negative | 98 | 0 |
| | Positive | 1 | 2 |
| *sul1* | | Phenotype | |
| | | Negative | Positive |
| Genotype | Negative | 96 | 4 |
| | Positive | 0 | 1 |
| *tetA/B/C/D/R* | | Phenotype | |
| | | Negative | Positive |
| Genotype | Negative | 86 | 0 |
| | Positive | 10 | 5 |

***Table 3.8:*** *Percentage concordance between antimicrobial resistance genotype and phenotype.*

| | Overall | G+ among P+ | P+ among G+ | G-among P- | P- among G- |
|---|---|---|---|---|---|
| *DHFR* | 99.0 | 100 | 77.6 | 98.9 | 100 |
| *sul1* | 96.0 | 20 | 100 | 100 | 95.9 |
| *tetA/B/C/D/R* | 89.8 | 100 | 33.3 | 89.4 | 100 |

**Overall,** percentage of cases for which genotype and phenotype agree; **G+ amongst P+,** percentage of the positive phenotypes that are confirmed by genotyping; **P+ amongst G+,** percentage of the positive genotypes that showed a positive phenotype; **G- amongst P-,** percentage of the negative phenotypes that are confirmed by genotyping; **P- amongst G-,** percentage of the negative genotypes that showed a negative phenotype

***Table 3.9****: Fisher's exact tests on contingency tables.*

| | p-value |
|---|---|
| *DHFR* | 0.00062 |
| *sul1* | 0.05051 |
| *tetA/B/C/D/R* | 0.00004 |

A higher level of genotypic antimicrobial resistance was found in *E. coli* isolates from humans in the Gambia, compared to what prevails in the monkey isolates (Figure 3.15). Notably, a range of beta-lactamase resistance genes were found among *E. coli* from humans in the Gambia ($bla_{OXA-1}$, $bla_{TEM-1B}$, $bla_{TEM-1C}$, $bla_{SHV-1}$), while only the $bla_{EC}$ gene occurred in my study isolates.

### 3.3.7 Prevalence of plasmid replicons

Eighty percent (81/101) of the study isolates harboured one or more plasmids. I detected the following plasmid replicon types: IncF (various subtypes), IncB/K/O/Z, I1, IncX4, IncY, Col plasmids (various subtypes) and plasmids related to p0111 (rep B) (Table 3.9). Long-read sequencing of six representative samples showed that the IncFIB plasmids encoded acquired antibiotic resistance, fimbrial adhesins and colicins (Table 3.10). Also, the IncFIC/FII, ColRNAI, Col156 and IncB/O/K/Z plasmids encoded fimbrial proteins and colicins. In addition, the IncX and Inc-I-Aplha encoded bundle forming pili *bfpB* and the heat-stable enterotoxin protein *StbB* respectively.

### 3.3.8 Polished assemblies of novel strains

I generated complete genome sequences of six novel sequence types of *E. coli* (ST8525, ST8527, ST8532, ST8826, ST8827 and ST8535) within the seven-allele scheme (Achtman) (Table 3.11). Although none of these new genomes encoded AMR genes, one of them (PapRG-04-4) contained an IncFIB plasmid encoding fimbrial proteins and a cryptic ColRNA plasmid. PHASTER identified thirteen intact prophages and four incomplete phage remnants (Table 3.12). Two pairs of genomes from Guinea baboons from different parks shared common prophages: one pair carrying PHAGE_Entero_933W, the other PHAGE-Entero_lambda.

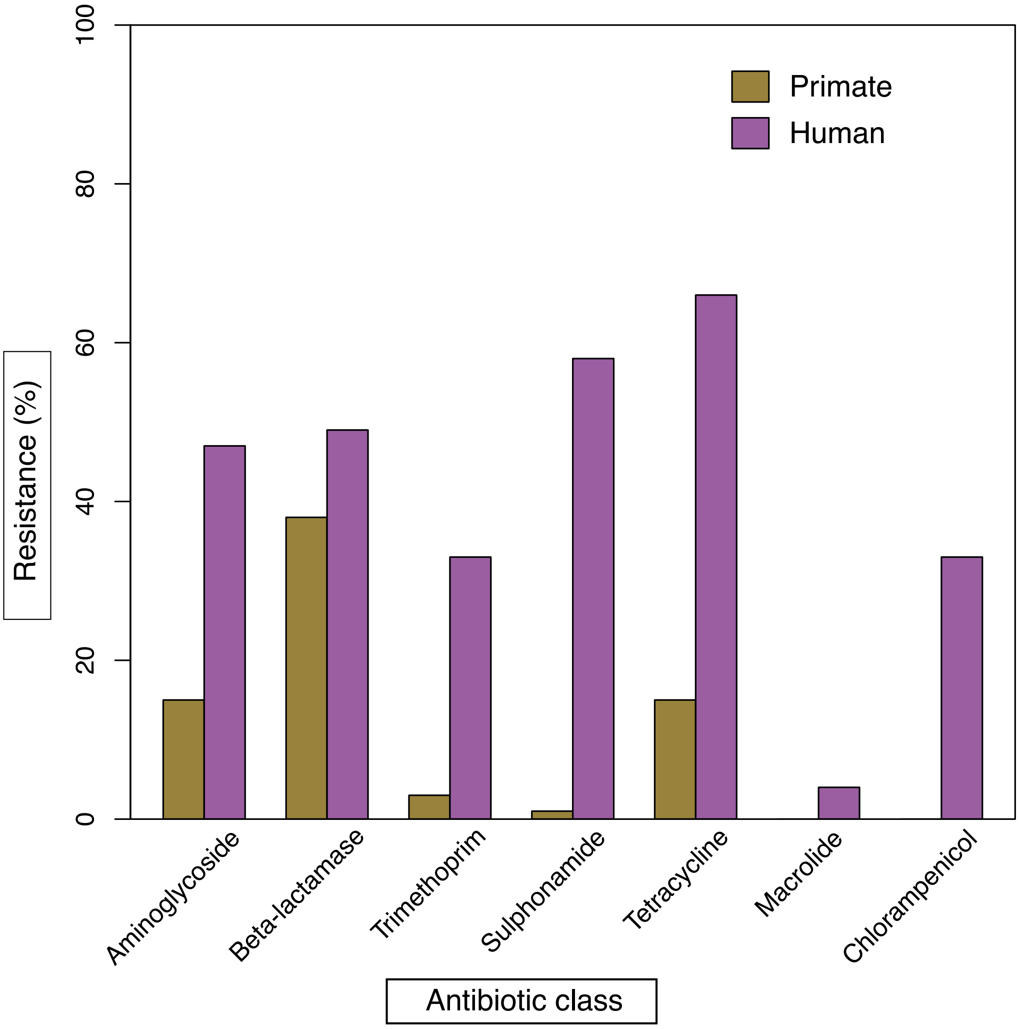


***Figure 3.15:*** *Comparison of the prevalence of antimicrobial resistance genotypes in E. coli isolated from humans in the Gambia to that found among my study isolates.*

The antimicrobial resistance genes detected were as follows: Aminoglycoside: *aph(6)-Id*, *ant aac(3)-IIa*, *ant(3'')-Ia*, *aph(3'')-Ib*, *aadA1*, *aadA2*; Beta-lactamase: *bla*$_{OXA-1}$, *bla*$_{TEM-1B}$, *bla*$_{TEM-1C}$, *bla*$_{SHV-1}$; Trimethoprim: *dfrA*; Sulphonamide: *sul1*, *sul2*; Tetracycline: *tet(A)*, *tet(B)*, *tet(34)*, *tet(D)*; Macrolide, *mph(A)*; Chloramphenicol, *catA1*. Screening of resistance genes was carried out using ARIBA ResFinder and confirmed by ABRicate (https://github.com/tseemann/abricate). A percentage identity of ≥ 90% and coverage of ≥ 70% of the respective gene length were taken as a positive result.

***Table 3.10:*** *Plasmid types detected among the study isolates.*

| Sample ID | Number detected | Predicted plasmids |
|---|---|---|
| PapRG-03-1 | 1 | IncFIB(AP001918) |
| PapRG-03-2 | 1 | IncFIB(AP001918) |
| PapRG-03-3 | 1 | IncFIB(AP001918) |
| PapRG-03-4 | 1 | IncFIB(AP001918) |
| PapRG-03-5 | 1 | IncFIB(AP001918) |
| PapRG-04-1 | 1 | IncFIB(AP001918) |
| PapRG-04-2 | 0 | |
| PapRG-04-4 | 2 | IncFIB(AP001918), ColRNAI |
| PapRG-04-5 | 0 | |
| PapRG-05-2 | 2 | IncFIA(HI1), IncFIB(pB171) |
| PapRG-05-3 | 0 | |
| PapRG-05-4 | 1 | IncFIB(AP001918) |
| PapRG-05-5 | 2 | IncFIA(HI1), IncFIB(pB171) |
| PapRG-06-1 | 0 | |
| PapRG-06-2 | 1 | IncFIB(AP001918) |
| PapRG-06-3 | 0 | |
| PapRG-06-4 | 1 | IncY |
| PapRG-06-5 | 0 | |
| ProbRG-07-1 | 2 | IncB/O/K/Z, /incFII(pCRY) |
| ProbRG-07-2 | 2 | IncB/O/K/Z, /incFII(pCRY) |
| ProbRG-07-3 | 2 | IncB/O/K/Z, /incFII(pCRY) |
| ProbRG-07-4 | 2 | IncB/O/K/Z, /incFII(pCRY) |
| ProbRG-07-5 | 2 | IncB/O/K/Z, /incFII(pCRY) |
| ChlosRG-12-1 | 0 | |
| ChlosRG-12-2 | 1 | IncFIB(AP001918) |
| ChlosRG-12-3 | 1 | IncFIB(AP001918) |
| ChlosRG-12-5 | 0 | |
| ChlosAN-13-1 | 2 | IncFIB(AP001918), IncFII(pCoo) |
| ChlosAN-13-2 | 1 | IncFIB(AP001918) |
| ChlosAN-13-4 | 0 | |
| ChlosAN-13-5 | 0 | |
| PapAN-14-1 | 2 | ColRNAI, IncFIB(AP001918) |
| PapAN-14-2 | 0 | |
| PapAN-14-3 | 7 | Col(BS512), ColRNAI, IncFIB(AP001918), IncFIA, IncFIB(pB171), IncFII(pHN7A8), IncFII |
| PapAN-14-4 | 7 | Col(BS512), ColRNAI, IncFIB(AP001918), IncFIA, IncFIB(pB171), IncFII(pHN7A8), IncFII |
| PapAN-14-5 | 7 | Col(BS512), ColRNAI, IncFIB(AP001918), IncFIA, IncFIB(pB171), IncFII(pHN7A8), IncFII |
| PapAN-15-1 | 6 | IncFIB(pB171), IncFIA, IncB/O/K/Z, IncFII([HN7A8), ColRNAI, Col(BS512) |
| PapAN-15-2 | 0 | |
| PapAN-15-3 | 7 | Col(BS512), ColRNAI, IncB/O/K/Z, IncFIA, IncFIB(pB171), IncFII(pHN7A8) |
| PapAN-15-4 | 0 | |
| PapAN-15-5 | 0 | |
| ChlosAN-17-1 | 1 | Col156 |
| ChlosAN-17-2 | 3 | IncFIB(AP001918), IncFII(pHN7A8), p0111 |
| ChlosAN-17-3 | 1 | Col156 |
| ChlosAN-17-4 | 1 | Col156 |
| ChlosAN-18-1 | 3 | IncFIB(AP001918), IncFII, IncY |
| ChlosAN-18-2 | 3 | IncFIB(AP001918), IncFII, IncY |
| ChlosAN-18-3 | 3 | IncFIB(AP001918), IncFII, IncY |
| ChlosAN-18-4 | 3 | IncFIB(AP001918), IncFII, IncY |
| ChlosAN-18-4 | 3 | IncFIB(AP001918), IncFII, IncY |
| ChlosAN-18-5 | 2 | IncFIB(AP001918), IncFII |
| ProbAN-19-2 | 2 | IncFIB(AP001918), IncFIC(FII) |
| ChlosBP-21-1 | 0 | |
| ChlosBP-21-2 | 0 | |
| ChlosBP-21-3 | 0 | |

***Table 3.9:*** *Plasmid types detected among the study isolates (continued).*

| | | |
|---|---|---|
| ChlosBP-21-4 | 0 | |
| ChlosBP-21-5 | 0 | |
| ChlosBP-23-1 | 0 | |
| ChlosBP-23-2 | 0 | |
| ChlosBP-23-3 | 1 | Col156 |
| ChlosBP-24-1 | 2 | Col156, IncY |
| ChlosBP-24-2 | 2 | Col156, IncY |
| ChlosBP-24-3 | 2 | Col156, IncY |
| ChlosBP-24-4 | 2 | Col156, IncY |
| ChlosBP-24-5 | 2 | Col156, IncY |
| ChlosBP-25-1 | 2 | Col156, IncY |
| ChlosBP-25-2 | 2 | Col156, IncY |
| ChlosBP-25-3 | 2 | Col156, IncY |
| ChlosBP-25-4 | 2 | Col156, IncY |
| ChlosBP-25-5 | 2 | Col156, IncY |
| ChlosM-29-1 | 3 | IncFIB(AP001918), ColRNAI, IncB/O/K/Z |
| ChlosM-29-2 | 3 | IncFIB(AP001918), ColRNAI, Col156 |
| PapM-31-1 | 0 | |
| PapM-31-2 | 2 | IncFIB(AP001918), IncFII(pHN7A8) |
| PapM-31-3 | 1 | IncFIB(AP001918) |
| PapM-31-4 | 3 | IncFIA(HI1), IncFIB(P001918), IncFII(pHN7A8) |
| PapM-31-5 | 1 | IncFIB(AP001918) |
| PapM-32-1 | 0 | Col156, IncFII(pSE11), IncI1_1_Alpha, IncX4 |
| PapM-32-2 | 4 | Col156, IncFII(pSE11), IncI1_1_Alpha, IncX4 |
| PapM-32-3 | 4 | Col156, IncFII(pSE11), IncI1_1_Alpha, IncX4 |
| PapM-32-4 | 4 | Col156, IncFII(pSE11), IncI1_1_Alpha, IncX4 |
| PapM-32-5 | 4 | Col156, IncFII(pSE11), IncI1_1_Alpha, IncX4 |
| PapM-33-1 | 3 | Col(MG828), IncFIB(AP001918), p0111 |
| PapM-33-2 | 3 | Col(MG828), IncFIB(AP001918), p0111 |
| PapM-33-3 | 3 | Col(MG828), IncFIB(AP001918), p0111 |
| PapM-33-4 | 0 | |
| PapM-33-5 | 3 | Col(MG828), IncFIB(AP001918), p0111 |
| PapM-34-1 | 1 | IncFIB(AP001918) |
| PapM-34-2 | 1 | IncFIB(AP001918) |
| PapM-34-3 | 1 | IncFIB(AP001918) |
| PapM-34-4 | 1 | IncFIB(AP001918) |
| PapM-36-1 | 1 | IncFIB(AP001918) |
| PapM-36-2 | 1 | IncFIB(AP001918) |
| PapKW-44-1 | 2 | ColRNAI, IncFIB(AP001918) |
| PapKW-44-2 | 2 | ColRNAI, IncFIB(AP001918) |
| PapKW-44-3 | 2 | ColRNAI, IncFIB(AP001918) |
| PapKW-44-4 | 2 | ColRNAI, IncFIB(AP001918) |
| ProbK-45-1 | 2 | Col156, IncB/O/K/Z |
| ProbK-45-2 | 2 | Col156, IncB/O/K/Z |
| ProbK-45-3 | 2 | Col156, IncB/O/K/Z |
| ProbK-45-4 | 2 | Col156, IncB/O/K/Z |
| ProbK-45-5 | 2 | Col156, IncB/O/K/Z |

***Table 3.11****: Characteristics and contents of plasmids as detected by long-read sequencing.*

| Plasmid type | Size (bp) | Non-human primate host | Virulence- or AMR-associated factors detected | Function |
|---|---|---|---|---|
| ColRNA | 5kb | PapRG-04-4 | - | |
| IncFIB(AP001918) | 168kb | PapRG-04-4 | Autotransporter adhesin *eha*G | Mediates biofilm formation and adhesion |
| IncB/O/K/Z | 95kb | PapRG-07-1 | *StbB* | Heat-stable enterotoxin |
| | | | Colicin-Ia | Bacteriocin |
| | | | Toxin-coregulated pilus E and T | Dissemination of plasmids and persistence of antibiotic resistance |
| IncFII(pCRY) | 28kb | PapRG-07-1 | Type IV secretion system protein *virB11* | Cytotoxic activity that is induced during stress |
| | | | Putative toxin *HigB2* | Toxic component of type II toxin-antitoxin system |
| IncFIB(AP001918) | 160kb | PapAN-14-1 | *sul1* | Sulphonamide resistance |
| | | | *ant (3”)_1a* | Aminoglycoside resistance |
| | | | *dfrA_1* | Trimethoprim resistance |
| | | | *tet*(*A*), *tet*(*B*), *tet*(*C*) | Tetracycline resistance proteins classes A, B and C |
| | | | *traA* | Pilin |
| | | | Colicin-Ia | Bacteriocin |
| ColMG828 | 4kb | PapAN-14-1 | mRNA interferase protein *relE* | Promotes cell death and growth arrest under stress conditions. |
| ColBS512 | 2kb | PapAN-15-1 | - | |
| IncFIB(pB171) | 146kb | PapAN-15-1 | Toxin *higB-2* | Toxic component of type II toxin-antitoxin system |
| IncFII(pHN7A8) | 62kb | PapAN-15-1 | Pilin protein *papB* | Adhesion |
| | | | K88 fimbrial protein AC | Adhesion |
| IncB/O/K/Z | 91kb | PapAN-15-1 | Type 1 fimbrial protein, *fimA* | Adhesion |
| | | | Putative fimbrial-like protein *elfG* | Adhesion |
| IncFIC/IncFII | 180kb | ChlosBP-25-1 | Colicin-Ia, colicin-M, colicin-A and colicin-B, | Bacteriocins |
| | | | P fimbrial pilin proteins, Fimbrial adhesins *pap*E, *papG*, *papK*, *prsF* and adhesin *yadA* | Adhesion |
| | | | S-fimbrial proteins | Adhesion |
| | | | *StbB* | Heat-stable enterotoxin |
| | | | *traA* Pilin protein | Adhesion |
| IncY | 66kb | ChlosBP-25-1 | Virulence regulon transcriptional activator (*virB*) | Transcription regulator for the invasion antigens *IpaB*, *IpaC* and *IpaD* |
| Col156 | 8kb | ChlosBP-25-1 | Colicin-E7, colicin-E2, lysis protein for colicin N | Bacteriocins |
| IncX | 87kb | PapM-32-4 | *bfpB* | Outer membrane lipoprotein required for biogenesis of bundle forming pili and EPEC adherence and autoaggregation |
| ColRNAI | 9kb | PapM-32-4 | Colicin E1 | Cytotoxic activity that is induced during stress |
| Inc-I-1-Alpha | 86kb | PapM-32-4 | *StbB* | Heat-stable enterotoxin |
| IncX | 34kb | PapM-32-4 | *virB1*, *virB4, virB8-virB11*, | Type IV secretion system proteins |

***Table 3.12:*** *Sequencing statistics of five novel strains sequenced by the Oxford Nanopore technology.*

| Assembly | #Chromosomal contigs | Total length (bp) | 7-allele ST | Length (chromosome) | Circularised | GC (%) | #Plasmidic contigs |
|---|---|---|---|---|---|---|---|
| ChlosBP-23-1 | 1 | 4769521 | 8527 | 4769521 | Yes | 50.44 | 0 |
| PapM-32-1 | 1 | 4866137 | 8532 | 4866137 | Yes | 50.57 | 0 |
| PapM-36-2 | 1 | 4863286 | 8535 | 4756164 | Yes | 50.66 | 1 |
| PapRG-04-4 | 1 | 5581923 | 8826 | 5581923 | Yes | 50.28 | 2 |
| PapRG-06-3 | 1 | 4703889 | 8827 | 4703889 | Yes | 50.74 | 0 |
| PapRG-06-5 | 1 | 4945561 | 8525 | 4945561 | Yes | 50.56 | 0 |
| **Assembly** | **Length (plasmid)** | | **Circularised** | **N50** | **CDSs** | **trNRAs** | **rRNAs** |
| ChlosBP-23-1 | | | | 4769521 | 4309 | 86 | 22 |
| PapM-32-1 | | | | 4866137 | 5241 | 80 | 22 |
| PapM-36-2 | 107122 | | Yes | 4756164 | 7818 | 90 | 22 |
| PapRG-04-4 | 167766 (IncFIB(AP001918); 4727 (ColRNAI) | | Yes | 5409430 | 5241 | 80 | 22 |
| PapRG-06-3 | | | | 4703889 | 4229 | 80 | 22 |
| PapRG-06-5 | | | | 4945561 | 4443 | 81 | 22 |

***Table 3.13:*** *Prophage sequences identified among the novel strains that were sequenced by Oxford Nanopore technology.*

| Sample ID | Region | Region Length | Region Position | Completeness | Score | #Total Proteins | Predicted Phage | GC % |
|---|---|---|---|---|---|---|---|---|
| PapRG-04-4 | 1 | 5.5Kb | 152364-157889 | Incomplete | 30 | 8 | PHAGE_Entero_933W_NC_000924 | 47.5 |
|  | 2 | 55.1Kb | 1597754-1652862 | Intact | 150 | 50 | PHAGE_Entero_P88_NC_026014 | 52.6 |
|  | 3 | 21.5Kb | 2098844-2120388 | Intact | 120 | 25 | PHAGE_Entero_P88_NC_026014 | 48.6 |
|  | 4 | 43.4Kb | 2457852-2501309 | Intact | 150 | 53 | PHAGE_Entero_cdtI_NC_009514 | 50.1 |
|  | 5 | 63.2Kb | 2845105-2908316 | Intact | 150 | 90 | PHAGE_Phage_Gifsy_1_NC_010392 | 49.4 |
|  | 6 | 58.5Kb | 4650128-4708686 | Intact | 150 | 72 | PHAGE_Shigel_SfII_NC_021857 | 49.4 |
| PapM-36-2 | 1 | 54.6Kb | 3512290-3566950 | Intact | 150 | 126 | PHAGE_Entero_phiP27_NC_003356 | 52.4 |
|  | 2 | 61.2Kb | 4413411-4474611 | Intact | 150 | 72 | PHAGE_Escher_pro147_NC_028896 | 51.1 |
| ChlosBP-23-1 | 1 | 10.3Kb | 1083628-1093936 | Incomplete | 10 | 7 | PHAGE_Escher_RCS47_NC_042128 | 51.6 |
|  | 2 | 11.2Kb | 1916437-1927681 | Incomplete | 40 | 16 | PHAGE_Salmon_118970_sal3_NC_031940 | 48.0 |
| PapM-32-1 | 1 | 5.5Kb | 187840-193365 | Incomplete | 30 | 9 | PHAGE_Entero_933W_NC_000924 | 47.5 |
|  | 2 | 60.5Kb | 2468395-2528918 | Intact | 150 | 62 | PHAGE_Entero_lambda_NC_001416 | 47.6 |
|  | 3 | 42.4Kb | 3156184-3198602 | Intact | 130 | 57 | PHAGE_Shigel_Sf6_NC_005344 | 47.7 |
|  | 4 | 49.9Kb | 3718827-3768727 | Intact | 150 | 53 | PHAGE_Shigel_SfII_NC_021857 | 50.9 |
| PapRG-06-3 | 1 | 30.9Kb | 1-30958 | Intact | 150 | 36 | PHAGE_Entero_P2_NC_001895 | 50.7 |
|  | 2 | 30.6Kb | 781886-812559 | Intact | 150 | 34 | PHAGE_Entero_lambda_NC_001416 | 48.4 |
|  | 3 | 11Kb | 3049355-3060369 | Intact | 107 | 16 | PHAGE_Entero_P4_NC_001609 | 48.9 |
| PapRG-06-5 | 1 | 27.9Kb | 1-27906 | Intact | 150 | 36 | PHAGE_Entero_P2_NC_001895 | 52.2 |
|  | 2 | 16.5Kb | 4342828-4359360 | Intact | 117 | 13 | PHAGE_Entero_P4_NC_001609 | 49.3 |

## 3.4 Discussion

In this Chapter, I have described the population structure of *E. coli* from diurnal non-human primates living in rural and urban habitats in the Gambia. Although my sample size was relatively small, I have recovered isolates that span the diversity previously described in humans and have also identified ten new sequence types (six of them now with complete genome sequences). This finding is significant, considering the vast number of *E. coli* genomes that have been sequenced to date (9,597 with MLST via sanger sequencing and 127, 482 via WGS as at 29th February 2020) [139].

Increasing contact between animal species facilitates the potential exchange of pathogens [553]. Accumulating data shows that ExPEC strains are frequently isolated from diseased companion animals and livestock—highlighting the potential for zoonotic as well as anthroponotic transmission [552, 554]. In a previous study, green monkeys from Bijilo Park were found to carry lineages of *Staphylococcus aureus* thought to be acquired from humans [537]. My analyses suggest some level of closeness between *E. coli* strains from humans and wild non-human primates—in one scenario, only seven cgMLST alleles separated a simian ST349 isolate from a human bloodstream isolate from Canada. This simian ST349 isolate was recovered from a green monkey in Abuko Nature Reserve, where tourists sometimes handfeed monkeys, despite prohibitions. A limitation of my study is that I could not sample *E. coli* from humans living in close proximity to the study primates. Comparisons between simian isolates and those from sympatric humans may shed light on possible transmission routes between humans and primates in this setting. My results also show that non-human primates harbour *E. coli* genotypes that are clinically important in humans, such as ST73, ST127 and ST681, yet are distinct from those circulating in humans—probably reflecting lineages that have existed in this niche for long periods.

I found that several monkeys were colonised with multiple STs, often encompassing two or more phylotypes. Colonisation with multiple serotypes of *E. coli* is common in humans [249, 413, 414, 505, 555]. My results indicate that a single monkey can carry as many as five STs. Sampling multiple colonies from single individuals also revealed within-host diversity arising from microevolution. However, I also found evidence suggesting acquisition in the same animal of multiple lineages of the same sequence type, although it is unclear whether this reflects a single transmission event involving more than one strain or serial transfers.

I found a relatively low prevalence of genotypic antimicrobial resistance among my study isolates, compared to the genotypic resistance observed among isolates sourced from humans in the Gambia—probably reflecting differing selective pressures from antibiotic use. The Gambia does not have national AMR surveillance data and background data on the use of antimicrobials is limited. However, a recent study on the aetiology of diarrhoea among children less than five years old reported the frequent use of trimethoprim/sulphamethoxazole in the treatment of diarrhoea in the Gambia  [515].  This probably accounts, at least in part, for the high rates of genotypic resistance to trimethoprim and sulphonamides among human *E. coli* isolates from the Gambia. The excretion of resistant bacteria and active antimicrobials from humans and domesticated animals and their persistence in the environment is known to facilitate the proliferation of AMR in the environment [296].

Antimicrobial resistance in wildlife is known to spread on plasmids through horizontal gene transfer [296, 556-558]. Given the challenge of resolving large plasmids using short-read sequences [559], I exploited long-read sequencing to document the contribution of plasmids to the genomic diversity that I observed in my study population. Consistent with previous reports [560], I found IncF plasmids which encoded antimicrobial resistance genes. Virulence-encoding plasmids, particularly colicin-encoding and the F incompatibility group ones, have long been associated with several pathotypes of *E. coli* [561]. Consistent with this, I found plasmids that contributed to the dissemination of virulence factors such as the heat-stable enterotoxin protein *StbB*, colicins and fimbrial proteins.

### 3.4.1 Limitations

This study could have been enhanced by sampling human populations living near those of my non-human primates; however, I compensated for this limitation by leveraging the wealth of genomes in publicly available databases. In addition, I did not sample nocturnal monkeys due to logistic challenges; however, these have more limited contact with humans than the diurnal species. Despite these limitations, however, my study provides insight into the diversity and population structure of *E. coli* among non-human primates in the Gambia, highlighting the impact of human continued encroachment on natural habitats and revealing important phylogenomic relationships between strains from humans and non-human primates.

# 4 CHAPTER FOUR: GENOMIC DIVERSITY OF *ESCHERICHIA COLI* ISOLATES FROM BACKYARD CHICKENS AND GUINEA FOWL IN THE GAMBIA

## 4.1 Introduction

The domestic chicken (*Gallus gallus domesticus)* is the most numerous bird on the planet, with an estimated population of over 22.7 billion—ten times more than any other bird [562]. Since their domestication from the red jungle fowl in Asia between 6,000 and 8,000 years ago [563, 564], chickens have been found almost everywhere humans live. Other poultry, such as turkeys, guinea fowl, pheasants, duck and geese, are derived from subsequent domestication events across Africa, Europe and the Americas [565]. For example, the helmeted guinea fowl (*Numida meleagris)* originated in West Africa, although domesticated forms of this bird are now found in many parts of the tropics.

Poultry are reared for meat, eggs and feathers [566]. Poultry production is classified into four sectors, based on the marketing of poultry products and the level of biosecurity [567]. Intensive poultry farming falls under sectors 1 to 3, characterised by moderate to high levels of biosecurity, while sector 4 pertains to the "backyard", "village" or "family" poultry system, with little or no biosecurity measures.

In rural backyard farming—prevalent in low- to middle-income countries such as the Gambia—a small flock of birds (between one and fifty) usually from indigenous breeds are allowed to scavenge for feed over a wide area during the daytime, with minimal supplementation, occasional provision of water and natural hatching of chicks. The poultry may be confined at night in rudimentary shelters to minimise predation, or birds may roost in owners' kitchens, family dwellings, tree-tops, or nest in the bush [568]. Urban and peri-urban backyard poultry farming—for example, in Australia, New Zealand (North Island), the US and in the UK—differs from rural backyard farming in that the birds are kept on an enclosed residential lot [569-571].

Backyard poultry fulfil important social, economic and cultural roles in many societies. Seventy percent of poultry production in low-income countries comes from backyard poultry [569]. The sale of birds and eggs generates income, while occasional consumption of poultry meat provides a source of protein in the diet. In traditional societies, domestic poultry meat is considered tastier than commercial broiler meat and, as it is perceived to be tougher in texture, is preferred for preparing dishes that require prolonged cooking [572]. It is estimated that meat and eggs from backyard poultry contribute about 30% of the total animal protein supply of households in low-income countries [573, 574]. In rural Gambia, backyard poultry can be offered as gifts for newlyweds or sold to solve family needs such as paying school fees, buying new clothes or other household needs [568]. Chickens may also be used as offering to a traditional healer, consumed when there is a guest, or during ceremonies. Urban and peri-urban poultry are kept mostly for home consumption of their eggs or meat, but also as pets or used for pest control [569, 575-577].

As discussed in the main introduction chapter, a sub-pathotype of ExPEC strains, known as Avian Pathogenic *E. coli* (APEC), causes colibacillosis—an extraintestinal disease in birds, with manifestations such as septicaemia, air sacculitis and cellulitis [578]. Avian colibacillosis results in high mortality and condemnation of birds, resulting in significant annual economic losses for the poultry industry [579]. As a result, antimicrobials are often used in intensive farming systems to prevent bacterial infections and treat sick birds—a practice that has been linked to the development of antimicrobial resistance (AMR) in poultry.

Although previous studies have focused on the detection of AMR and documented the emergence of multiple-drug resistance (MDR) in this niche [580-584], little is known about the population structure of *E. coli* in rural backyard poultry. The Gambia does not have genomic data on *E. coli* from poultry prior to this study and data on the circulating MLST types among poultry *E. coli* strains from the sub-Saharan Africa is limited. However, reports from Ghana, Senegal and Nigeria have indicated the prevalence of ST624, ST69, ST540, ST7473, ST155, ST297, ST226, ST10, ST3625 and ST58 among *E. coli* isolates from commercial poultry [585-587]. Given the increased exposure to humans, the natural environment and other animals, the population of *E. coli* in birds raised under the backyard system may differ considerably from those reared in intensive systems. It is also possible that the lineages of *E. coli* within local genotypes of rural poultry might differ between geographical regions. The absence of biosecurity measures in backyard poultry farming increases the potential for zoonotic transmission of pathogenic and/or antimicrobial-resistant strains to humans.

In a recent study of commercial broiler chickens, multiple colony sampling revealed that a single broiler chicken could harbour up to nine sequence types of *E. coli* [588]. However, within-host diversity of *E. coli* in backyard poultry—particularly in guinea fowl—has not been well studied and so we do not know how many lineages of *E. coli* can co-colonise a single backyard bird. To address these gaps in our knowledge, I exploited whole-genome sequencing to investigate the genomic diversity and burden of AMR among *E. coli* isolates from backyard chickens and guinea fowl in rural Gambia, West Africa.

## 4.2 Materials and methods

### 4.2.1 Study population

The study population comprised ten local-breed chickens and nine guinea fowl from a village in Sibanor in the Western Region of the Gambia (Table 4.1). Sibanor covers an area of approximately 90 km$^2$ and is representative of rural areas in the Gambia [589]. It has a population of about 10,000. Most of the villagers are subsistence farmers growing peanuts, maize and millet. Households within this community comprise extended family units of up to fifteen people, which make up the "compound". All guinea fowl were of the pearl variety, characterised by purplish-grey feathers dotted with white.

### 4.2.2 Sample collection

The sampling was done in November 2016. Poultry birds were first observed in motion for the presence of any abnormalities. Healthy-looking birds were procured from eight contiguous households within 0.3-0.4 km of each other and transported to the Abuko Veterinary Station, the Gambia in an air-conditioned vehicle. A qualified veterinarian then euthanised the birds and removed their caeca under aseptic conditions. These were placed into sterile falcon tubes and flash-frozen on dry ice in a cooler box. The samples were transported to the Medical Research Council Unit the Gambia at the London School of Hygiene and Tropical Medicine labs in Fajara, where the caecal contents were aseptically emptied into new falcon tubes for storage at -80°C within 3 h. A peanut-sized aliquot was taken from each sample into a 1.8 ml Nunc tube containing 1 ml of Skimmed Milk Tryptone Glucose Glycerol (STGG) transport and storage medium (Oxoid, Basingstoke, UK), vortexed at 4200rpm for 2 min and frozen at -80°C. Figure 4.1 summarises the sample processing flow—the methods for which have been described in Chapter Two.

## 4.3 Results

### 4.3.1 Study population

I analysed nineteen caecal samples obtained from ten chickens and nine guinea fowl. Fifteen out of the nineteen (79%) samples yielded growth of *E. coli* on culture, from which 68 colonies were recovered (five colonies from each of thirteen birds, two from a single bird and one colony from another bird).

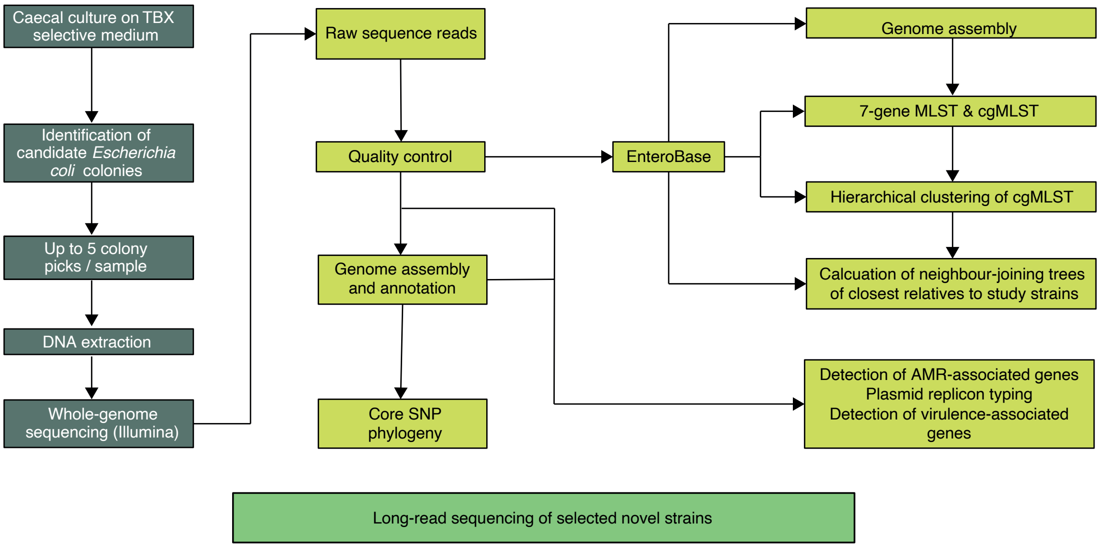

Caecal culture on TBX selective medium
Identification of candidate *Escherichia coli* colonies
Up to 5 colony picks / sample
DNA extraction
Whole-genome sequencing (Illumina)
Raw sequence reads
Quality control
Genome assembly and annotation
Core SNP phylogeny
EnteroBase
Genome assembly
7-gene MLST & cgMLST
Hierarchical clustering of cgMLST
Calcuation of neighbour-joining trees of closest relatives to study strains
Detection of AMR-associated genes
Plasmid replicon typing
Detection of virulence-associated genes
Long-read sequencing of selected novel strains

***Figure 4.1:*** *Study sample-processing flow diagram.*

### 4.3.2 Sequence type and phylogroup distribution

I recovered 28 seven-allele sequence types (STs), of which ST155 was the most common (22/68, 32%). Four of the STs were novel—two from chickens and two from guinea fowl. The allelic profiles of the novel strains are provided in [590], File S2. Seventeen of the 28 STs have been previously isolated from humans and or other vertebrates, five (ST942, ST2165, ST2461, ST4392 and ST5826) have not been seen in humans before and one (ST6025) occurred in only one other isolate in EnteroBase, beside the study strain. However, the source of isolation of this other isolate was not available (Table 4.2). The isolates were spread over phylogroups B1, A, E and D, but most belonged to phylogroups B1 and A, which are home to strains associated with human intestinal infections and avian colibacillosis [591, 592] (Figure 4.2). Hierarchical clustering resolved the study strains into 22 cgMLST complexes, indicating a high level of genomic diversity (File S2 in [590]).

I generated complete, circular genome assemblies of the two novel sequence types isolated from guinea fowl: ST10654 (GF3-3) and ST9286 (GF4-3). Although neither strain encoded AMR genes or plasmids, GF3-3 contained three prophages (two intact, one incomplete), while GF4-3 harboured four prophages (three intact, one incomplete) (Table 4.3).

### 4.3.3 Within-host genomic diversity and transmission of strains

Several birds (12/19, 63%) were colonised by two or more STs; in most cases, the STs spanned more than two phylotypes (Table 4.1). In two chickens, all five colony picks belonged to distinct

STs. I observed some genetic diversity among multiple colonies of the same ST recovered from the same host (Table 4.4). Most of these involved variants that differed by 0-4 SNPs, i.e., variation likely to have arisen due to within-host evolution. However, in one instance, pair-wise SNP differences (ranging from 4 to 255) suggested independent acquisition of distinct clones. Pair-wise SNP analysis also suggested transmission of strains between chickens and between chicken and guinea fowl (Table 4.5 and 4.6) from the same household (File S4 in [590]).

***Table 4.1:*** *Characteristics of the study population.*

| Sample ID | Poultry species | Gender | Household | Colony picks | Recovered Sequence Types (No. of colonies per ST) | Phylogroup distribution (ST(s) per phylogroup) |
|---|---|---|---|---|---|---|
| C1 | Chicken | Rooster | 1 | No growth | | |
| C2 | Chicken | Hen | 3 | 1 | 155 (1) | B1 (155) |
| C3 | Chicken | Rooster | 2 | 5 | 155 (1), 48 (1), 746 (1) 2461 (1), 542 (1) | A (48, 746, 2461, 542), B1 (155) |
| C4 | Chicken | Rooster | 2 | 5 | 1423 (1), 337 (1), 9285* (1), 540 (1), 58 (1) | A (540), B1 (1423, 337, 9285*, 58) |
| C5 | Chicken | Hen | 2 | 2 | 155 (2) | B1 (155) |
| C6 | Chicken | Rooster | 2 | 5 | 155 (3), 9284* (2) | B1 (155), E (9284*) |
| C7 | Chicken | Rooster | 3 | 5 | 155 (4), 602 (1) | B1 (155, 602) |
| C8 | Chicken | Rooster | 4 | 5 | 5286 (1), 2772 (2), 6186 (1), 2165 (1) | A (5286), B1 (2772, 6186, 2165) |
| C9 | Chicken | Hen | 5 | No growth | | |
| C10 | Chicken | Rooster | 5 | No growth | | |
| GF1 | Guinea fowl | Rooster | 1 | 5 | 540 (5) | A (540) |
| GF2 | Guinea fowl | Rooster | 1 | 5 | 155 (4), 540 (1) | A (540), B1 (155) |
| GF3 | Guinea fowl | Rooster | 3 | 5 | 540 (2), 443 (1), 6025 (1), 10654* (1) | A (540), B1 (443), D (6025), E (10654) |
| GF4 | Guinea fowl | Rooster | 6 | 5 | 155 (4), 9286* (1) | B1 (155, 9286) |
| GF5 | Guinea fowl | Hen | 6 | 5 | 155 (2), 4392 (1), 86 (1), 942 (1) | B1 (155, 4392, 86, 942) |
| GF6 | Guinea fowl | Hen | 1 | 5 | 540 (1), 2067 (4) | A (540), B1 (2067) |
| GF7 | Guinea fowl | Rooster | 2 | 5 | 212 (4), 155 (1) | B1 (155, 212) |
| GF8 | Guinea fowl | Rooster | 7 | No growth | | |
| GF9 | Guinea fowl | Rooster | 8 | 5 | 2614 (2), 295 (1) 196 (1), 2067 (1) | B1 (2614, 295, 196) |
| Total | | | | 68 | | |

Novel sequence types are designated by an asterisk (*); No growth indicates where *E. coli* was not isolated.

### 4.3.4 Prevalence of AMR, virulence factors and plasmid replicons among the study isolates

Twenty isolates (20/68, 29%) harboured at least one AMR gene and sixteen (16/68, 24%) were positive for genes predicted to convey resistance to three or more classes of antibiotics (Figure 4.3; File S5 in [590]). Fourteen of these sixteen isolates belonged to ST155—representing 64% (14/22) of the ST155 isolates recovered in this study. Notably, none of the clinically important beta-lactamase resistance genes commonly found among multi-drug resistant clones were detected among my study isolates—with only the class A broad-spectrum beta-lactamase resistance genes (bla$_{TEM-1A}$/bla$_{TEM-1B}$) observed among 26% (18/68) of the study isolates. Phenotypic resistance was confirmed in >50% of the isolates tested.

Interestingly, isolates encoding resistance to three or more classes of antibiotics also harboured more genes encoding putative virulence factors than did less-resistant isolates (Figure 4.2). Overall, 125 unique virulence-associated genes were detected from the study isolates (File S6 in [590]). Notably, the virulence and AMR profiles of co-colonising STs tended to differ from each other.

One or more plasmid replicons were detected in 69% (47/68) of the study isolates, with seventeen plasmid types detected overall (File S7 in [590]). IncF plasmids were the most common. A single isolate carried the col156 virulence plasmid. The multi-drug resistant isolates often co-carried large IncF plasmids (IncFIA_1, ~27kb; IncFIB(AP001918)_1, ~60kb; IncFIC(FII)_1, ~56kb). Scrutiny of annotated assemblies revealed that resistance genes were often co-located on the same contig as one of the IncF plasmids. In three birds (Guinea fowl 2, Guinea fowl 5 and Guinea fowl 7), co-colonising strains (belonging to different STs) shared the same plasmid profile. The results of ARIBA ResFinder, PlasmidFinder and VFDB were 100% concordant with those produced by ABRicate for my study isolates.

### 4.3.5 Population dynamics of study strains

Hierarchical clustering analyses highlighted some genomic relationships between strains from poultry and those from humans (Table 4.7); however, this warrants further investigation using samples collected from poultry and humans living in close proximity from the same setting. Significant among these were ST2772 and ST4392, which were separated from human isolates belonging to these STs by just 41 and 68 alleles in the core-genome MLST scheme respectively

(Figures 4.4 and 4.5). Similarly, ST86, ST6186 and ST602 were closest to isolates from livestock (Figures 4.6-4.8), suggesting possible interactions of strains between livestock species.

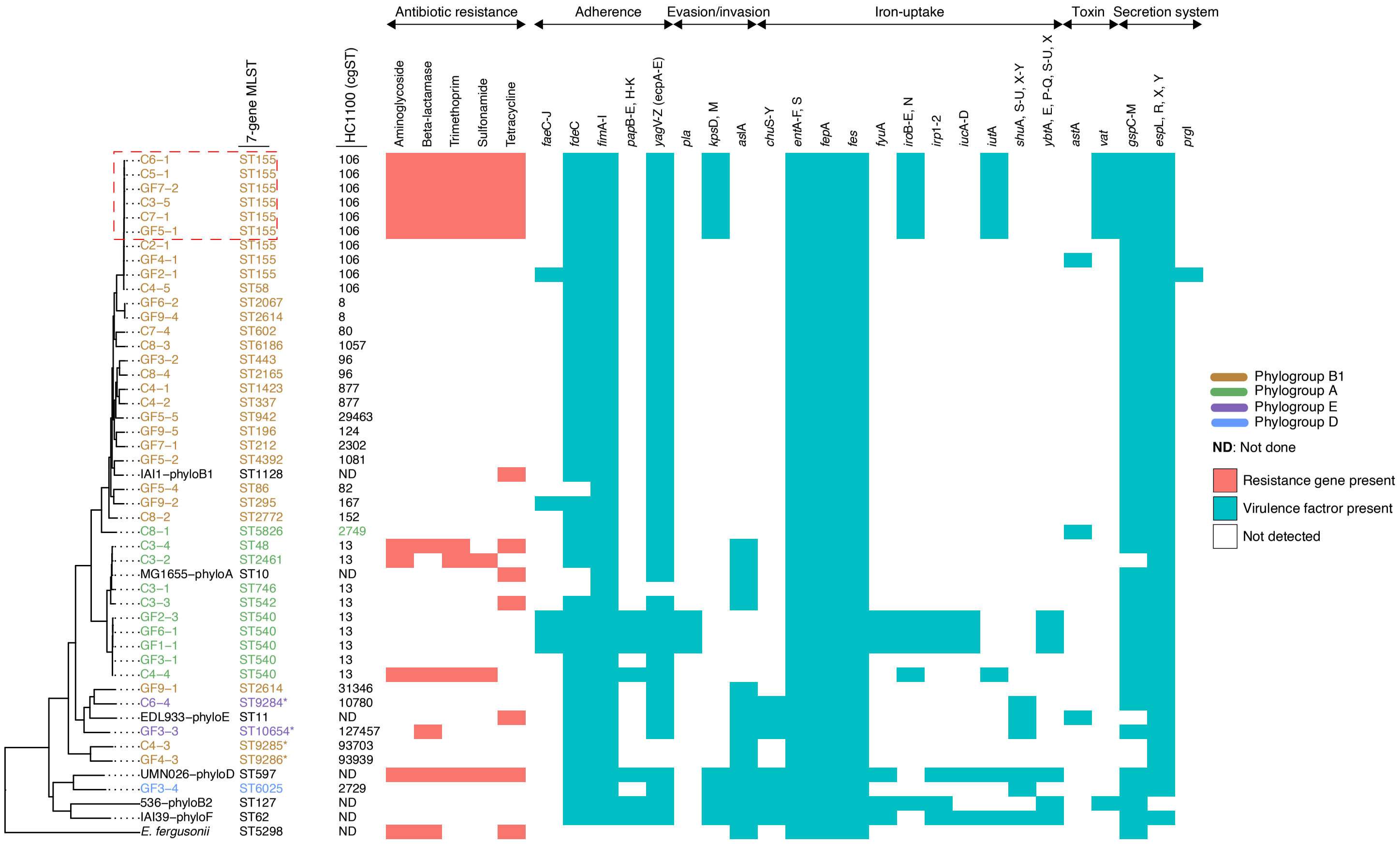
By contrast, three of the novel STs from this study (ST10654, ST9285, ST9286) were genetically distinct from anything else in the public domain. These belonged to unique HC1100 clusters in the cgMLST scheme and did not have any relatives in the seven-allele MLST scheme, even after allowing for two mismatches. Two of these (ST10654 from Guinea fowl 3 and ST9286 from Guinea fowl 4) now have complete genomic assemblies (Table 4.3).

***Table 4.2:*** *Prevalence of the study sequence types in EnteroBase.*

| ST | Source | Phylotype | Prevalence in EnteroBase |
|---|---|---|---|
| 48 | Chicken | A | Human, livestock, Celebes ape |
| 58 | Chicken | B1 | Human, livestock, poultry |
| 86 | Guinea fowl | B1 | Human, livestock, companion animal, poultry |
| 155 | Chicken, Guinea fowl | B1 | Human, poultry, mink, livestock |
| 196 | Guinea fowl | B1 | Human, livestock, companion animal, environment |
| 212 | Guinea fowl | B1 | Human, poultry, deer, companion animal |
| 295 | Guinea fowl | B1 | Human, poultry, livestock, companion animal, environment, food, |
| 337 | Chicken | B1 | Human, rhinoceros, poultry, environment (soil and water) |
| 443 | Guinea fowl | B1 | Human, environment, livestock |
| 540 | Chicken, Guinea fowl | A | Human, environment (water and sewage), livestock, poultry, gull, rabbit, plant, oyster, fish |
| 542 | Chicken | A | Human, livestock, poultry |
| 602 | Chicken | B1 | Human, poultry, livestock, bird, fish, reptile |
| 746 | Chicken | A | Human, poultry, fish, livestock, environment (water) |
| 942 | Guinea fowl | B1 | Environment, food, companion animal, livestock |
| 1423 | Chicken | B1 | Human, reptile, livestock |
| 2067 | Guinea fowl | B1 | Human, environment |
| 2165 | Chicken | B1 | Livestock, companion animal, reptile, bird |
| 2461 | Chicken | A | Sheep, poultry |
| 2614 | Guinea fowl | B1 | Human |
| 2772 | Chicken | B1 | Human, livestock, environment |
| 4392 | Guinea fowl | B1 | Livestock, wild animal, companion animal |
| 5826 | Chicken | A | Poultry |
| 6025 | Guinea fowl | D | Unknown source§ |
| 6186 | Chicken | B1 | Livestock, environment |
| 9284 | Chicken | E | Novel |
| 9285 | Chicken | B1 | Novel |
| 9286 | Guinea fowl | B1 | Novel |
| 10654 | Guinea fowl | D | Novel |

§ ST6025 occurred in only one other isolate in EnteroBase (source of isolation unknown), beside the study strain.

**Figure 4.2:** *A maximum-likelihood phylogeny depicting the genetic relationships among my study isolates.*



*A maximum-likelihood phylogeny of my study isolates reconstructed with RAxML, based on non-repetitive, non-recombinant core SNPs, using a general time-reversible nucleotide substitution model with 1000 bootstrap replicates. The tip labels indicate the sample names, with the respective Achtman sequence types (STs) and HC1100 (cgST complexes) indicated next to the sample names. The colour codes indicate the respective phylogroups to which the isolates belong. The outgroup and the other E. coli reference genomes denoting the major E. coli phylogroups are in black. Asterisks (*) are used to indicate novel STs. Overlaid on the tree are the predicted antimicrobial resistance genes and virulence factors for each isolate. The virulence genes are grouped according to their function. Chicken isolates are denoted 'C' and guinea fowl samples 'GF', with the suffix indicating the colony pick. I have not shown multiple colonies of the same Achtman ST recovered from a single bird – in such instances, only one representative isolate is shown. Nor have I shown virulence factors that were detected only in the reference genomes. The red box highlights multi-drug-resistant isolates that concurrently harbour putative fitness and colonization factors that are important for invasion of host tissues and evasion of host immune defences. The full names of virulence factors and their known functions are provided in File S6 in Foster-Nyarko et al. (2020) (Reference 585).*

***Table 4.3:*** *A summary of the sequencing statistics of two novel sequence types derived from guinea fowl B: Prophage types detected from long-read sequences using PHASTER (Reference 523).*

**A**

| Assembly | ST | Total length (bp) | Circularised | GC (%) | N50 | CDSs | tRNAs | rRNAs | Repeat region | tmRNA |
|---|---|---|---|---|---|---|---|---|---|---|
| GF3-3 | 9286 | 4706754 | Yes | 50.63 | 4706754 | 5284 | 86 | 22 | 1 | 1 |
| GF4-3 | 10654 | 4821968 | Yes | 50.51 | 4821968 | 4324 | 85 | 22 | 1 | 1 |

**B**

| Sample ID | Region | Region Length | Completeness | #Total Proteins | Region Position | Predicted Phage | GC % |
|---|---|---|---|---|---|---|---|
| GF4-3 | 1 | 27Kb | Incomplete | 15 | 2194854-2221904 | PHAGE_Entero_YYZ_2008_NC_011356(2) | 47.71 |
|  | 2 | 31.4Kb | Intact | 31 | 2225927-2257412 | PHAGE_Entero_mEp460_NC_019716(22) | 51.39 |
|  | 3 | 38.9Kb | Intact | 51 | 3642598-3681512 | PHAGE_Entero_sfV_NC_003444(37) | 48.55 |
|  | 4 | 30.2Kb | Intact | 53 | 4059624-4089830 | PHAGE_Mycoba_32HC_NC_023602(1) | 46.22 |
| GF3-3 | 1 | 26.8Kb | Intact | 32 | 139-27030 | PHAGE_Escher_pro483_NC_028943(25) | 52.68 |
|  | 2 | 35.1Kb | Intact | 48 | 1303385-1338558 | PHAGE_Salmon_Fels_2_NC_010463(37) | 50.23 |
|  | 3 | 5.9Kb | Incomplete | 9 | 3220895-322618 | PHAGE_Bacill_G_NC_023719(2) | 47.27 |

***Table 4.4:*** *Within-host single nucleotide polymorphism diversity between multiple genomes of the same ST recovered from the same bird.*

| Sample ID | Sequence type (ST) | Colonies per ST | Pair-wise SNP distances between multiple colonies of the same ST |
|---|---|---|---|
| C5 | 155 | 2 | 0 |
| C6 | 155 | 3 | 0 |
| C6 | 9284 | 2 | 4 |
| C7 | 155 | 4 | 0 |
| C8 | 2772 | 2 | 4 |
| GF1 | 540 | 5 | 0-3 |
| GF2 | 155 | 4 | 0 |
| GF3 | 540 | 2 | 2 |
| GF4 | 155 | 4 | 0-4 |
| GF5 | 155 | 2 | 0 |
| GF6 | 2067 | 4 | 0 |
| GF7 | 212 | 4 | 4-255 |
| GF9 | 2614 | 2 | 0 |

"C" denotes chickens and "GF" denotes guinea fowl.

***Table 4.5:*** *Single nucleotide polymorphism differences between isolates recovered from Chicken 3, Chicken 5, Chicken 6 and Guinea fowl 7.*

| | C3-5 | C5-1 | C5-2 | GF7-2 | C6-1 | C6-2 | C6-3 |
|---|---|---|---|---|---|---|---|
| C3-5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C5-1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C5-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| GF7-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C6-1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C6-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C6-3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Chicken samples are denoted by “C” and guinea fowl samples by “GF”. All the isolates in this transmission network encoded resistance to ≥ 3 classes of antimicrobials.

***Table 4.6:*** *Single nucleotide diversity differences between isolates recovered from guinea fowls 1, 2 and 6.*

| | GF1-1 | GF1-2 | GF1-3 | GF1-4 | GF1-5 | GF2-3 | GF6-1 |
|---|---|---|---|---|---|---|---|
| GF1-1 | 0 | 2 | 3 | 1 | 1 | 2 | 3 |
| GF1-2 | 2 | 0 | 3 | 1 | 1 | 2 | 3 |
| GF1-3 | 3 | 3 | 0 | 2 | 2 | 3 | 2 |
| GF1-4 | 1 | 1 | 2 | 0 | 0 | 1 | 2 |
| GF1-5 | 1 | 1 | 2 | 0 | 0 | 1 | 2 |
| GF2-3 | 2 | 2 | 3 | 1 | 1 | 0 | 3 |
| GF6-1 | 3 | 3 | 2 | 2 | 2 | 3 | 0 |

### 4.3.6 The global prevalence of strains and AMR among avian *E. coli* isolates

Phylogenomic analyses of 4,846 poultry *E. coli* isolates from all over the world revealed that ST155 is common among poultry isolates from Africa and South America (Figures 4.9 and 4.10). In contrast, ST117 is prevalent among poultry isolates from Europe and North America (Figures 4.11 and 4.12), with ST156 and ST254 being the most common *E. coli* STs found in poultry from Asia and Oceania respectively (Figures 4.13 and 4.14).

The phylogenetic analyses revealed that ST155 strains from Africa were dispersed among other ST155 isolates from the rest of the world; however, the majority of ST155 strains from this study belonged to a tight genomic cluster, comprised of isolates from poultry and livestock from sub-Saharan Africa (separated by 38-39 alleles), except for a single isolate sourced from poultry in the US. In the cgMLST scheme, all my study ST155 isolates fell into four HC100 sub-clusters (100 alleles difference) (Figure 4.15). The largest sub-cluster (sub-cluster 1, HC100_43137) comprised ST155 isolates from this study and isolates from Uganda and Kenya; while sub-clusters 2

(HC100_73903), 3 (HC100_73905) and 4 (HC100_93719) occurred in the Gambia only, although distantly related to isolates from humans and a companion animal (Figures 4.16-4.19).

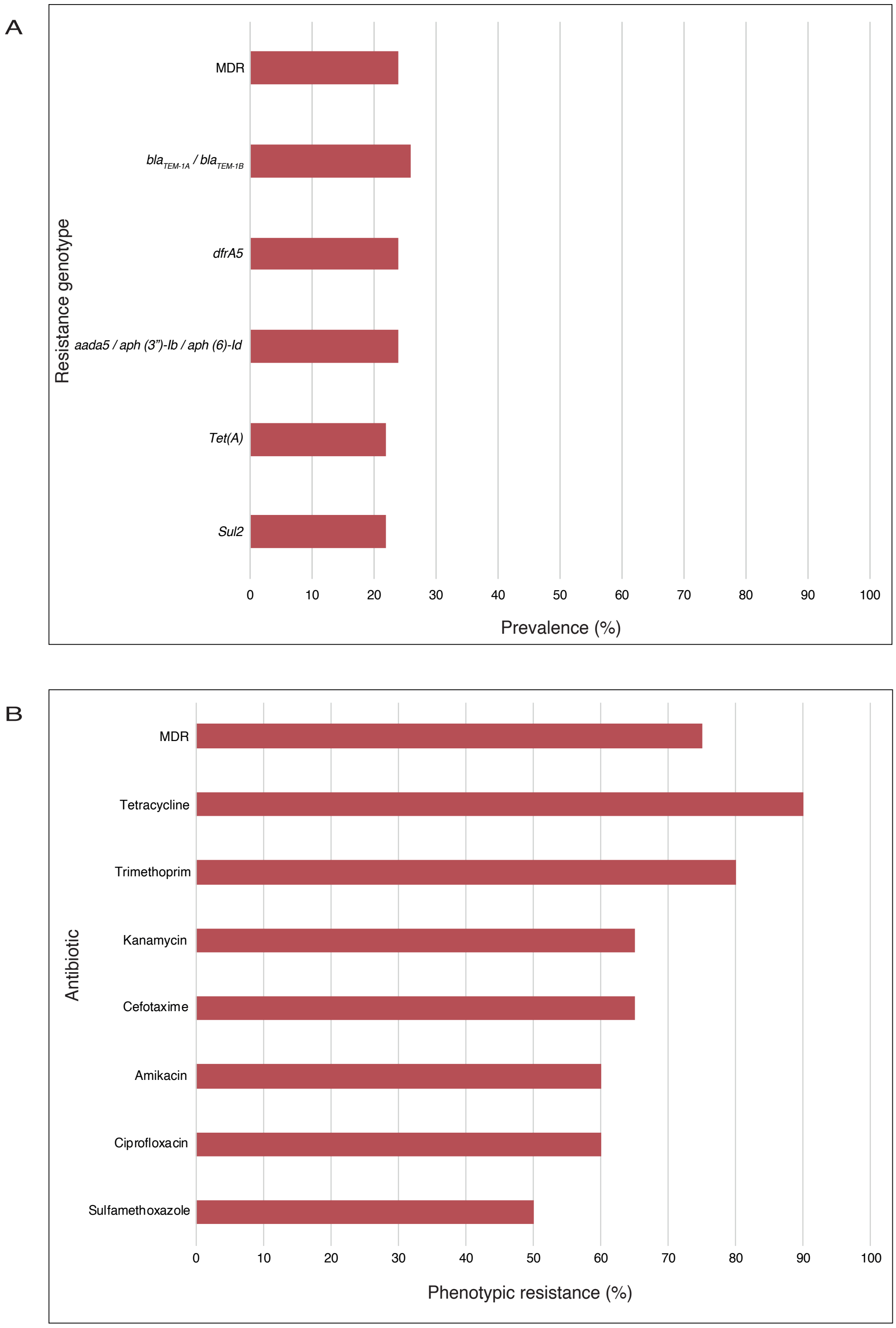


***Figure 4.3: A:** A bar graph showing the prevalence of resistance genes found among the study isolates.*
This was using the core Virulence Factors Database (virulence factors), ResFinder (AMR) and PlasmidFinder (plasmid-associated genes) databases, with a cut-off percentage identity of ≥ 90% and coverage of ≥ 70%. The full list of the resistance genes that were detected is presented in File S5 in [590]. **B:** A bar graph depicting the prevalence of phenotypic antimicrobial resistance in 20 isolates. The results were interpreted using the recommended breakpoint tables from EUCAST (http://www.eucast.org) or the Clinical Laboratory Standards Institute (https://www.clsi.org) (Performance Standards for Antimicrobial Susceptibility Testing (28th Information Supplement, M100-S28) where EUCAST cut-off values were not available.

***Table 4.7:*** *Closest relatives to the Gambian poultry strains.*

| 7-gene ST | cgST HC100 sub-cluster | Study poultry host | Neighbour host | Neighbour's country of isolation | Allelic distance |
|---|---|---|---|---|---|
| ST9286 | NA | Guinea fowl | Chicken | Gambia (this study) | 945 |
| ST9285 | NA | Chicken | Guinea fowl | Gambia (this study) | 945 |
| ST10654 | NA | Guinea fowl | Unknown avian source | Kenya | 1324 |
| ST155 | 43137 | Chicken and Guinea fowl | Poultry | US | 32-34 |
| ST2772 | NA | Chicken | Human | Kenya | 41 |
| ST6186 | NA | Chicken | Livestock | US | 58 |
| ST540 | 10207 | Guinea fowl | Human | UK | 59 |
| ST58 | 25133 | Chicken | Unknown | Unknown | 59 |
| ST2461 | 93699 | Chicken | Human | Kenya | 64 |
| ST2165 | 12281 | Chicken | Food | Kenya | 66 |
| ST4392 | NA | Guinea fowl | Human | UK | 68 |
| ST602 | NA | Chicken | Livestock | US | 70 |
| ST540 | 70056 | Chicken | Food | UK | 72 |
| ST540 | 1320 | Guinea fowl | Poultry | US | 73 |
| ST942 | NA | Guinea fowl | Environment (tap water) | Australia | 76 |
| ST212 | NA | Guinea fowl | Seagull | Australia | 81 |
| ST5826 | NA | Chicken | Water | UK | 91 |
| ST1423 | 27957 | Chicken | Reptile | US | 96 |
| ST337 | 73054 | Chicken | Reptile | US | 96 |
| ST196 | NA | Guinea fowl | Human | Kenya | 102 |
| ST155 | 93719 | Chicken | Tanzania | Human | 106 |
| ST86 | NA | Guinea fowl | US | Livestock | 131 |
| ST155 | 73905 | Guinea fowl | Companion animal | US | 137 |
| ST542 | 93732 | Chicken | Poultry | US | 148 |
| ST746 | NA | Chicken | Poultry | US | 148 |
| ST295 | NA | Guinea fowl | Human | Mexico | 162 |
| ST48 | 93724 | Chicken | Unknown | UK | 163 |
| ST542 | 93697 | Chicken | Environment (soil/dust) | US | 194 |
| ST155 | 73903 | Guinea fowl | Nepal | Human | 195 |
| ST443 | 93721 | Guinea fowl | Unknown | Unknown | 224 |
| ST6025 | NA | Guinea fowl | Unknown | US | 245 |
| ST2614 | NA | Guinea fowl | Human | China | 284 |
| ST9284 | NA | Chicken | Environment (soil/dust) | North America | 293 |
| ST2067 | NA | Guinea fowl | Human | Gambia | 458 |

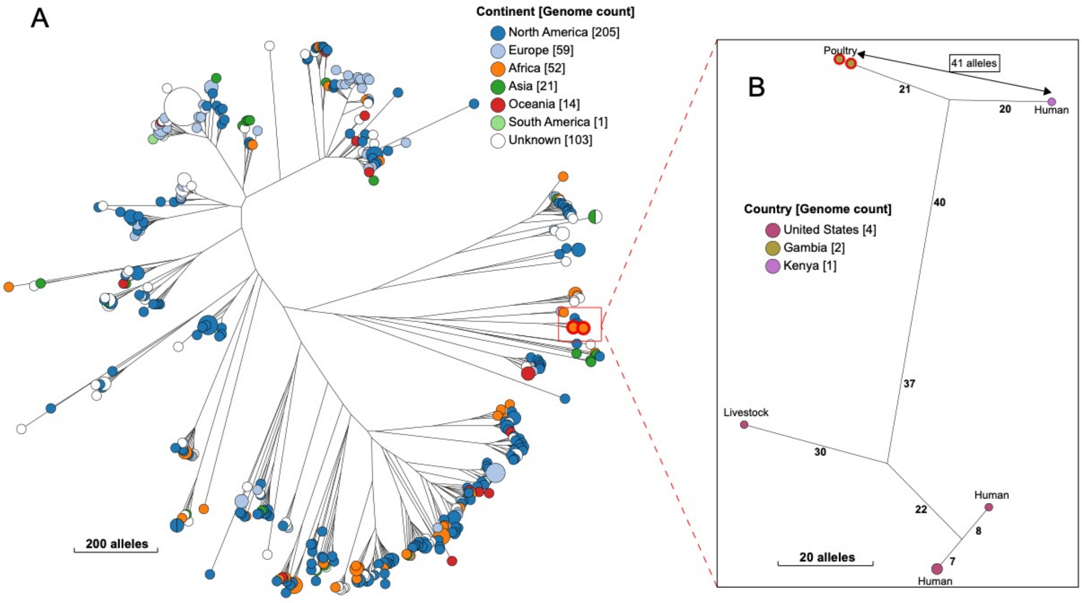


***Figure 4.4:*** *Population structure of ST2772.*
A NINJA neighbour-joining tree showing the phylogenetic relationship between my study ST2772 (Achtman) strain and all other publicly available genomes that fell within the same HC1100 cluster (cgST complex). The locations of the isolates are displayed, with the genome counts displayed in square brackets. The branch lengths are annotated with the allelic distances separating the genomes. Strains from this study are highlighted in red. The sub-tree (**B**) shows the closest relatives to the study strains, with the allelic distance separating them displayed with the arrow (41 alleles).

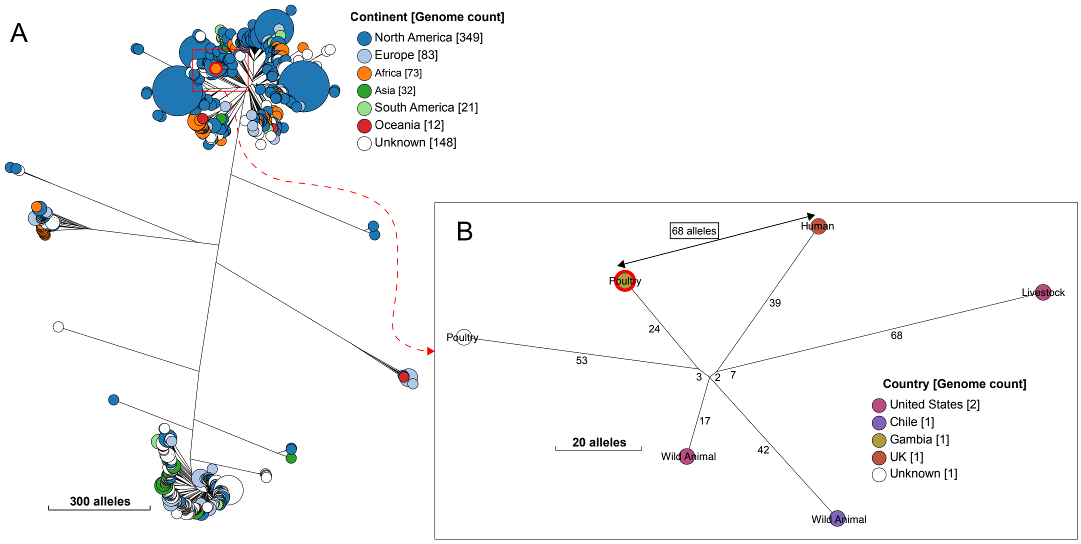


***Figure 4.5:*** *Population structure of ST4392.*
A NINJA neighbour-joining tree showing the phylogenetic relationship between the avian ST4392 (Achtman) strain from this study and all other publicly available genomes that cluster together at HC1100 level (cgST complex). The legend shows the continent of isolation of the isolates, with genome counts displayed in square brackets. Gambian poultry strains are highlighted in red. The study ST strain is separated from a human ST4392 isolate by 68 alleles, as shown in the subtree (**B**).

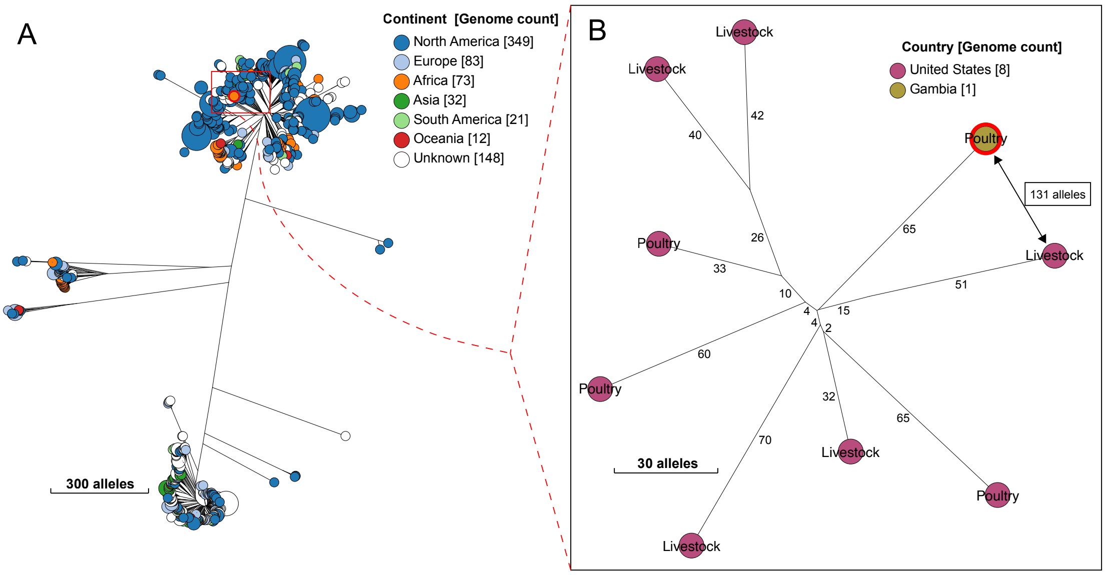


***Figure 4.6:*** *NINJA phylogenetic trees showing the closest neighbours to avian ST86 isolates from this study.* The nearest relatives occurred in livestock, depicted with the arrow in the subtree (**B**). The legend indicates the location of isolation, with the genome count displayed in square brackets.

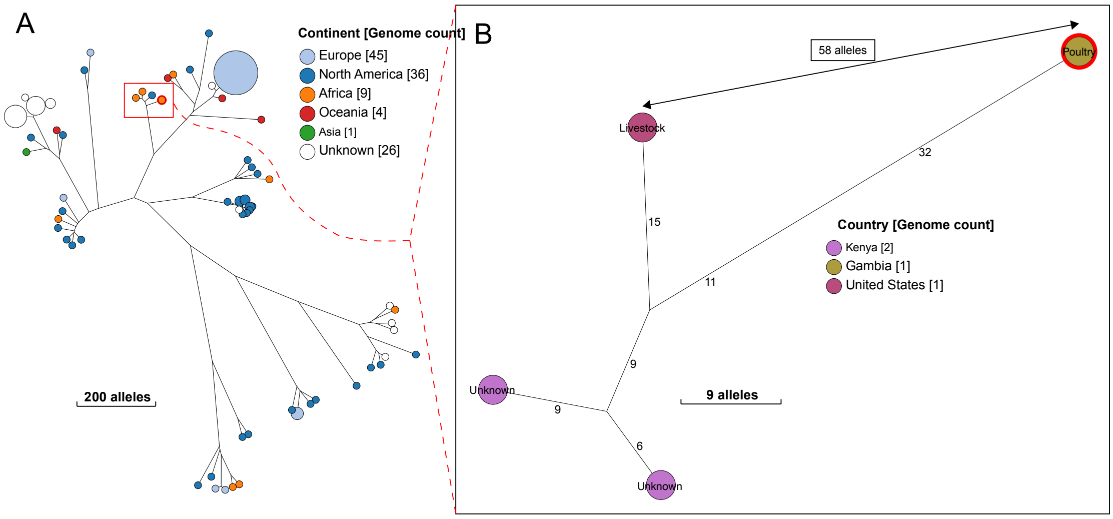


***Figure 4.7:*** *A NINJA phylogenetic tree showing the closest neighbours to avian ST6186 isolates from this study.* (**A**). The nearest relatives occurred in livestock from Kenya (**B**), separated by 58 alleles (depicted with the arrow). The branch lengths display the allelic A NINJA phylogenetic tree showing the closest neighbours to avian ST6186 isolates from this study (A). The nearest relatives occurred in livestock from Kenya (**B**), separated by 58 alleles (depicted with the arrow). The branch lengths display the allelic distance between the genomes. The legend indicates the location of isolation, with the genome count displayed in square brackets.

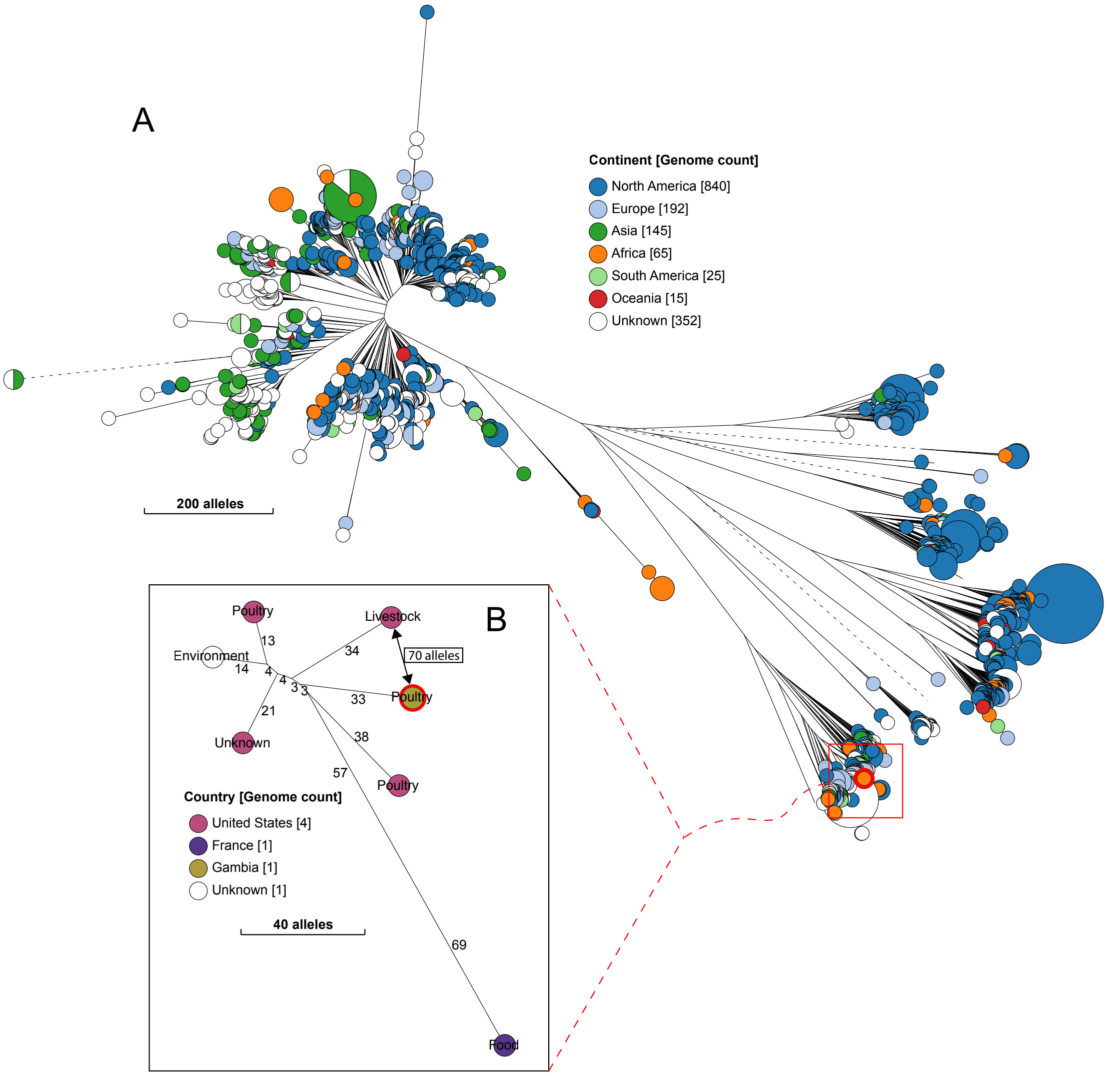


***Figure 4.8:*** *A NINJA phylogenetic tree showing the closest neighbours to avian ST602 isolates from this study.*
The nearest relatives were isolated from livestock from the US (B), separated by 70 alleles (depicted with the arrow). The branch lengths display the allelic distance between the genomes. The legend indicates the location of isolation, with the genome count displayed in square brackets.

Antimicrobial resistance was high across the continents, with the highest prevalence of MDR in South America (100/131, 77%), followed by Asia (175/249, 70%), then Africa (392/591, 66%) (Table 4.8; File S8 in [590]). Of note, the highest percentages of resistance globally were that of broad-spectrum beta-lactamases, while the lowest percentages of resistance were to colistin (Table 4.8). Interestingly, the prevalence of colistin resistance was highest in Europe but did not occur in Oceania and North America.

***Table 4.8:*** *Global prevalence of AMR genes.*

| | **Europe** | **Africa** | **South America** | **North America** | **Asia** | **Oceania** |
|---|---|---|---|---|---|---|
| Tetracycline | 564/752, 75% | 559/591, 95% | 108/131, 83% | 2480/2975, 83% | 228/249, 92% | 132/148, 90% |
| Aminoglycoside | 303/752, 40% | 378/591, 64% | 94/131, 72% | 1497/2975, 50% | 172/249, 69% | 56/148, 38% |
| Beta-lactamase | 303/752, 40% | 246/591, 42% | 127/131, 98% | 933/2975, 31% | 157/249, 63% | 61/148, 41% |
| Sulphonamide | 338/752, 45% | 377/591, 64% | 84/131, 65% | 1174/2975, 39% | 167/249, 67% | 52/148, 35% |
| Trimethoprim | 192/752, 25% | 353/591, 52% | 58/131, 45% | 176/2975, 6% | 143/249, 57% | 66/148, 45% |
| Chloramphenicol | 303/752, 40% | 69/591, 13% | 36/131, 28% | 69/2975, 2% | 131/249, 53% | 0/148, 0% |
| Quinolone | 51/752, 7% | 144/591, 24% | 24/131, 18% | 17/2975, 1% | 74/249, 30% | 0/148, 0% |
| Lincosamide | 57/752, 8% | 0/591, 0% | 12/131, 9% | 0/2975, 0% | 14/249, 6% | 1/148, 1% |
| Macrolide | 20/752, 3% | 79/591, 13% | 3/131, 2% | 30/2975, 1% | 92/249, 37% | 0/148, 0% |
| Fosfomycin | 8/752, 1% | 4/591, 1% | 31/131, 24% | 19/2975, 1% | 71/249, 29% | 0/148, 0% |
| Streptogrammin | 0/752, 0% | 0/591, 0% | 23/131, 18% | 0/2975, 0% | 0/249, 0% | 0/148, 0% |
| Colistin | 29/752, 4% | 0/591, 0% | 9/131, 7% | 0/2975, 0% | 119/249, 48% | 0/148, 0% |
| MDR | 406/752, 54% | 392/591, 66% | 100/131, 77% | 1236/2975, 42% | 175/249, 70% | 56/148, 44% |

The full list of resistance genes that were detected is presented in File S6 of [590].

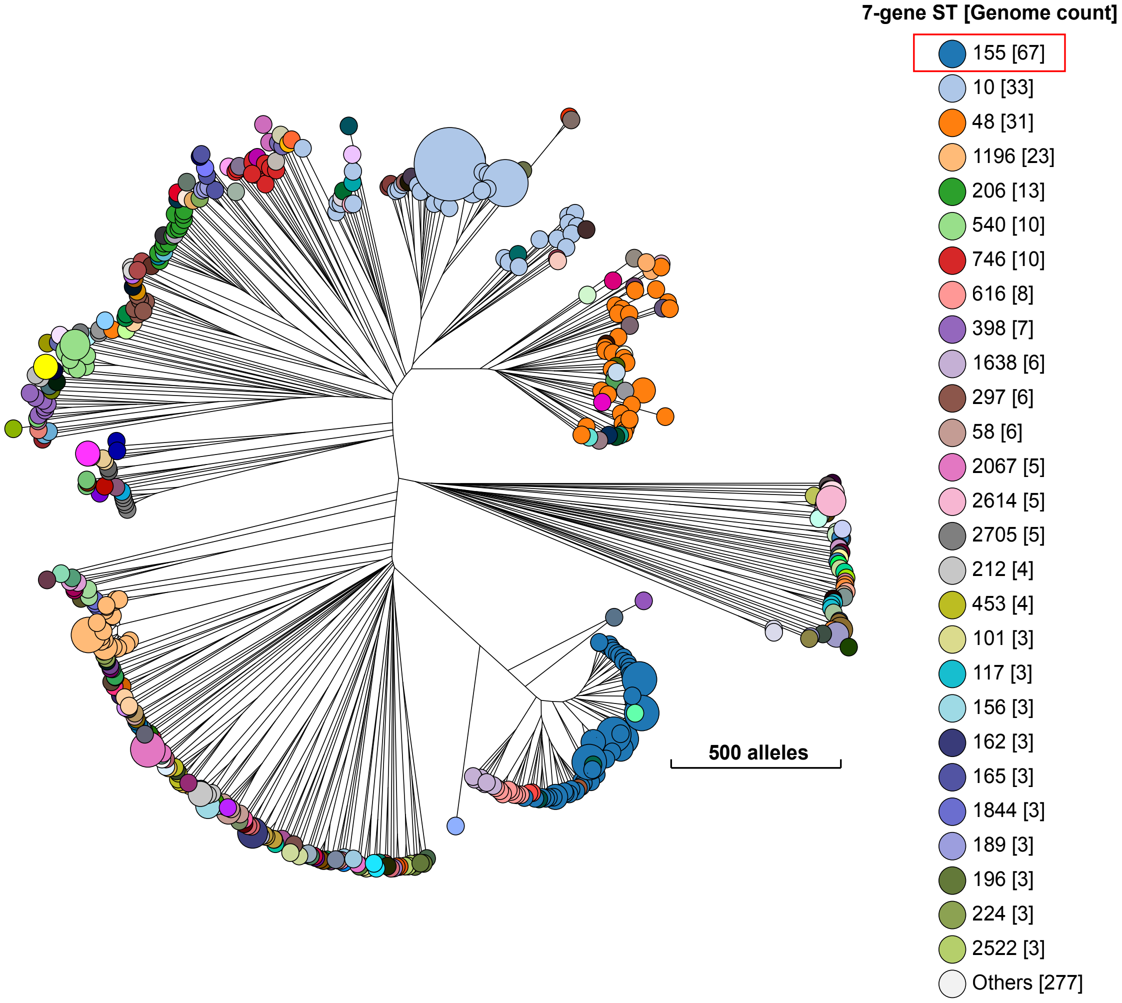


***Figure 4.9:*** *A NINJA neighbour-joining tree of all publicly available E. coli poultry isolates from Africa, showing the prevalence of Achtman sequence types (STs).*
The dominant ST is highlighted with a red box. The legend displays the top 27 STs, with the respective genome counts displayed in square brackets.

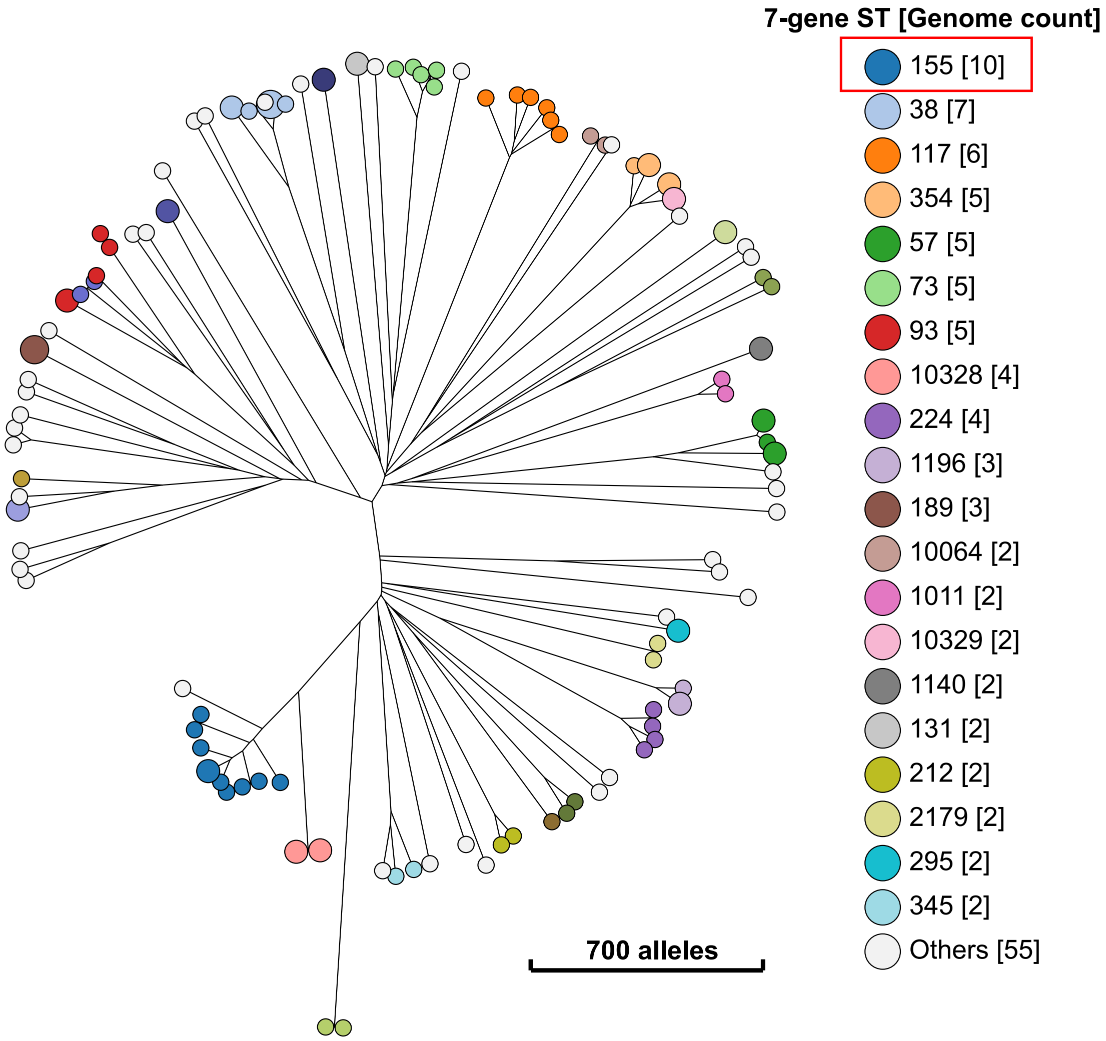


***Figure 4.10:*** *A NINJA neighbour-joining tree of all publicly available E. coli poultry isolates from South America, depicting the prevalence of Achtman sequence types (STs).*

The most common ST found among E. coli isolates from this continent is ST155 (highlighted with a red box), similar to Africa (see Figure 4.9). The top 20 STs are displayed in the legend, with the respective genome counts displayed [in square brackets].

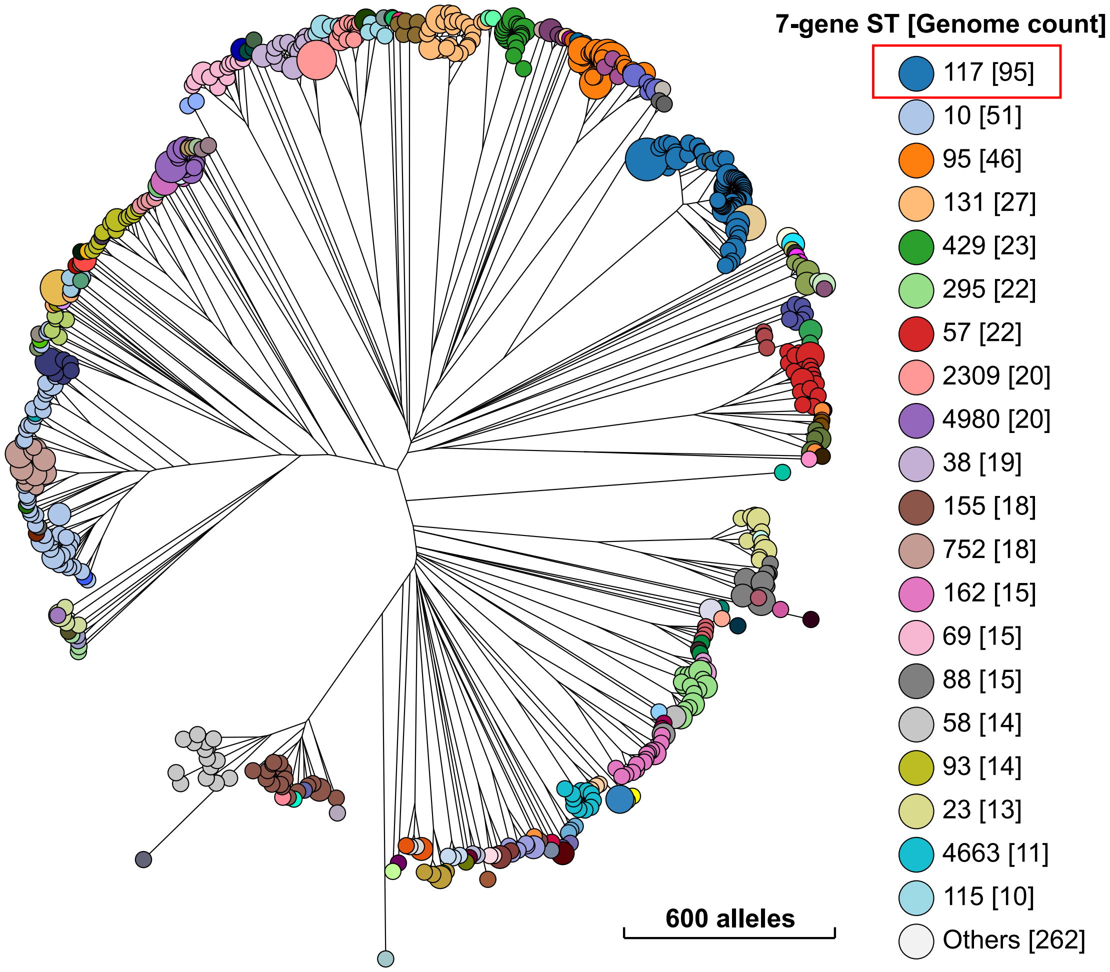


***Figure 4.11:*** *A NINJA neighbour-joining tree of all publicly available E. coli poultry isolates from Europe, depicting the prevalence of Achtman sequence types (STs).*
The top 20 STs are displayed in the legend, with the most common ST among poultry isolates from this continent (ST117) highlighted with a red box. The respective genome count per ST is also displayed.

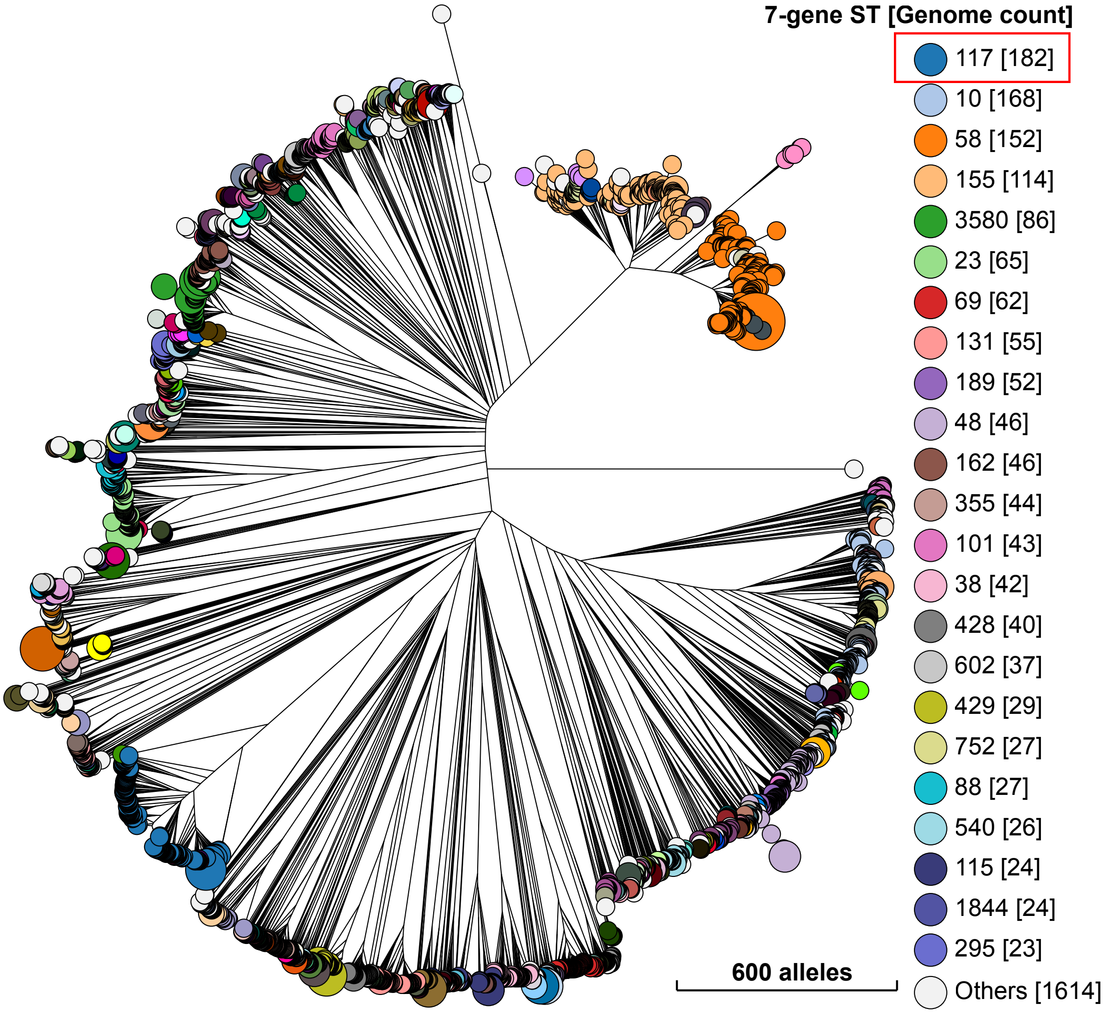


***Figure 4.12:*** *A NINJA neighbour-joining tree of all publicly available E. coli poultry isolates from North America, showing the prevalence of Achtman sequence types (STs).*

The most common ST among poultry isolates from this continent is ST117 (highlighted with a red box). The legend displays the top 23 STs, with the respective genome counts displayed next to the STs.

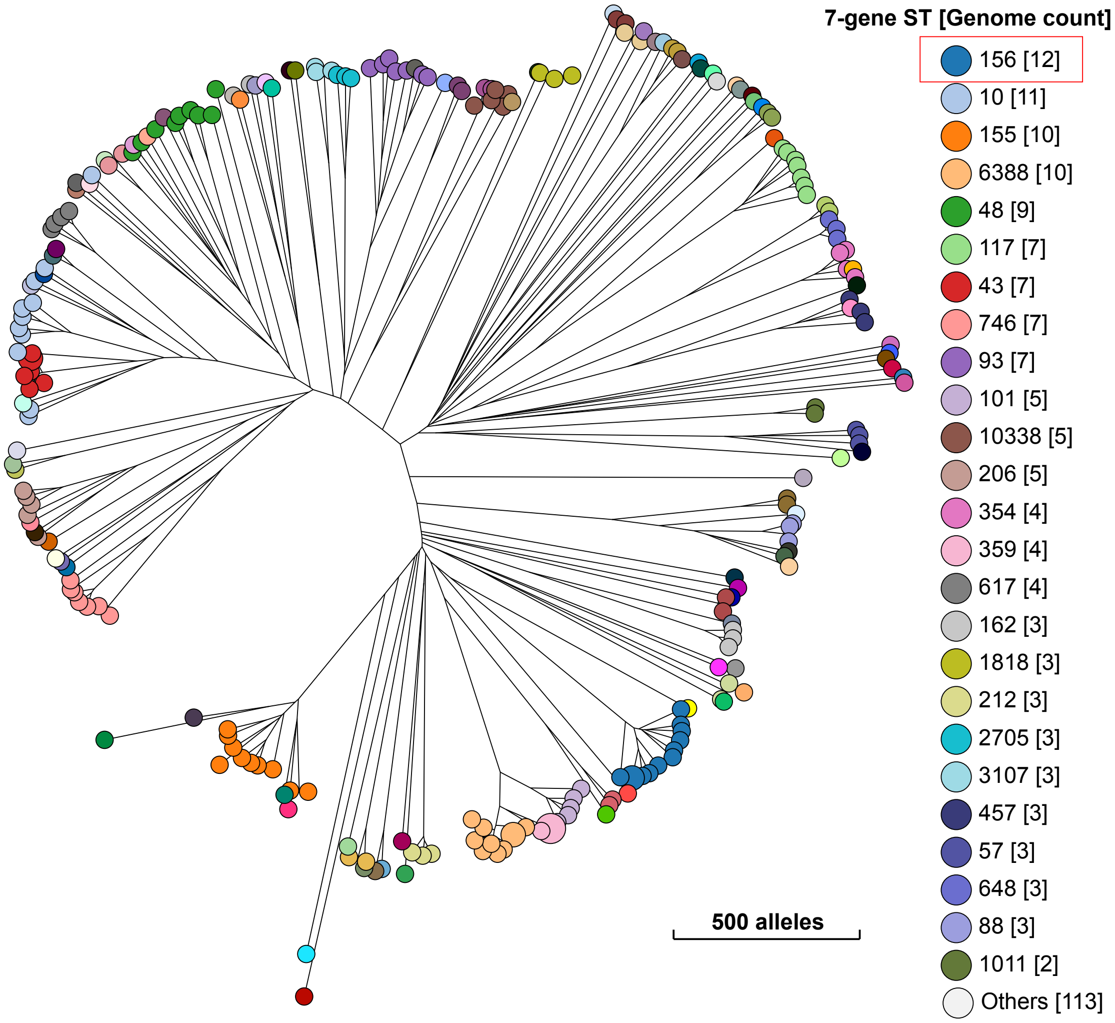


***Figure 4.13:*** *A NINJA neighbour-joining tree of all publicly available E. coli poultry isolates from Asia, showing the prevalence of Achtman sequence types (STs).*

The most common ST among poultry isolates from this continent is ST156 (highlighted with a red box). The legend displays the top 25 STs, with the respective genome counts displayed next to the STs.

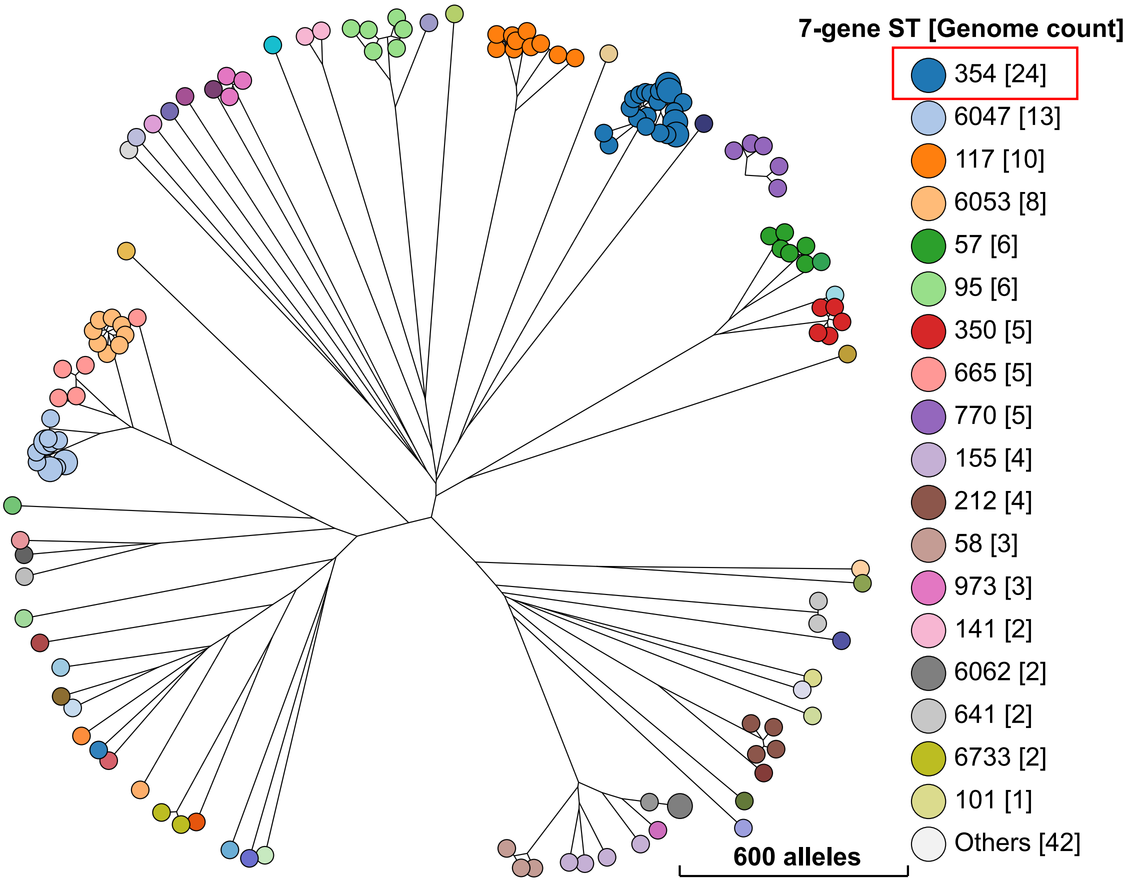
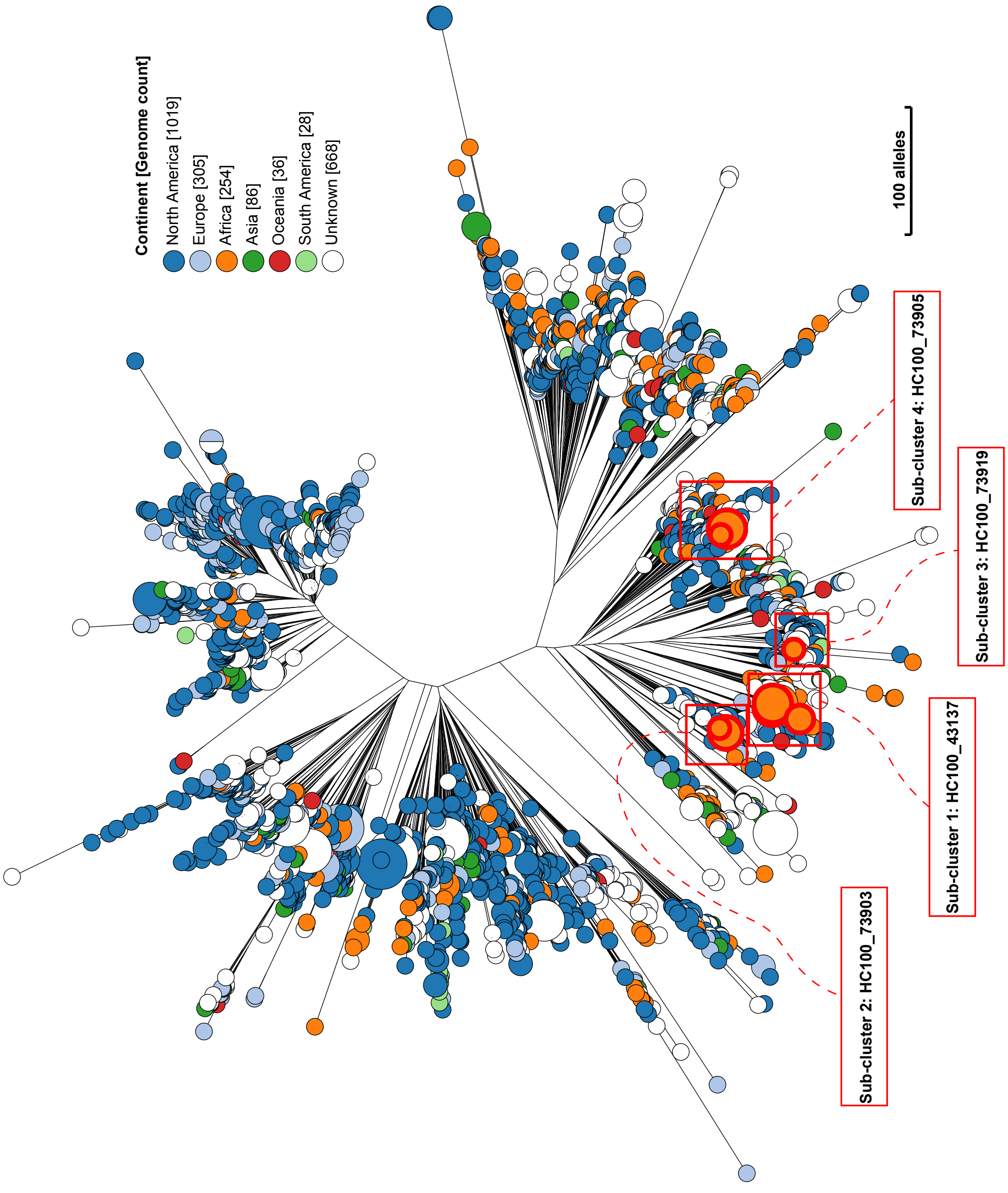


***Figure 14.14:** A NINJA neighbour-joining tree of all publicly available E. coli poultry isolates from Oceania, depicting the prevalence of Achtman sequence types (STs).*

The most common ST found among *E. coli* isolates from this continent is ST354 (highlighted with a red box). The first 18 STs are displayed in the legend, with the respective genome counts displayed.



*A phylogenetic tree showing the global distribution of E. coli ST155 isolates. My study ST155 isolates are highlighted with red circles. Hierarchical clustering resolved four sub-clusters, encompassing my Gambian ST155 strains , displayed in red boxes. The legend displays the locations of the isolates, with the genome counts depicted in square brackets.*

***Figure 4.15:*** *A phylogenetic tree showing the global distribution of E. coli ST155 isolates.*

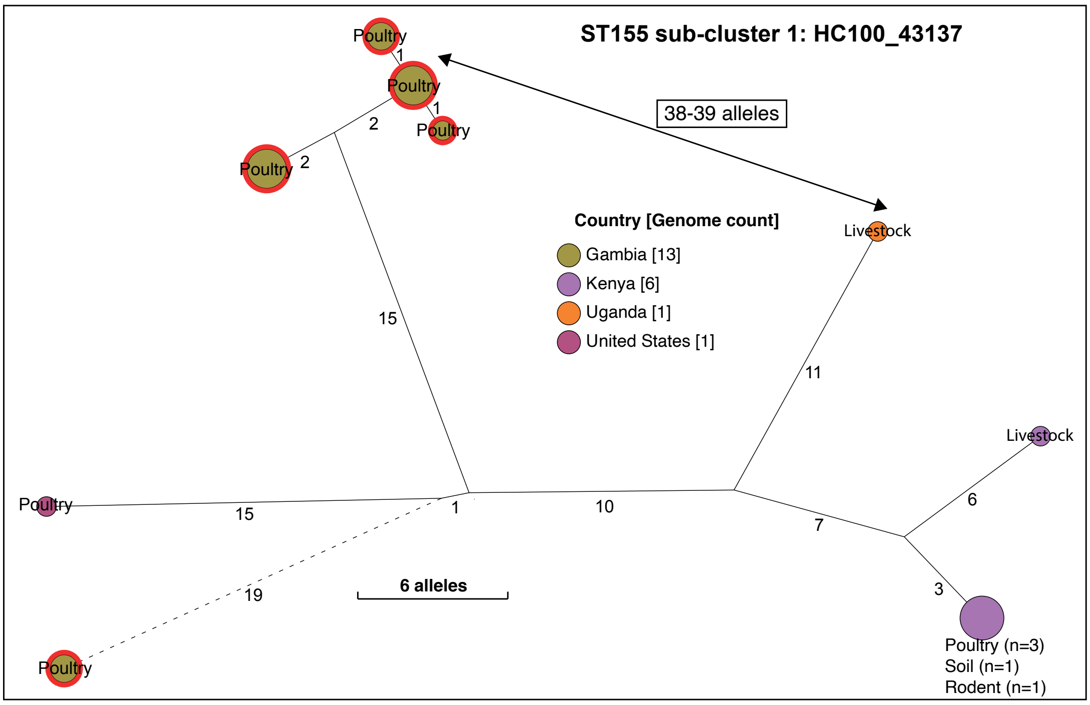


***Figure 4.16:*** *Population structure of ST155.*

NINJA phylogenetic trees showing the largest sub-clusters for the study ST155 population within the cgMLST hierarchical clustering scheme (the HC100_43137 sub-cluster) encompassing most of the study ST155 isolates (13/22, 59%), which were closely related to isolates from poultry and livestock in sub-Saharan Africa (separated by 38-39 alleles).

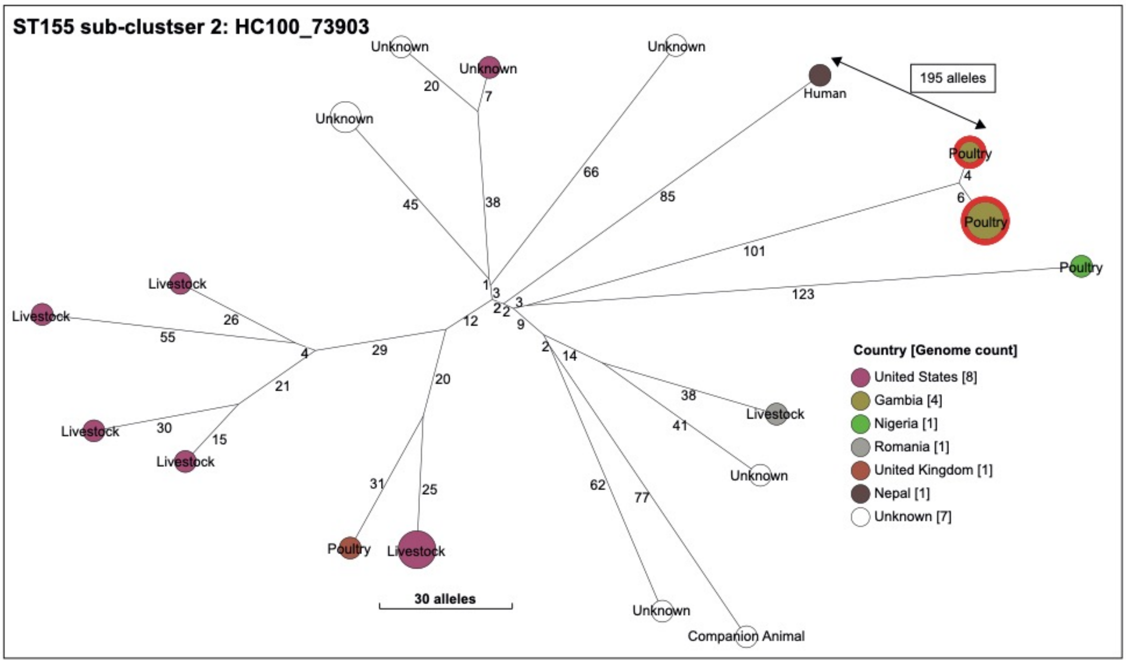


***Figure 4.17:*** *Phylogenetic relationships of isolates within the ST155 sub-cluster 2.*

A NINJA phylogenetic tree depicting the second sub-cluster within the study ST155 population (sub-cluster HC100_73903), comprising strains unique to the Gambia, although distantly related to isolates from humans

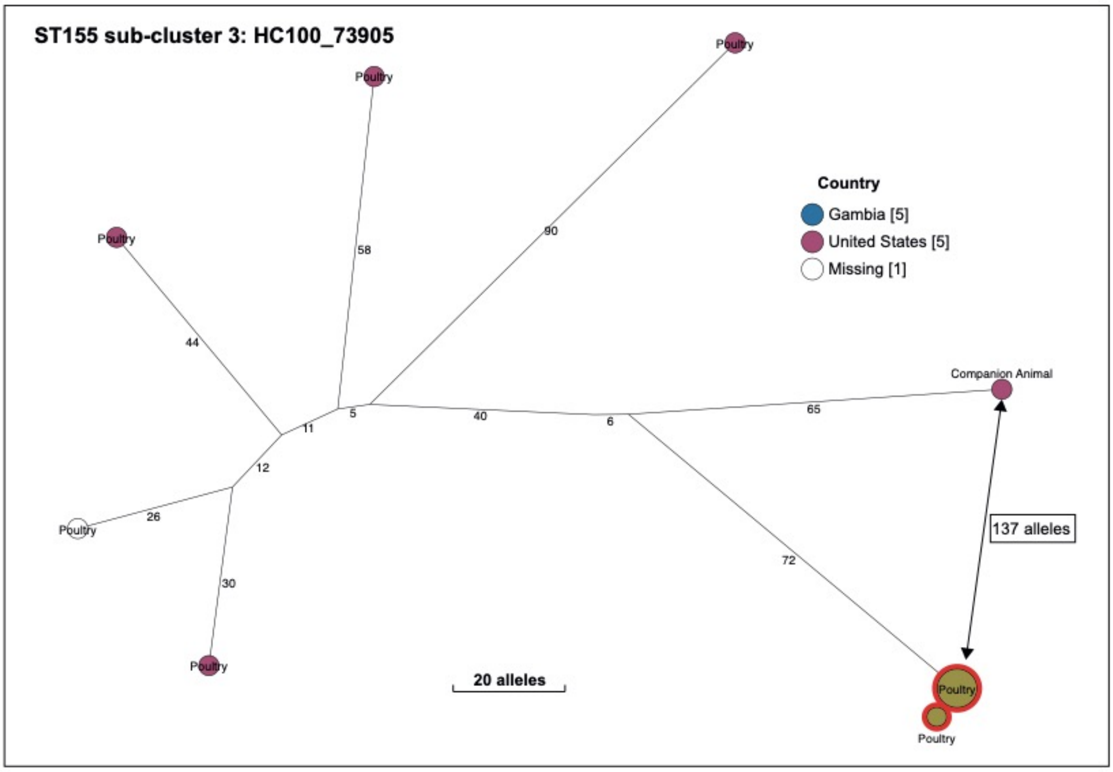


***Figure 4.18:*** *Phylogenetic relationships of isolates within the ST155 sub-cluster 3.*
A NINJA phylogenetic tree depicting the third sub-cluster within the study ST155 population (sub-cluster HC100_73905), comprised of strains unique to the Gambia, although distantly related to an isolate from a companion animal.

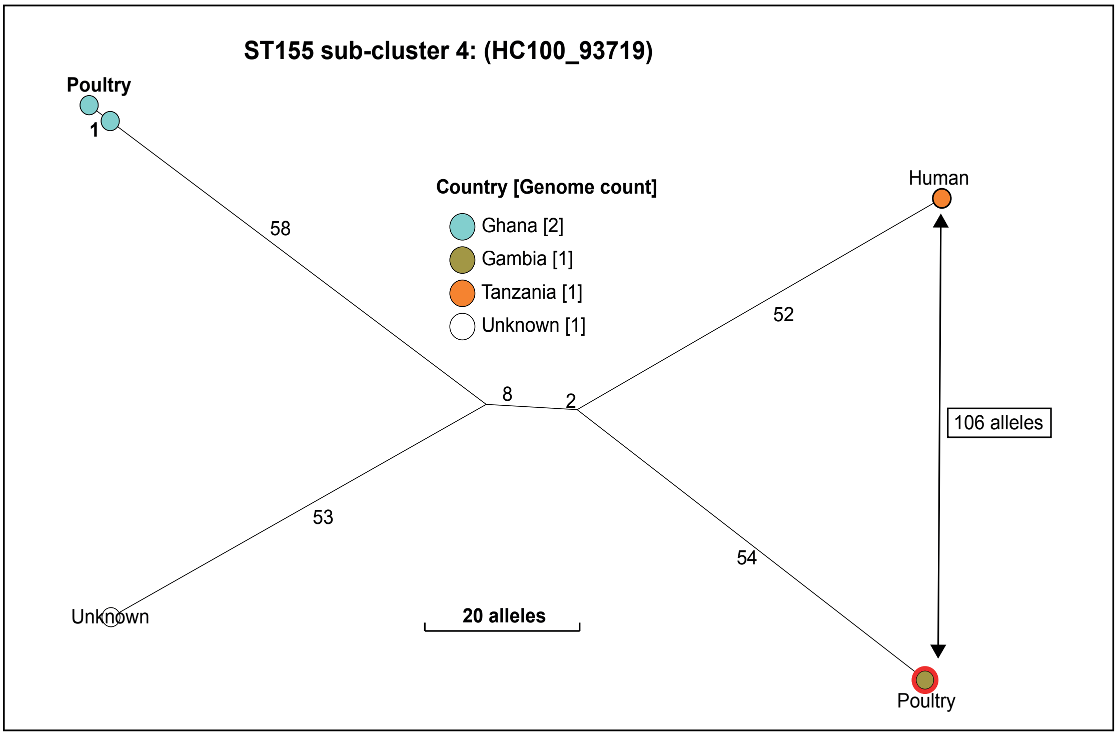


***Figure 4.19:*** *Phylogenetic relationships of isolates within the ST155 population sub-cluster 4.*
A NINJA phylogenetic tree depicting the fourth sub-cluster within the study ST155 population (sub-cluster HC100_93719), comprised of strains unique to the Gambia, although distantly related to an isolate from a human.

## 4.4 Discussion

Here, I have described the genomic diversity of *E. coli* from backyard chickens and guinea fowl reared in households in rural Gambia, West Africa. Backyard poultry from this rural setting harbour a remarkably diverse population of *E. coli* strains that encode antimicrobial-resistance genes and virulence factors important for infections in humans. Furthermore, I provide evidence of sharing of strains (including MDR strains) from poultry to poultry and between poultry, livestock and humans, with potential implications for public health.

My results reflect the rich diversity that exists within the *E. coli* population from backyard poultry. Although my sample size was small (19 birds), I recovered as many as 28 sequence types of *E. coli*, four of which have not been seen before. Three of the novel STs differed by >945 alleles from their nearest relative. Two of these now have complete assemblies. Also, some of the strains from this study were found in unique cgMLST HierCC clusters containing strains only from this study.

My results confirm previous reports that phylogroups B1 and A are dominant phylogroups among *E. coli* isolates from both intensive and backyard poultry [585, 593-595]. Hierarchical clustering analysis suggested that ST155 is common in African poultry. However, most of my study ST155 strains belong to a unique cgMLST cluster containing closely related (38-39 alleles differences and so presumably recently diverged) isolates from poultry and livestock from sub-Saharan Africa, suggesting that strains can be exchanged between livestock and poultry in this setting.

Rural backyard poultry can act as a source of transmission of infections to humans, due to the absence of biosecurity and daily contact with humans [596]. Indirect contact might occur through food or through contact with faeces, for example by children who are often left to play on the ground [597].

I observed a high prevalence of AMR genes among *E. coli* isolates sourced from African poultry. Similarly, high rates of genotypic MDR were detected among poultry *E. coli* isolates from the rest of the world, with ESBL (various types) being the most significant resistant gene detected. Poultry-associated ESBL genes have also been found among human clinical isolates [598]. Strikingly, most of my ST155 isolates encoded resistance to ≥3 classes of clinically relevant antibiotics, with the highest percentages to *bla*$_{\text{TEM-1}}$ beta-lactamase and tetracycline. This is worrying, as beta-lactamase-positive isolates are often resistant to several other classes of antibiotics [599, 600].

My results are consistent with previous studies that reported ST155 isolates to be commonly associated with MDR [601, 602], and with reports of a low prevalence of ESBL in backyard poultry. For example, in a study that compared the prevalence of ESBL genes in backyard poultry and commercial flocks from West Bengal, India, none of the 272 *E. coli* isolates from backyard birds harboured any ESBL gene [603], while 30% of commercial birds carried ESBL genes. The absence of resistance in that study was attributed to a lack of exposure to antimicrobials. Similarly, *E. coli* from organic poultry in Finland were reported to be highly susceptible to most of the antimicrobials studied and no ESBL resistance was detected [604].

Although tetracycline is commonly used in poultry farming for therapeutic purposes [605], resistance to this antibiotic is known to be prevalent in poultry, even in the absence of the administration of this antibiotic [215]. My results also suggest that IncF plasmids may play a role in the dissemination of AMR in my study population. A limitation of my study is that due to Covid-19 restrictions, I could not perform conjugation assays to confirm the association of these plasmids with the observed resistance genes and the mobilisability of the plasmids and thus, the potential for exchange among co-colonising strains in a single host.

Many sub-Saharan countries lack clear guidelines on the administration of antibiotics in agriculture, although an increasing trend in the veterinary use of antimicrobials has been documented [606]. The usage of antimicrobials in developing countries is likely to increase because of increasingly intensive farming practices [195]. Europe has banned the use of antimicrobials as growth promoters since 2006 [194] and the use of all essential antimicrobials for prophylaxis in animal production since 2011 [607]. However, AMR may be less well controlled in other parts of the world.

Although APEC strains span several phylogroups (A, B1, B2 and D) and serogroups [592], the majority of APEC strains encode virulence genes associated with intestinal or extra-intestinal disease in humans. These include adhesion factors, toxins, iron-acquisition genes and genes associated with serum resistance, such as *fyuA*, *iucD*, *iroN*, *iss*, *irp2, hlyF*, *vat, kpsM* and *ompT*. Although APEC isolates present different combinations of virulence factors, each retains the capability to cause colibacillosis [578, 608]. I did not detect haemolysin or serum survival genes in my study isolates; however, I recovered some of the known markers of intestinal and extraintestinal virulence in some study isolates, such as the enteroaggregative *E. coli* heat-stable enterotoxin and the vacuolating autotransporter toxin (*vat*, *astA*), invasion and evasion factors (*kpsM*, *kpsD*, *pla*) and adherence factors (*fim* and *pap* genes) that are associated with intestinal and extraintestinal infections in humans. Thus, these strains could cause disease in humans, should they gain access to the appropriate tissues.

Several birds were colonised with two or more STs and at least two phylotypes of *E. coli.* This level of diversity is probably a consequence of the frequent exposure of backyard poultry to the environment, livestock and humans. Co-colonisation of single hosts with multiple strains may facilitate the spread of AMR- and virulence-associated genes from resistant strains to other bacteria via both horizontal and vertical gene transfer [609]. A high co-colonisation rate of *E. coli* has been described in humans [505, 506] and in non-human primates [610], involving pathogenic strains of *E. coli*. Recently, Li *et al* reported three to nine sequence types of colistin-resistant *E. coli* to co-exist within a single broiler chicken [588]. Here, I report co-colonisation with different lineages of *E. coli* in backyard chickens and guinea fowl. Unsurprisingly, co-colonising strains often had different AMR and virulence patterns.

### 4.4.1 Limitations

An obvious limitation of my study is the small sample size. This study could have also been enhanced by sampling *E. coli* from humans within close proximity to my backyard birds, however, I could not perform an analysis of *E. coli* from sympatric humans from my study setting due to logistic reasons and the opportunistic nature of my study. However, the inclusion of publicly available sequences strengthens my analysis and inference of the population of *E. coli* in this setting. I also could not perform phenotypic susceptibility testing on all isolates. I acknowledge that a minor percentage of genotypic resistance predictions fail to correspond with phenotypic resistance [611].

Taken together, my results indicate a rich diversity of *E. coli* within backyard poultry from the Gambia, characterised by strains with a high prevalence of AMR and the potential to contribute to infections in humans. This, coupled with the potential for the exchange of strains between poultry and livestock within this setting, might have important implications for human health and warrants continued surveillance.

# 5 CHAPTER FIVE: GENOMIC DIVERSITY OF *ESCHERICHIA COLI* ISOLATES FROM HEALTHY CHILDREN FROM RURAL GAMBIA

## 5.1 Introduction

As discussed in the main introduction (Chapter 1), ease of culture and genetic tractability account for the unparalleled status of *Escherichia coli* as "the biological rock star", driving advances in biotechnology [12], while also providing critical insights into biology and evolution [35]. However, *E. coli* is also a widespread commensal, as well as a versatile pathogen, linked to diarrhoea (particularly in the under-fives), UTI, neonatal sepsis, bacteraemia and multi-drug resistant infection in hospitals [612-614]. Yet, most of what we know about *E. coli* stems from the investigation of laboratory strains, which fail to capture the ecology and evolution of this key organism "in the wild" [615]. What is more, most studies of non-lab strains have focused on pathogenic strains or have been hampered by low-resolution PCR methods, so we have relatively few genomic sequences from commensal isolates, particularly from low- to middle-income countries [359, 498, 616-620].

As already discussed in Chapter 1, we have a broad understanding of the population structure of *E. coli,* with eight significant phylogroups loosely linked to ecological niche and pathogenic potential (B2, D and F linked to extraintestinal infection; A and B1 linked to severe intestinal infections such as haemolytic-uraemic syndrome) [14, 109, 118, 592]. All phylogroups can colonise the human gut, but it remains unclear how far commensals and pathogenic strains compete or collaborate—or engage in horizontal gene transfer—within this important niche [609, 621].

Although clinical microbiology typically relies on single-colony picks (which has the potential to underestimate species diversity and transmission events), within-host diversity of *E. coli* in the gut is crucial to our understanding of inter-strain competition and co-operation and also for accurate diagnosis and epidemiological analyses. Pioneering efforts using serotyping, molecular typing and whole-genome sequencing have shown that normal individuals typically harbour more than one strain of *E. coli*, with one individual carrying 24 distinct clones [249, 413-415, 505]. More recently, whole-genome sequencing has illuminated molecular epidemiological investigations [498], for example, studies of the transmission of ESBL-encoding *E. coli*, multidrug-resistant *Acinetobacter baumannii* and the genomic surveillance of multidrug-resistant *E. coli* carriage. Whole-genome data has also been applied to studies of *E. coli* adaptation during and after infection [622, 623], as well as the intra-clonal diversity in healthy hosts [502].

There are two plausible sources of within-host genomic diversity. Although a predominant strain usually colonises the host for extended periods [624], successful immigration events mean that incoming strains can replace the dominant strain or co-exist alongside it as minority populations [395]. Strains originating from serial immigration events are likely to differ by hundreds or thousands of single-nucleotide polymorphisms (SNPs). Alternatively, within-host evolution can generate clouds of intra-clonal diversity, where genotypes differ by just a handful of SNPs [413].

Most relevant studies have been limited to Western countries, except for a recent report from Tanzania [414], so little is known about the genomic diversity of *E. coli* in sub-Saharan Africa. The Global Enteric Multicenter Study (GEMS) [515, 516] has documented a high burden of diarrhoea attributable to *E. coli* (including *Shigell*a) among children from the Gambia, probably as a result of increased exposure to this organism through poor hygiene and frequent contact with animals and the environment. GEMS was a prospective case-control study which investigated the aetiology of moderate-to-severe diarrhoea in children aged less than five years residing in sub-Saharan Africa and South Asia. In the Gambia, children with moderate-to-severe diarrhoea seeking care at the Basse Health centre in the Upper River Division of the country were recruited, with one to three matched control children randomly selected from the community along with each case.  In also facilitating access to stool samples from healthy Gambian children, the GEMS study has given us a unique opportunity to study within-host genomic diversity of commensal *E. coli* in this

## 5.2 Materials and methods

### 5.2.1 Study population

I initially selected 76 faecal samples from three- to five-year-old (36-59 months) asymptomatic Gambian children, who had been recruited into the GEMS study [515] as healthy controls from 1st December 2007, to 3rd March 2011. Samples had been collected according to a previously described sampling protocol [625] and the results of the original study are publicly available at ClinEpiDB.org. Ten of the original 76 samples were depleted and were therefore unavailable for processing in this study. Of the remaining 66 stools, 62 had previously tested positive for *E. coli*. GEMS isolated three *E. coli* colonies per stool sample but pooled these into a single tube for frozen storage. Thus, I needed to re-culture the stools with multiple colony picks, as the original isolate collection was unsuitable for the investigation of within-host diversity. Archived stool samples were retrieved from -80°C storage and allowed to thaw on ice. A 100-200 mg aliquot from each sample was transferred aseptically into 1.8ml Nunc tubes for microbiological processing as below in the main methods section.

### 5.2.2 Inclusion of publicly available human isolates from the Gambia

Publicly available *E. coli* sequences in EnteroBase (http://enterobase.warwick.ac.uk/species/index/ecoli) [139] were included for comparative analysis, including 23 previously sequenced isolates obtained from diarrhoeal cases recruited in the GEMS study in the Gambia (File S2 in [626]).

### 5.2.3 Chapter-specific processes and overall sample processing workflow

A detailed explanation of the methods employed for the culture and isolation of *E. coli*, genomic DNA extraction, whole-genome sequencing and the subsequent bioinformatics analyses is presented in the main methods chapter (Chapter Two). Specific processes and analysis that pertained to only the human *E. coli* isolates analysed in this chapter are described below. Also, a flow chart that summarises the overall study sample processing workflow is presented (Figure 5.1).

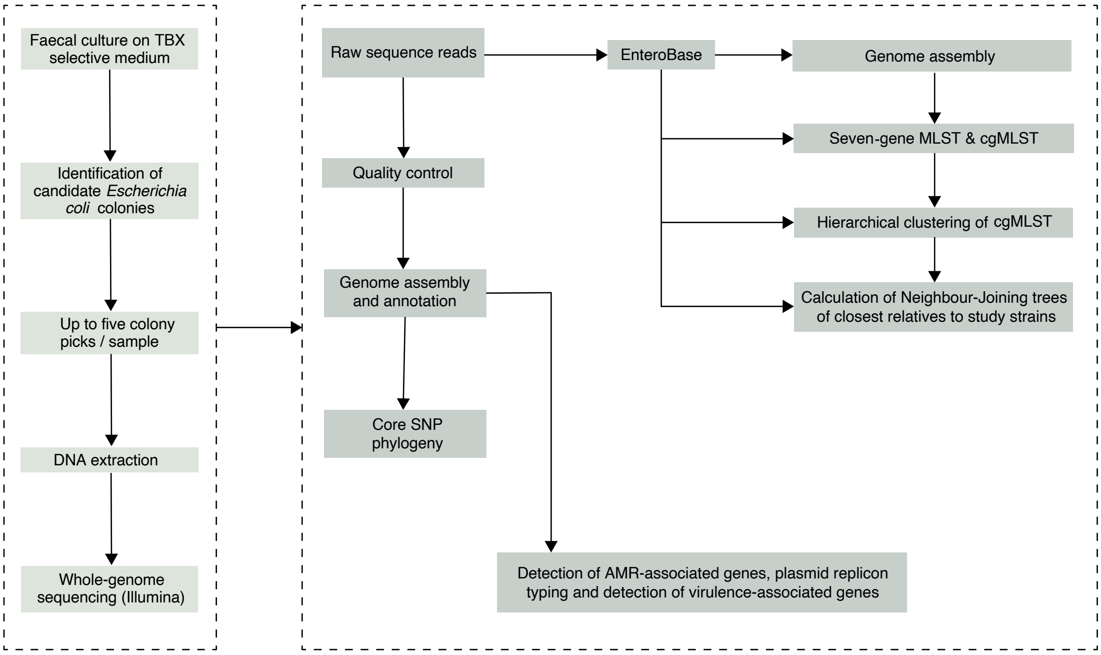

Faecal culture on TBX selective medium
Identification of candidate *Escherichia coli* colonies
Up to five colony picks / sample
DNA extraction
Whole-genome sequencing (Illumina)
Raw sequence reads
Quality control
Genome assembly and annotation
Core SNP phylogeny
EnteroBase
Genome assembly
Seven-gene MLST & cgMLST
Hierarchical clustering of cgMLST
Calculation of Neighbour-Joining trees of closest relatives to study strains
Detection of AMR-associated genes, plasmid replicon typing and detection of virulence-associated genes

***Figure 5.1:*** *A flowchart summarising the study sample processing workflow.*

### 5.2.4 Isolates that were sequenced twice

Following Dixit *et al.* [413], I sequenced a random selection of ten isolates twice, using DNA obtained from independent cultures, to help in the determination of clones and the analysis of within-host variants (Table 5.3).

***Table 5.1:*** *List of the sample clones for which two independent cultures were obtained and sequenced, to find the SNPs between the same clones.*

| Individual | Sample | Coverage | N50 | Total length | # Contigs |
|---|---|---|---|---|---|
| 32 | H-32_1 (1) | 25 | 229898 | 4806091 | 70 |
|  | H-32_1 (2) | 52 | 131282 | 4816766 | 83 |
| 34 | H-34_1 (1) | 42 | 154777 | 4806091 | 81 |
|  | H-34_1 (2) | 33 | 105841 | 4644423 | 116 |
| 34 | H-34_3 (1) | 30 | 258099 | 4639908 | 54 |
|  | H-34_3 (2) | 52 | 182362 | 4646674 | 76 |
| 36 | H-36_4 (1) | 50 | 351245 | 4879323 | 36 |
|  | H-36_4 (2) | 76 | 263944 | 4884447 | 50 |
| 37 | H-37_4 (1) | 34 | 134993 | 5388081 | 176 |
|  | H-37_4 (2) | 45 | 92748 | 5379306 | 215 |
| 37 | H-37_5 (1) | 33 | 92298 | 5274674 | 228 |
|  | H-37_5 (2) | 41 | 60606 | 5294739 | 242 |
| 38 | H-38_1 (1) | 22 | 152501 | 5327666 | 123 |
|  | H-38_1 (2) | 43 | 116988 | 5350616 | 134 |
| 38 | H-38_5 (1) | 49 | 166358 | 5333851 | 126 |
|  | H-38_5 (2) | 45 | 104003 | 5346499 | 165 |
| 39 | H-39_2 (1) | 34 | 192437 | 4997502 | 185 |
|  | H-39_2 (2) | 55 | 156538 | 5039316 | 166 |
| 41 | H-41_2 (1) | 75 | 185894 | 4872981 | 92 |
|  | H-41_2 (2) | 128 | 185391 | 4893458 | 104 |

### 5.2.5 Determination of immigration events and within-host variants

For the whole genome sequences of the strains sequenced twice, I used SPAdes v3.13.2 [314] to assemble each set of reads and map the raw sequences from one sequencing run to the assembly of the other run and vice versa, as described previously [413]. Briefly, mapping was done using the BWA-MEM algorithm v0.7.17-r1188 under default parameters to generate a SAM alignment. This was then converted to BAM files using Samtools view v1.9 [335], sorted and indexed. Next, variants were called and written to a VCF file using Samtools mpileup and the "view" module of BCFtools (which is part of the Samtools v1.9 package) and visualised in Tablet v1.19.09.13 [627]. The number of SNPs and their positions were determined and compared between the two steps, counting only those SNPs that were detected in both sets of reads as accurate.

In line with [413], isolates belonging to different STs recovered from the same host were considered to be separate strains derived from independent exposures and immigration events.

As described in [413], I determined the number of SNP differences that existed between assemblies of the same isolate that were sequenced on two separate occasions, to determine if multiple isolates of the same ST from a single host were distinct variants (clones). If the SNP difference between two isolates belonging to the same ST recovered from the same host was less than the SNP difference between the sequences of the same isolate sequenced on two separate occasions, then the two isolates were taken to represent replicate copies of the same clone. Otherwise, they were considered as within-host variants (separate, distinct clones of the same strain)—provided the SNP differences between such distinct clones were no more than eleven SNPs. This cut-off was chosen based on an estimated mutation rate of 1.1 SNP per genome per year [407], assuming equal rates of mutation in both genomes being compared. Based on these data, I inferred replicate clones with SNP differences of greater than 11 SNPs to represent a divergence of more than five years. Thus, it seems implausible that such replicate clones would have emerged from within-host evolution, considering the age of the study participants (<5 years old).

I produced a contingency table to summarise the distribution of variants derived from migration events and within-host evolution and visualised this using a clustered bar graph. I then performed Fisher's exact test to investigate the association between phylogroup and the distribution of variants (migration versus within-host evolution). I based my calculations on the assumption of independence among the observed phylogroups—that is, the finding of one phylogroup does not preclude or predict the co-occurrence of another.

## 5.3 Results

### 5.3.1 Population structure

The study population included 27 females and 39 males (Table 5.3). All but one reported the presence of a domestic animal within the household. Twenty-one samples proved positive for the growth of *E. coli*, yielding 88 isolates (File S4 in [626]). I detected 37 seven-allele sequence types (STs) among the isolates, with a fairly even distribution (Figure 5.2). Five STs were completely novel (ST9274, ST9277, ST9278, ST9279 and ST9281). These study strains were scattered over all the eight main phylogroups of *E. coli*: A (27%), B1 (32%), B2 (9%), D (15%), C and F (5% each), E (1%) and the cryptic Clade I (7%), although the majority belonged to phylogroups A and B1 (Table 5.4). Hierarchical clustering of core genomic STs revealed twenty-seven cgST clonal complexes (File S4 in [626]). The raw genomic sequences of the study isolates have been deposited in the NCBI SRA under the BioProject ID PRJNA658685 (accession numbers SAMN15880274 to SAMN15880361).

***Table 5.2:*** *Characteristics of the study population.*

| Sample ID | Lab ID | Age (months) | Gender | Bristol stool index | Domestic animal within household | Enrolment date |
|---|---|---|---|---|---|---|
| 102135 | H1 | 43 | Female | Thick liquid | Goat, sheep | 18-Feb-09 |
| 102650 | H2 | 45 | Female | Soft | Goat, sheep, donkey | 27-Jul-09 |
| 103296 | H3 | 44 | Male | Soft | Goat, horse, donkey, rodent | 27-Apr-10 |
| 103298 | H4 | 44 | Male | Formed | Sheep, fowl, horse, donkey, rodent | 27-Apr-10 |
| 103621 | H5 | 37 | Female | Soft | Sheep, fowl, rodent | 01-Sep-10 |
| 103650 | H6 | 48 | Female | Soft | Fowl, donkey, rodent | 29-Sep-10 |
| 103649 | H7 | 45 | Female | Soft | Goat, sheep, fowl, horse, rodent | 29-Sep-10 |
| 103071 | H8 | 53 | Male | Formed | Goat, sheep, fowl | 15-Jan-10 |
| 103622 | H9 | 39 | Female | Soft | Goat, sheep | 01-Sep-10 |
| 100167 | H10 | 40 | Female | Soft | Goat, sheep, fowl | 01-Feb-08 |
| 100217 | H11 | 57 | Male | Formed | Cat, fowl, horse, rodent | 21-Feb-08 |
| 100230 | H12 | 51 | Male | Soft | Goat, sheep, cat, fowl, rodent | 28-Feb-08 |
| 100612 | H13 | 55 | Female | Formed | Goat, sheep, dog, fowl, horse, donkey, rodent | 16-Aug-08 |
| 100162 | H14 | 47 | Female | Thick liquid | Sheep, horse, donkey, rodent | 30-Jan-08 |
| 102255 | H15 | 42 | Male | Formed | Goat, sheep, fowl, horse, donkey, rodent | 26-Mar-09 |
| 102250 | H16 | 39 | Male | Formed | Fowl | 25-Mar-09 |
| 102114 | H17 | 54 | Male | Formed | Rodent | 12-Feb-09 |
| 102123 | H18 | 37 | Female | Soft | Goat, sheep, fowl, rodent | 14-Feb-09 |
| 103282 | H19 | 43 | Male | Formed | Goat, sheep, dog, cat, cow, fowl, | 22-Apr-10 |
| 100817 | H20 | 44 | Male | Soft | Dog, fowl | 03-Dec-08 |
| 100816 | H21 | 40 | Male | Soft | Goat, sheep, cow, fowl, horse, donkey, rodent | 03-Dec-08 |
| 102836 | H22 | 47 | Male | Thick liquid | Fowl, rodent | 12-Oct-09 |
| 102837 | H23 | 41 | Male | Thick liquid | Sheep, fowl, rodent | 12-Oct-09 |
| 102843 | H24 | 44 | Male | Soft | Fowl, rodent | 13-Oct-09 |
| 102907 | H25 | 36 | Male | Soft | Goat, sheep, fowl | 05-Nov-09 |
| 102905 | H26 | 37 | Male | Soft | Goat, sheep, fowl | 05-Nov-09 |
| 102262 | H27 | 38 | Male | Formed | Goat, sheep, rodent | 01-Apr-09 |
| 102728 | H28 | 41 | Male | Soft | Goat, fowl | 24-Aug-09 |
| 102729 | H29 | 41 | Male | Soft | Goat, dog, cat, fowl, donkey | 24-Aug-09 |
| 100806 | H30 | 55 | Male | Soft | Goat, sheep, dog, fowl | 21-Nov-08 |
| 102053 | H31 | 37 | Female | Formed | Cow, fowl, donkey, rodent | 29-Jan-09 |
| 102052 | H32 | 38 | Female | Formed | Goat, sheep, cow, fowl, donkey, rodent | 29-Jan-09 |
| 102511 | H33 | 37 | Male | Soft | Fowl, horse, donkey, rodent | 19-Jun-09 |
| 102649 | H34 | 37 | Male | Soft | Fowl, horse, donkey, rodent | 27-Jul-09 |
| 102454 | H35 | 52 | Male | Soft | Sheep, fowl, donkey, rodent | 02-Jun-09 |
| 102459 | H36 | 51 | Male | Formed | Goat, sheep, dog, cat, cow, horse, donkey, rodent | 04-Jun-09 |
| 100303 | H37 | 58 | Male | Formed | Sheep, fowl | 08-Apr-08 |
| 100320 | H38 | 42 | Female | Formed | Sheep, fowl, rodent | 19-Apr-08 |
| 100319 | H39 | 45 | Female | Formed | Goat, sheep, fowl, rodent | 17-Apr-08 |
| 103081 | H40 | 39 | Female | Thick liquid | Goat, sheep, fowl, horse, donkey, rodent | 20-Jan-10 |
| 103082 | H41 | 39 | Female | Thick liquid | Goat, sheep, fowl, horse, donkey, rodent | 20-Jan-10 |
| 100663 | H42 | 36 | Male | Thick liquid | Goat, sheep, fowl, donkey | 10-Sep-08 |
| 100072 | H43 | 51 | Female | Formed | Goat, cow, fowl, rodent | 03-Jan-08 |
| 103171 | H44 | 36 | Female | Soft | Goat, sheep, rodent, fowl, rodent | 18-Feb-10 |

***Table 5:3:*** *Characteristics of the study population (continued).*

| Sample ID | Lab ID | Age (months) | Gender | Bristol stool index | Domestic animal within household | Enrolment date |
|---|---|---|---|---|---|---|
| 103172 | H45 | 36 | Female | Soft | Goat, sheep, fowl, rodent | 18-Feb-10 |
| 103292 | H46 | 39 | Male | Soft | Goat, sheep, fowl | 23-Apr-10 |
| 102952 | H47 | 36 | Male | Soft | Goat, sheep, fowl, rodent | 20-Nov-09 |
| 102953 | H48 | 37 | Male | Soft | Goat, sheep, fowl, rodent | 20-Nov-09 |
| 102964 | H49 | 40 | Female | Formed | Goat, fowl, rodent | 26-Nov-09 |
| 102966 | H50 | 37 | Female | Formed | Goat, sheep, fowl, horse, donkey, rodent | 22-Apr-10 |
| 103281 | H51 | 44 | Male | Formed | Goat, sheep, dog, cat, fowl | 22-Apr-10 |
| 100540 | H52 | 43 | Male | Soft | Goat, sheep, fowl, rodent | 22-Jul-08 |
| 103123 | H53 | 38 | Male | Soft | Sheep | 03-Feb-10 |
| 103124 | H54 | 36 | Male | Soft | Fowl | 03-Feb-10 |
| 102089 | H55 | 38 | Female | Soft | Goat, cow, fowl, horse, donkey, rodent | 05-Feb-09 |
| 103297 | H56 | 38 | Male | Soft | Goat, sheep, fowl, horse, donkey, rodent | 27-Apr-10 |
| 102251 | H57 | 39 | Male | Formed | Fowl | 25-Mar-09 |
| 103602 | H58 | 38 | Female | Formed | Goat, sheep, cow, fowl | 26-Aug-10 |
| 103600 | H59 | 39 | Female | Formed | Goat, sheep, fowl | 26-Aug-10 |
| 100026 | H60 | 49 | Female | Soft | Goat, sheep, cow, fowl | 14-Dec-07 |
| 102102 | H61 | 47 | Female | Opaque watery | None | 11-Feb-09 |
| 102263 | H62 | 38 | Male | Formed | Horse, donkey, rodent | 01-Apr-09 |
| 103070 | H63 | 58 | Male | Soft | Goat, sheep, fowl | 15-Jan-10 |
| 103130 | H64 | 40 | Male | Soft | Sheep, fowl | 03-Feb-10 |
| 102051 | H65 | 36 | Female | Formed | Goat, sheep, dog, cat, cow, fowl, donkey, rodent | 29-Jan-09 |
| 102524 | H66 | 36 | Male | Soft | Goat, sheep, fowl, horse, donkey, rodent | 24-Jun-09 |

*Table 5.4: Phylogroup and sequence types of the distinct clones isolated from each study subject.*

| Host | Colony or isolate number | | | | | Number of distinct genotypes (clones) | Migration events | Within-host evolution events |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | Phylotype (number of events) | Phylotype (number of events) |
| H-2 | A (9274) | A (9274) | A (9274) | A (9274) | A (9274) | 1 | A (1) | 0 |
| H-9 | A (2705) | A (2705) | A (2705) | D (2914) | B1 (29) | 3 | A (1), D (1), B1 (1) | 0 |
| H-15 | B2 (9277) | B2 (9277) | B2 (9277) | Clade I (747) | Clade I (747) | 3 | B2 (1), Clade I (1) | Clade I (1) |
| H-18 | D (38) | D (38) | B1 (9281) | A (9274) | | 4 | D (1), B1 (1), A (1) | D (1) |
| H-21 | B1 (58) | B1 (58) | B1 (223) | A (540) | D (1204) | 4 | B1(2) A (1), D (1) | 0 |
| H-22 | B1 (316) | B1 (316) | B1 (316) | B1 (316) | | 2 | B (1) | B1(1) |
| H-25 | A (181) | A (181) | A (181) | A (181) | B1 (337) | 4 | A (1), B1 (1) | A (2) |
| H-26 | B1 (641) | B1 (2741) | A (10) | A (398) | | 4 | B1(2), A (1), D (1) | 0 |
| H-28 | B1 (469) | B1 (469) | B1 (469) | B1 (469) | | 2 | B1(1) | B1(1) |
| H-32 | B1 (101) | B1 (101) | B1 (101) | B1 (2175) | A (10) | 3 | B1(2), A (1) | 0 |
| H-34 | B1 (603) | B1 (603) | B1 (603) | B1 (1727) | A (10) | 4 | B1(2), A (1) | B1(1) |
| H-35 | A (226) | | | | | 1 | A (1) | 0 |
| H-36 | F (59) | F (59) | F (59) | F (59) | E (9278) | 4 | F (1), E (1) | F (1) |
| H-37 | D (5148) | D (5148) | D (5148) | D (5148) | D (5148) | 3 | D (1) | D (2) |
| H-38 | D (394) | D (394) | D (394) | D (394) | B1 (58) | 4 | D (1), B1(1) | D (2) |
| H-39 | B2 (452) | B2 (452) | B2 (452) | B2 (452) | B2 (452) | 2 | B2(1) | B2 (1) |
| H-40 | B1 (155) | | | | | 1 | B1(1) | 0 |
| H-41 | A (43) | A (43) | A (43) | A (43) | B1 (9283) | 2 | A (1), B1(1) | 0 |
| H-48 | Clade I (485) | Clade I (485) | Clade I (485) | Clade I (485) | | 3 | Clade I (1) | 0 |
| H-50 | C (410) | C (410) | C (410) | C (410) | B1 (515) | 2 | C (1), B1(1) | 0 |
| H-55 | A (9279) | | | | | 1 | A (1) | 0 |

### 5.3.2 Within-host diversity

Just a single ST colonised nine individuals, six carried two STs, four carried four STs and two carried six STs. I found 56 distinct genotypes, which equates to an average of 2.7 genotypes per host. Two individuals (H-18 and H-2) shared an identical strain belonging to ST9274 (zero SNP difference) (File S5 in [626], yellow highlight), suggesting recent transfer from one child to another or recent acquisition from a common source. I observed thirteen within-host variants in ten hosts (intra-clonal diversity) (subjects H-15, H-18, H-22, H-25, H-28, H-34, H36, H37, H-38 and H-39), compared to forty-one immigration events (Tables 5.4 and 5.5).

Overall, immigration events accounted for the majority (76%) of variants (Figure 5.3). The proportion of migration versus within-host evolution events did not appear to be affected by phylogroup (p=0.42). Twenty-two percent of within-host mutations represented synonymous changes, 43% were non-synonymous mutations, while 31% occurred in non-coding regions and 4% represented stop-gained mutations (File S3 in [626]). On an average, Ka/Ks ratios were greater than 1, which seems to suggest that these mutations were under positive Darwinian selection—indicating that most of the mutations were likely to have little effect on fitness. However, these remain to be investigated further. Also, the observed non-synonymous mutations were spread across genes with a variety of functions, including metabolism, transmembrane transport, pathogenesis and iron import into the cell. However, the bulk (42%) occurred in genes involved in metabolism. The average number of SNPs among within-host variants was 5 (range 0-18) (Table 5.5). However, in two subjects (H36 and H37), pairwise distances between genomes from the same ST (ST59 and ST5148) were as large as 14 and 18 SNPs respectively (File S5 in [626], grey highlight).

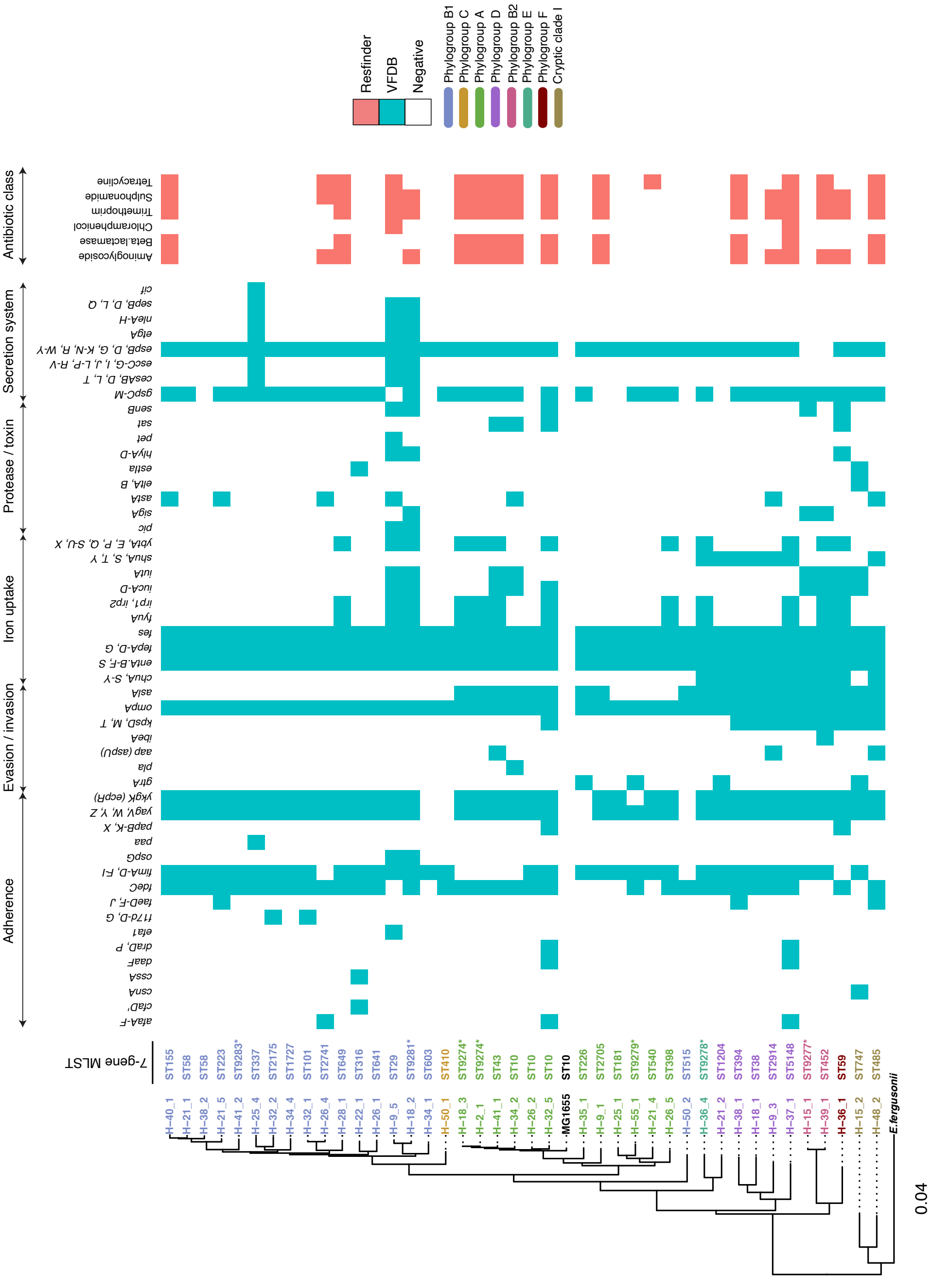
### 5.3.3 Accessory gene content and relationships with other strains

A quarter of my isolates were most closely related to commensal strains from humans, with smaller numbers most closely related to human pathogenic strains or strains from livestock, poultry or the environment (Table 5.6). One isolate was most closely related to a canine isolate from the UK. Three STs (ST38, ST10 and ST58) were shared by my study isolates and diarrhoeal isolate from the GEMS study (Figure 5.4), with just eight alleles separating my commensal ST38 strain from a diarrhoeal isolate from the GEMS study (Figure 5.5).

For ST10 and ST58, hierarchical clustering placed the commensal strains from this study into separate clusters from the pathogenic isolates from diarrhoeal cases, indicating that they were genetically distinct to each other. Yet, the closest relative of my study ST58 strain was an extraintestinal strain isolated from the blood of a 69-year-old male (87 alleles differences, Figure 5.7). Also, the resident ST10 isolates recovered from this study (H-26_2, H-34_2 and H-32_5) had their closest neighbours in isolates from livestock (83 and 111 alleles each) and a human isolate of an unspecified sample source (18 alleles differences) respectively (Table 5.6).

I detected 130 genes encoding putative virulence factors across the 88 study isolates (Figure 5.2; File S8 in [626]). Notable among these were genes associated with pathogenesis in Enteroaggregative *E. coli* and *Salmonella* referred to as the Serine Protease Autotransporters of

*Enterobacteriaceae* (SPATEs) [628], such as *sat* (13%), *sigA* (11%) and *pic* (1%). In addition, eight isolates harboured known markers of Enteropathogenic *E. coli* (*eltAB* or *estA*).



*The tree was reconstructed with RAxML, using a general time-reversible nucleotide substitution model and 1,000 bootstrap replicates. The genome assembly of E. coli str. K12 substr. MG1655 was used as the reference, and the tree rooted using the genomic assembly of E. fergusonii as an outgroup. The sample names are indicated at the tip, with the respective Achtman sequence types (ST) indicated beside the sample names. The respective phylogroups the isolates belong to are indicated with colour codes as displayed in the legend. E. coli reference genome is denoted in black. Asterisks (*) are used to indicate novel STs. The predicted antimicrobial resistance genes and putative virulence factors for each isolate are displayed next to the tree, with the virulence genes clustered according to their function. Multiple copies of the same strain (ST) isolated from a single host are not shown. Instead, we have shown only one representative isolate from each strain. Virulence and resistance factors were not detected in the reference strain either.*

***Figure 5.2 :*** *A maximum-likelihood tree depicting phylogenetic relationships among study isolates.*

***Table 5.3:*** *Pairwise SNP distances between variants arising from within-host evolution.*

| Host | Sequence type (ST) | Colonies per ST | Pairwise SNP distances between multiple colonies of the same ST |
|---|---|---|---|
| H2 | 9274 | 5 | 0-9 |
| H9 | 2705 | 3 | 0-1 |
| H15 | 9277 | 3 | 0-1 |
| H15 | 747 | 2 | 3 |
| H18 | 38 | 2 | 3 |
| H21 | 58 | 2 | 0 |
| H22 | 316 | 4 | 0-3 |
| H25 | 181 | 4 | 1-5 |
| H28 | 469 | 4 | 0-3 |
| H32 | 101 | 3 | 1-9 |
| H34 | 603 | 3 | 2-8 |
| H36 | 59 | 4 | 0-14 |
| H37 | 5148 | 5 | 2-18 |
| H38 | 394 | 4 | 1-3 |
| H39 | 452 | 5 | 0-2 |
| H41 | 43 | 4 | 0-1 |
| H48 | 485 | 4 | 1-9 |
| H50 | 410 | 4 | 0 |

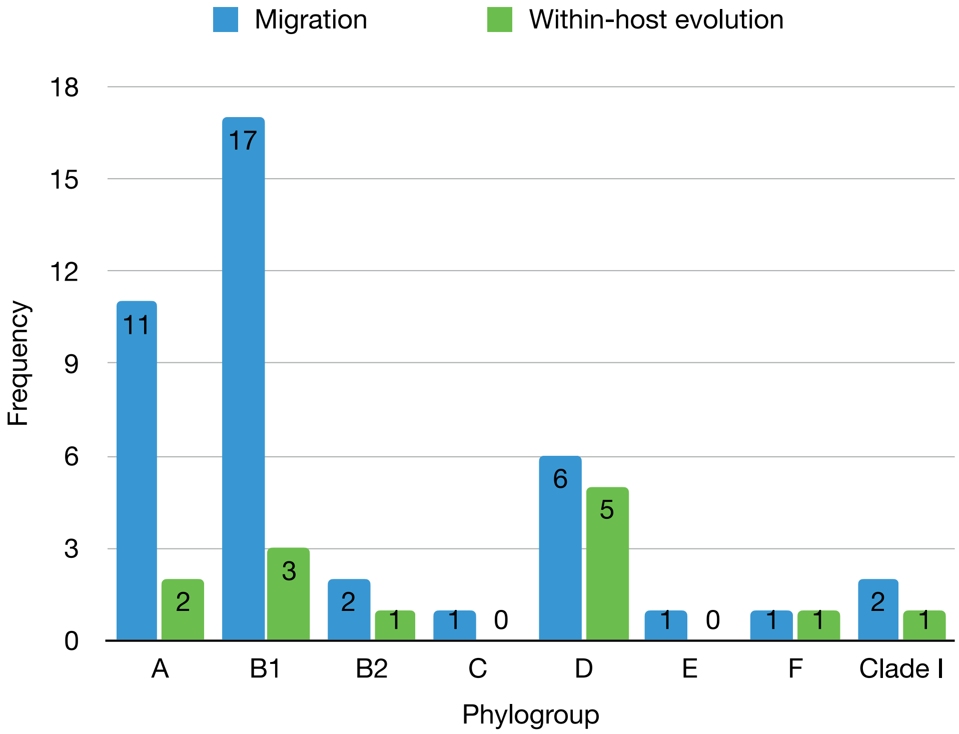


***Figure 5.3:*** *The distribution of variants inferred to have arisen from immigration events compared to those generated by within-host evolution by phylogroup.*

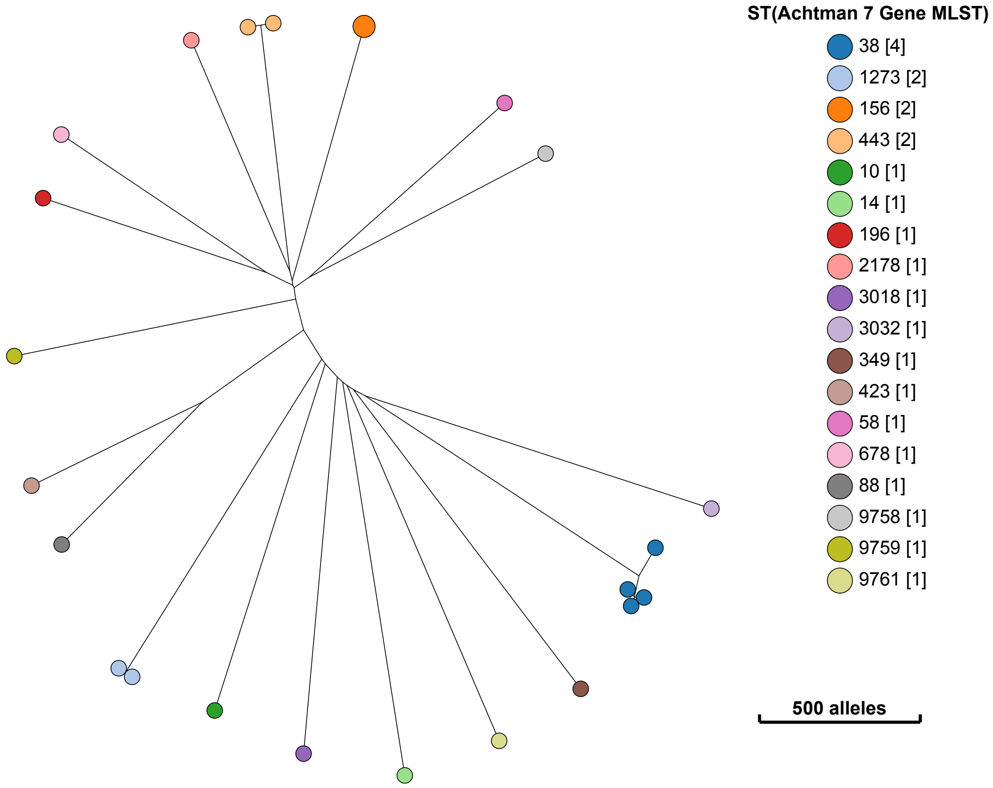
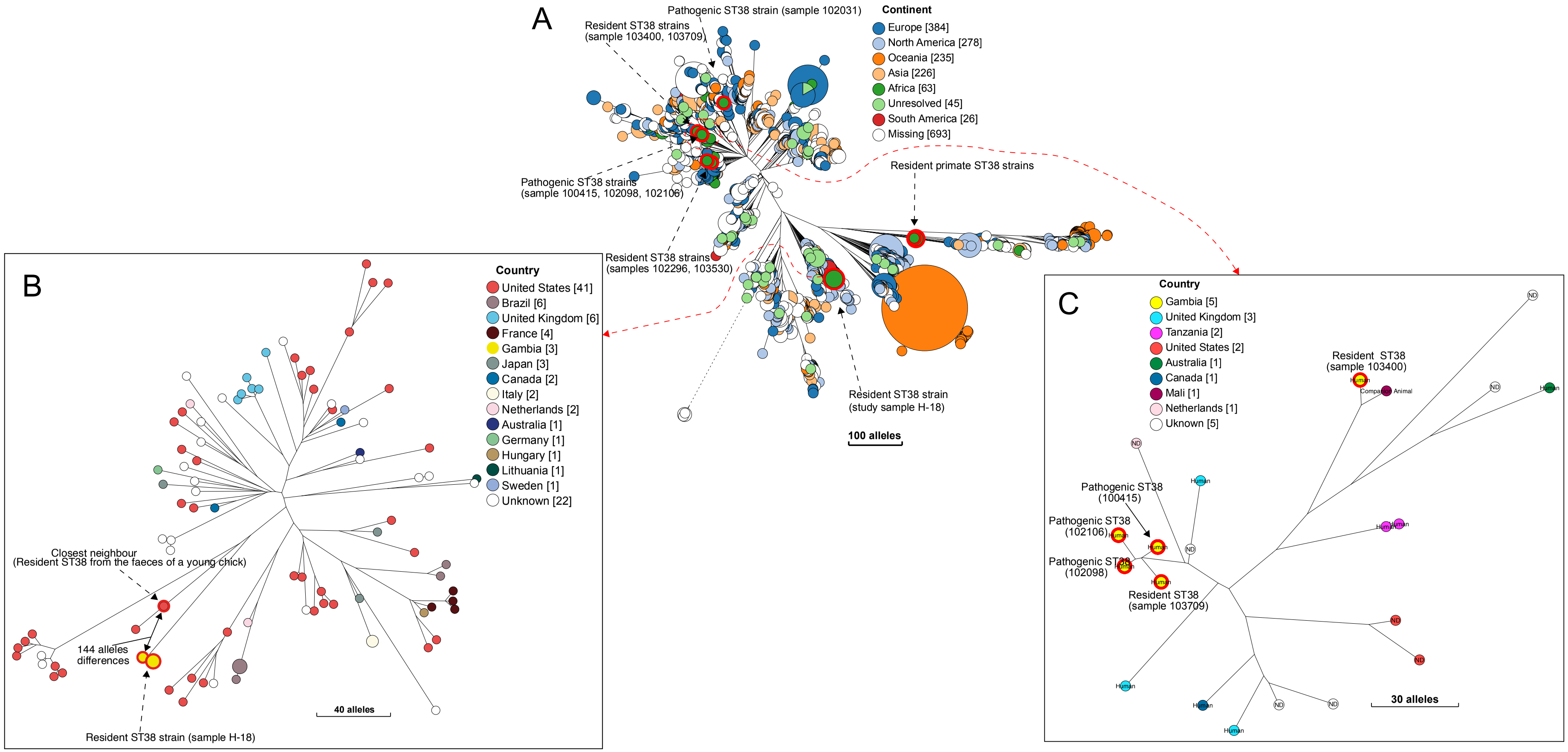


***Figure 5.4:*** *A Neighbour-joining phylogenetic tree depicting the genetic relationships among twenty-four strains isolated from diarrhoeal cases in the GEMS study.*

The Sequence types identified in these isolates are shown in the legend, with the genome count displayed in square brackets next to the respective sequence types. Three STs (ST38, ST58 and ST10) overlapped with what I found among commensal strains from this study (see Figure 5.5 and 5.6).

***Table 5.4:*** *Closest relatives to the study isolates.*

| Sample ID | 7-gene ST | Neighbour host | Neighbour status | Neighbour's country of isolation | Allelic distance |
|---|---|---|---|---|---|
| H-32_5 | 10 | Human | Unknown | UK | 18 |
| H-36_1 | 59 | Human | Unknown | UK | 18 |
| H-39_1 | 452 | Human | Commensal | UK | 26 |
| H-9_1 | 2705 | Livestock | Commensal | China | 29 |
| H-18_3 | 9274 | Human | Commensal | Unknown | 34 |
| H-2_1 | 9274 | Human | Commensal | Unknown | 34 |
| H-22_1 | 316 | Human | Commensal | UK | 35 |
| H-38_1 | 394 | Human | Pathogen (cystitis) | US | 39 |
| H-25_4 | 337 | Human | Unknown | Mali | 43 |
| H-37_1 | 5148 | Human | Pathogen (diarrhoea) | Ecuador | 43 |
| H-26_1 | 641 | Livestock | Commensal | US | 46 |
| H-26_5 | 398 | Poultry | Commensal | Kenya | 47 |
| H-48_2 | 485 | Human | Commensal | Tanzania | 57 |
| H-15_1 | 9277 | Human | Commensal | Zambia | 68 |
| H-15_2 | 747 | Human | Commensal | Egypt | 72 |
| H-28_1 | 469 | Human | Commensal | Kenya | 77 |
| H-21_2 | 1204 | Avian | Commensal | Kenya | 81 |
| H-34_2 | 10 | Livestock | Commensal | UK | 83 |
| H-38_2 | 58 | Human | Pathogen (bloodstream infection) | Australia | 87 |
| H-34_4 | 1727 | Unknown | Unknown | Unknown | 89 |
| H-35_1 | 226 | Human | Commensal | China | 93 |
| H-21_1 | 58 | Unknown | Unknown | Unknown | 98 |
| H-21_4 | 540 | Human | Unknown | Belgium | 100 |
| H-32_2 | 2175 | Livestock | Commensal | UK | 100 |
| H-26_2 | 10 | Livestock | Commensal | US | 111 |
| H-32_1 | 101 | Unknown | Unknown | Unknown | 111 |
| H-50_2 | 515 | Environment | Commensal | Canada | 117 |
| H-41_1 | 43 | Unknown | Unknown | Unknown | 120 |
| H-26_4 | 2741 | Human | Commensal | Germany | 126 |
| H-50_1 | 410 | Livestock | Commensal | US | 140 |
| H-18_1 | 38 | Poultry | Commensal | US | 144 |
| H-21_5 | 223 | Unknown | Unknown | Unknown | 145 |
| H-40_1 | 155 | Unknown | Unknown | US | 146 |
| H-41_2 | 9283 | Environment | Commensal | US | 191 |
| H-36_4 | 9278 | Avian | Commensal | Kenya | 208 |
| H-9_3 | 2914 | Canine | Commensal | UK | 272 |
| H-9_5 | 29 | Unknown | Unknown | Unknown | 288 |
| H-34_1 | 603 | Laboratory |  | UK | 325 |
| H-55_1 | 9279 | Environment | Commensal | Unknown | 333 |
| H-18_2 | 9281 | Unknown | Unknown | France | 430 |
| H-25_1 | 181 | Human | Commensal | Tanzania | 607 |



*A: A NINJA neighbour-joining tree showing the population structure of E. coli ST38, drawn using the genomes found in the core-genome MLST hierarchical cluster at HC1100, which corresponds to ST38 clonal complex. The size of the nodes represents the number of isolates per clade. The geographical locations where isolates were recovered are displayed in the legend; with the genome counts shown in square brackets. The study resident ST38 strains and the pathogenic ST38 strains recovered from GEMS cases are highlighted with red circles around the nodes.* ***B:*** *The closest neighbour to a pathogenic strain reported in GEMS [reference 30] is shown to be a commensal isolate recovered from a healthy individual. The size of the nodes represents the number of isolates per clade. The geographical locations where isolates were recovered are displayed in the legend; with the genome counts shown in square brackets. Red circles around the nodes are used to highlight the study resident ST38 strains and the pathogenic ST38 strains recovered from GEMS cases within this cluster.* ***C:*** *The closest relatives to the commensal ST38 strain recovered from this study is shown (red highlights), with the number of core-genome MLST alleles separating the two genomes displayed. The geographical locations where isolates were recovered are displayed in the legend; with the genome counts shown in square brackets, with the size of the nodes depicting the number of isolates per clade.*

***Figure 5.5:*** *Population structure of ST38.*

Several strains (across all phylogroups) also harboured virulence genes associated with intestinal or extraintestinal disease in humans, including adhesins, invasins, toxins and iron-acquisition genes such as *fyuA*, several *fim* and *pap genes*, *iroN*, *irp1*,2, *ibeA* and *aslA*. I did not detect any of the well-known markers of EPEC (*eae, bfpA, stx1*, or *stx2*) (Figure 5.2, File S3 in [626]). The prevalence of some virulence factors involved in invasion/evasion, iron uptake, adherence and secretion systems appeared to be more or less likely to occur in one or a few phylotypes (p≤0.05) as follows (File S9 in [626]). The iron acquisition genes *chuA, S-Y* and *shuA, S, T, Y* were found to be present in all cases for phylogroup D (n=5) and absent in virtually all cases for phylogroups A (n=13) and B1 (n=16).

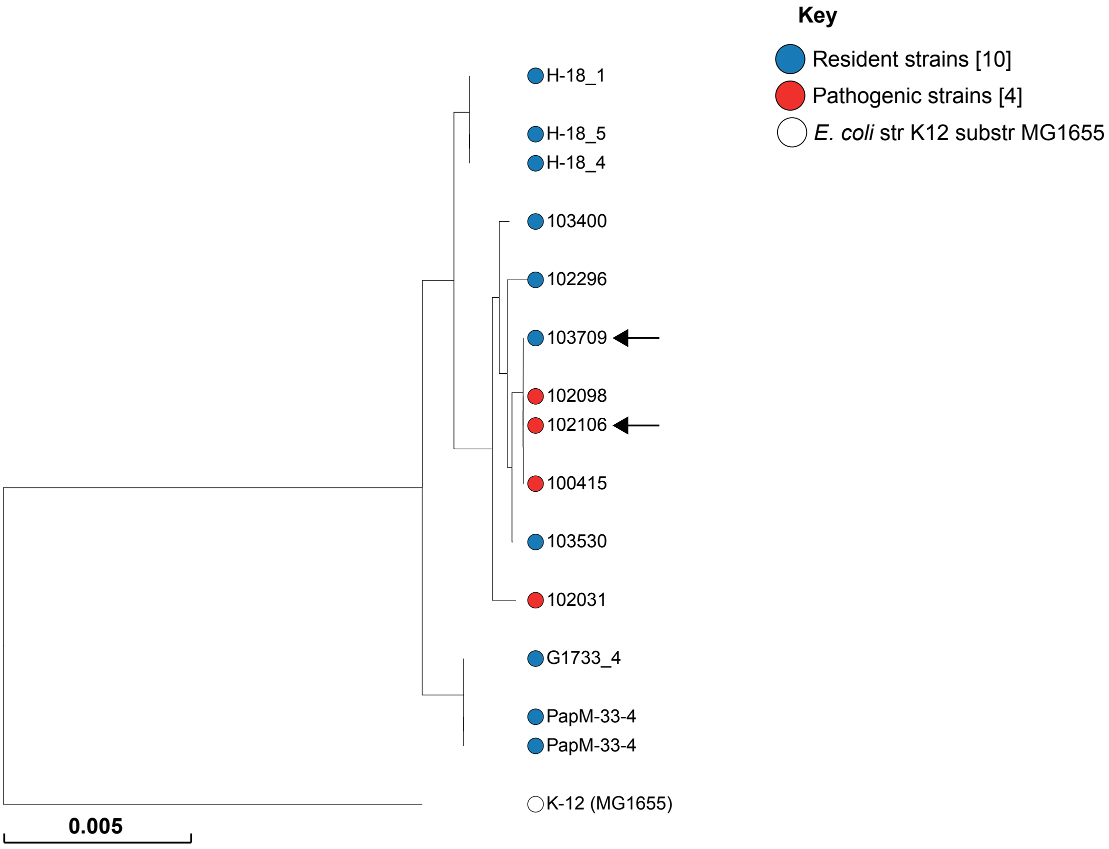
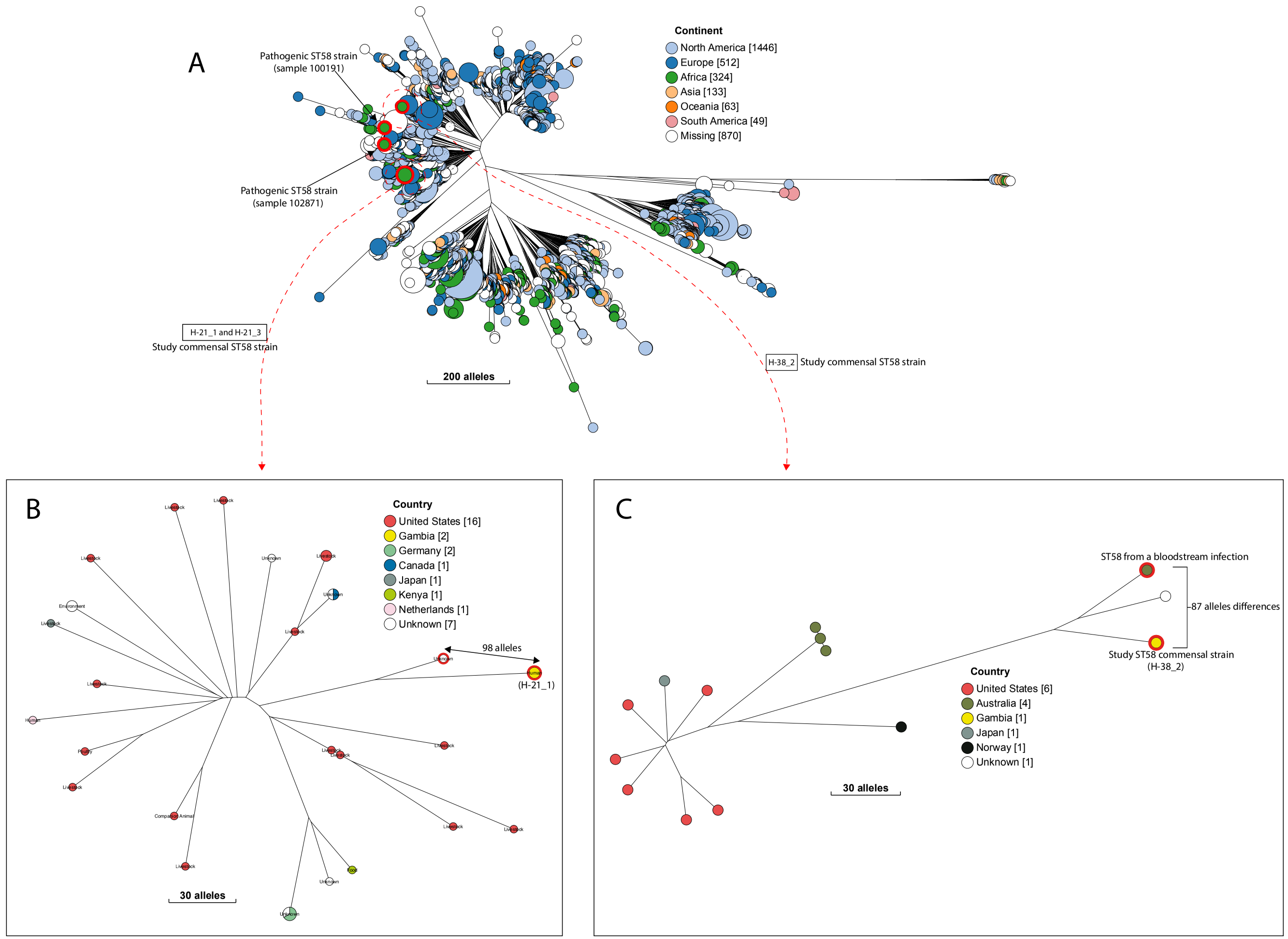


***Figure 5.6:*** *A maximum -likelihood Phylogenetic tree reconstructed using the genomes found in the Cluster 5.5C.*

A maximum likelihood phylogenetic tree reconstructed using the genomes found in cluster 5C in Figure 5.5 above, comprising both pathogenic and commensal ST38 strains, depicting the genetic relationships between strain 100415 (pathogenic) and 103709 (commensal) (highlighted by the arrows). The nodes are coloured to depict the status of the strains as pathogenic (red) or commensal (blue).

On the other hand, *iutA* and *iucA-D* were observed in the two cases from phylogroup B2 and absent from all samples from phylogroup D (n=5). The invasion/evasion genes *kpsD, M, T* and *aslA* were found to be present in almost all cases for phylogroups D (n=5), B2 (n=2) and Clade I (n=2) and absent in B1 (n=16). The secretion system gene cluster *espB, D, G, K-N, R, W-Y* was observed in all cases except the two belonging to phylogenetic group B2. The protease gene *sigA* was absent from most samples, except two samples from phylotype B2. The adherence gene *fdeC* was observed in all cases for phylotype D (n=5) and most for B1 (n=16). More than half of the isolates encoded resistance to three or more clinically relevant classes of antibiotics such as aminoglycosides, penicillins, trimethoprim, sulphonamides and tetracyclines (Figure 5.8). The most common resistance gene network was -*aph(6)-Id_1-sul2* (41% of the isolates), followed by *aph(3'')-Ib_5-sul2* (27%) and *bla-*$_{TEM}$*-aph(3'')-Ib_5* (24%) (Figure 5.9). Most isolates (67%) harboured two or more plasmid types (Figure 5.10).



*The population structure of ST58.*

*A: A NINJA neighbour-joining tree depicting the population structure of E. coli ST58, drawn using the genomes found that clustered together in the same in HC1100 hierarchical cluster in the core-genome MLST scheme. Commensal ST58 strains from this study and Gambian pathogenic ST58 isolates from GEMS are highlighted in red. The geographical locations where isolates were recovered are displayed in the legend; with the genome counts shown in square brackets. The size of the nodes represents the number of isolates per clade.* ***B*** *and* ***C****: The closest relatives to the study ST58 strains are shown.*

*Geographical locations where isolates were recovered are displayed in the legend, with the genome counts displayed in square brackets. The red highlights around the nodes depict the study commensal ST58 strains and their closest neighbours. The size of the nodes represents the number of isolates per clade, and the geographical locations where isolates were recovered are displayed in the legend; with the genome counts shown in square brackets.*

***Figure 5.7:*** *The population structure of ST58.*

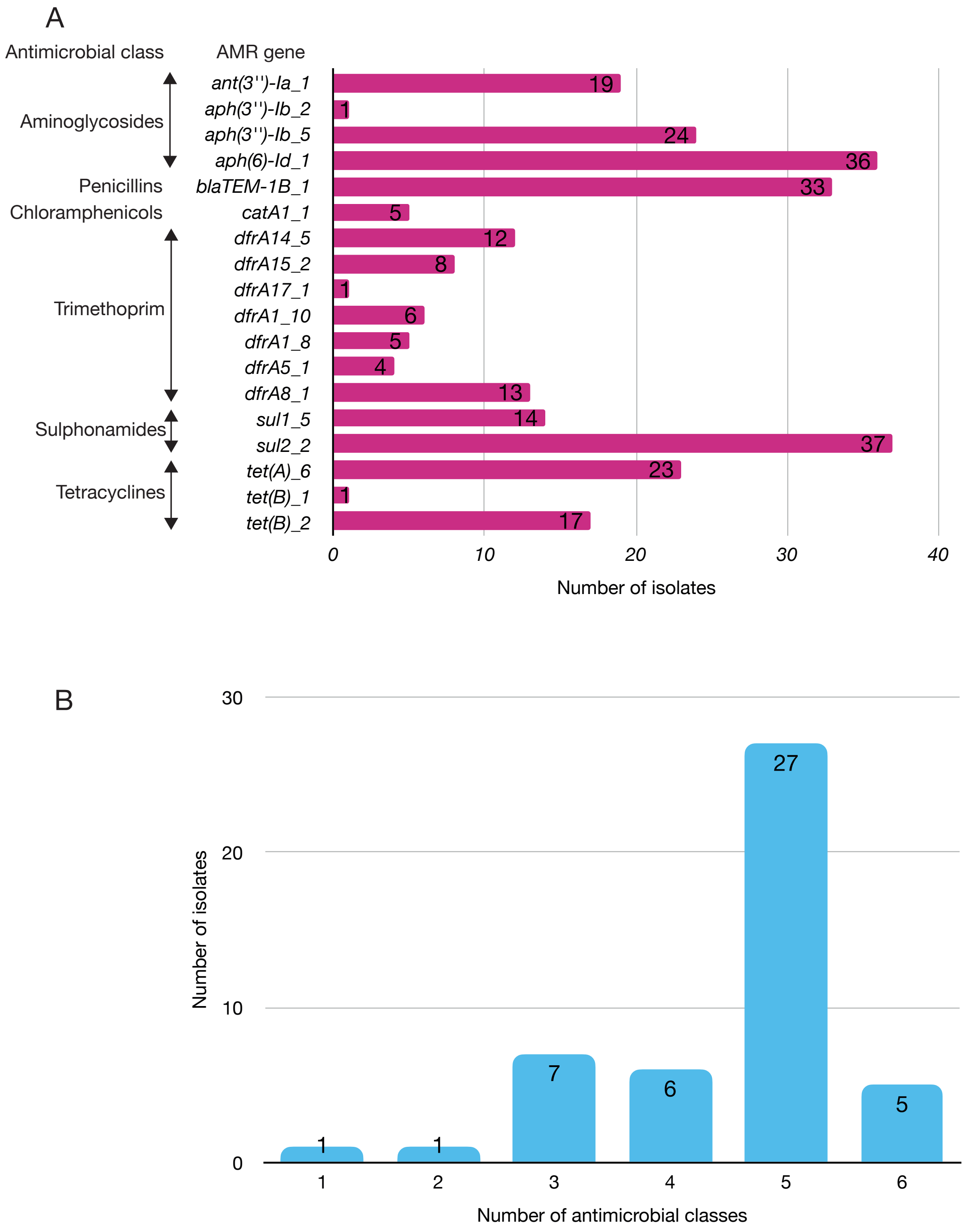


***Figure 5.8:*** *The prevalence of antimicrobial-associated genes detected in the isolates.*
**(A)** The y-axis shows the prevalence of the detected AMR-associated genes in the study isolates, grouped by antimicrobial class. **(B)** A histogram depicting the number of antimicrobial classes to which resistance genes were detected in the corresponding strains.

Of the 24 plasmid types detected, IncFIB was the most common (41%), followed by col156 (19%) and IncI_1-Alpha (15%). Nearly three-quarters of the multi-drug resistant isolates carried IncFIB (AP001918) plasmids (~50kb), suggesting that these large plasmids disseminate resistance genes within my study population.

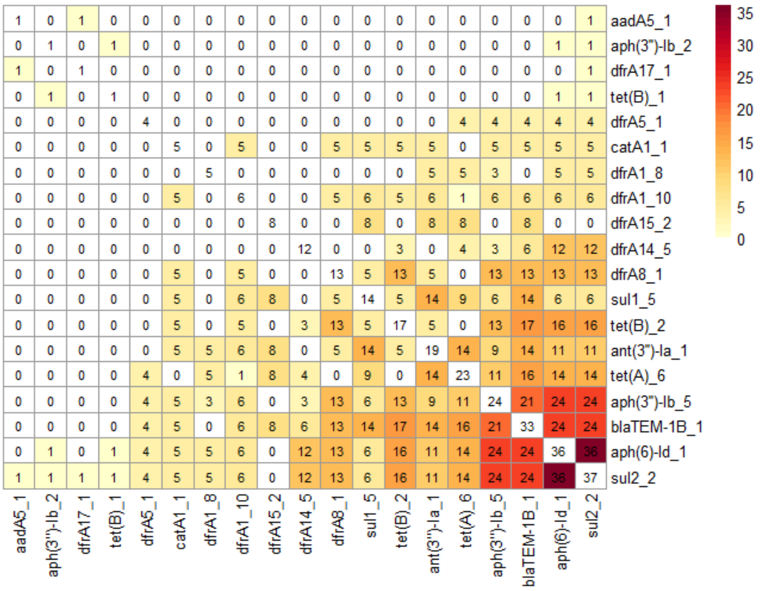


***Figure 5.9:*** *A co-occurrence matrix of acquired antimicrobial resistance genes detected in the study isolates.*
The diagonal values show how many isolates each individual gene was found in, while the intersections between the columns represent the number of isolates in which the corresponding antimicrobial resistance genes co-occurred.

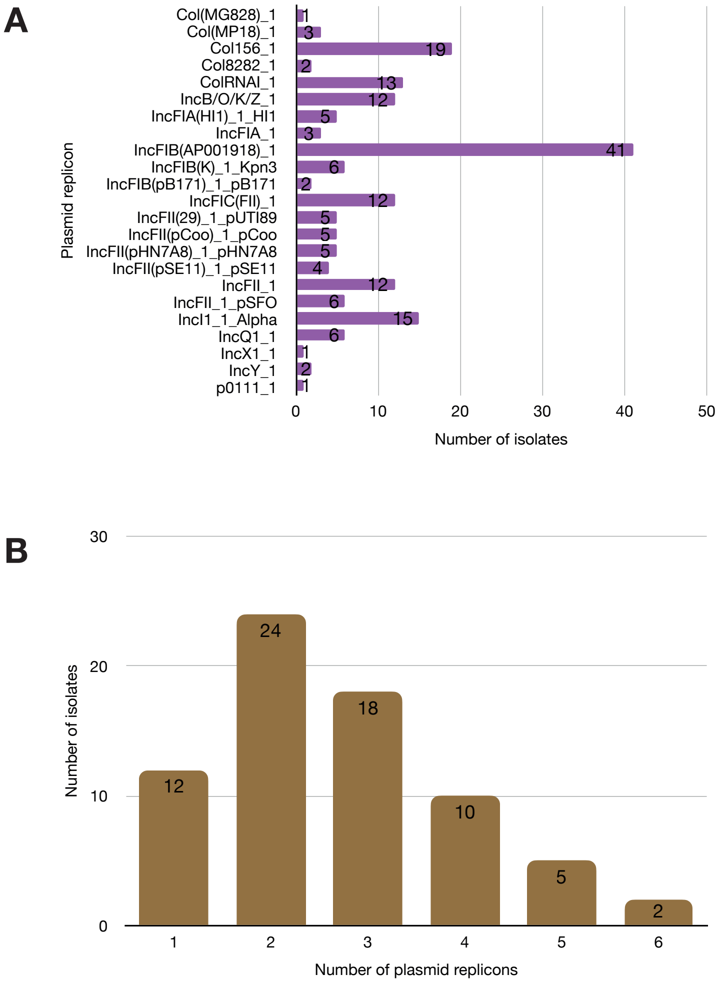


***Figure 5.10:*** *Prevalence of plasmid replicons among the study isolates.*
**(A)** Plasmid replicons detected in the study isolates. **(B)** A histogram depicting the number of plasmids co-harboured in a single strain.

## 5.4 Discussion

This study provides an overview of the within-host genomic diversity of *E. coli* in healthy children from a rural setting in the Gambia, West Africa. Surprisingly, I was able to recover *E. coli* from only 34% of stools which had previously tested positive for *E. coli* in the original study. This low rate of recovery may reflect some hard-to-identify effect of long-term storage (nine to thirteen years) or the way the samples were handled, even though they were kept frozen and thawed only just before culture.

Several studies have shown that sampling a single colony is insufficient to capture *E. coli* strain diversity in stools [413-415]. Lidin-Janson *et al.* [506] claim that sampling five colonies provides a >99% chance of recovering dominant genotypes from single stool specimens, while Schlager *et al.* [505] calculate that sampling twenty-eight colonies provides a >90% chance of recovering minor genotypes. My results confirm the importance of multiple-colony picks in faecal surveillance studies, as over half (57%) of my strains would have been missed by picking a single colony.

I recovered strains encompassing all eight major phylotypes of *E. coli,* however, the majority fell into the A and B1 phylogenetic groups, in line with previous reports that these phylogroups dominate in stools from people in low- and middle-income countries [629, 630]. Although not fully understood, there appear to be host-related factors that influence the composition of *E. coli* phylogroups in human hosts. For example, the establishment of strains belonging to phylogroups E or F seems to favour subsequent colonisation by other phylotypes, compared to the establishment of phylogroup B2 strains, which tend to limit the heterogeneity within individual hosts [631]. Geographical differences have also been reported, with phylogroups A and B1 frequently dominating the stools of people living in developing countries [629, 630]. Conversely, phylogroup B2 and D strains appear to be pervasive among people living in developed countries [632, 633]. These locale-specific patterns in the distribution of *E. coli* phylotypes have been attributed to differences in diet and climate [629, 630].

The prevalence of putative virulence genes in most of my isolates highlights the pathogenic potential of commensal intestinal strains—regardless of their phylogroup—should they gain access to the appropriate tissues, for example, the urinary tract. My results complement previous studies reporting genomic similarities between faecal *E. coli* isolates and those recovered from UTIs [622, 634].

I found that within-host evolution plays a minor role in the generation of diversity in my study population. This might be due to the low prevalence of B2 strains, which are thought to inhibit the establishment of strains from other phylogroups, as discussed above [631]; or it may indicate that members of phylogroups A and B1 might favour a more heterogeneous composition of *E. coli* phylotypes in stools of healthy individuals. However, this remains to be properly investigated, as I did not find statistical evidence that the distribution of variants (independent migration versus within-host evolution) was influenced by phylogroup. My findings are in line with that of Dixit *et al.* [413], who reported that 83% of diversity originates from immigration events and with epidemiological data suggesting that the recurrent immigration events account for the high faecal diversity of *E. coli* in the tropics [2].

The estimated mutation rate for *E. coli* lineages is around one SNP per genome per year [407], so that two genomes with a most recent common ancestor in the last five years would be expected to be around ten SNPs apart. However, in two subjects, pairwise distances between genomes from the same ST (ST59 and ST5148) were large enough (14 and 18 respectively) to suggest that they might have arisen from independent immigration events. It remains possible that the mutation rate was higher than expected in these lineages, although I found no evidence of damage to DNA repair genes. Alternatively, the observed mutations may have arisen from within-strain recombination events. Co-colonising variants belonging to the same ST tended to share an identical virulence, AMR and plasmid profile, signalling similarities in their accessory gene content.

The sources of novel variation that account for within-host diversity include point mutation and small insertions or deletions (indels), large indels and the loss or acquisition of mobile genetic elements. Among the variants inferred to have been derived from within-host evolution, I observed dominance of mutations that were predicted to result in changes in protein function, in the form of missense mutations and non-sense mutations (leading to a premature stop codon). Although the mutations appeared to be heterogeneously distributed, a higher number was observed in genes associated with metabolism. These appeared to be under positive selection, although it remains to be seen if these changes confer any effects on fitness. It will be desirable to investigate this in future studies. Due to the cross-sectional nature of my sampling, I was unable to analyse the dynamics of strain gain or loss and variation in gene content over time. Homologous recombination has also been noted to contribute to the generation of diversity [482, 635], however, I detected and remove recombinant regions prior to phylogenetic reconstruction and thus focused on my analysis on SNPs.

More than half of my study isolates encode resistance to three or more classes of antimicrobials echoing the high rate of MDR (65%; confirmed by phenotypic testing) in the GEMS study; although none of the ESBL genes commonly associated with multi-drug resistant clones (such as bla$_{OXA}$ and bla$_{CTXM)}$ were found among my study isolates. IncFIB (AP001918) was the most common plasmid Inc type from my study, in line with the observation that IncF plasmids are frequently associated with the dissemination of resistance [560]. However, a limitation of my study is that I did not perform phenotypic antimicrobial resistance testing, although Doyle *et al.* [611] reported that only a small proportion of genotypic AMR predictions are discordant with phenotypic results.

Comparative analyses confirm the heterogeneous origins of the strains reported here, documenting links to other human commensal strains or isolates sourced from livestock or the environment. This is not surprising, as almost all study participants reported that animals are kept in their homes and children in rural Gambia are often left to play on the ground, close to domestic animals such as pets and poultry [597]. My results show that the commensal *E. coli* population in the gut of healthy children in rural Gambia is richly diverse, with the independent immigration and establishment of strains contributing to the bulk of the observed diversity.

### 5.4.1 Limitations

An obvious limitation to my study is the low recovery of *E. coli* from frozen stools—which potentially implies I may have underestimated the extent of genetic diversity present within my study population. A further limitation is the inability to perform phenotypic antimicrobial resistance tests as mentioned earlier, due to Covid-19 restrictions. Also, it is unclear whether the population diversity I observed reflects what pertains in adult populations; however, I note previous studies suggesting that the intestinal microbiota in children aged three to five years old is reasonably stable and persists through adulthood [636]. Moreover, the nature of my study meant that I did not have sufficient data to explore the spatial diversity of *E. coli* within my study population.

Although solely observational, my study paves the way for future studies aimed at a mechanistic understanding of the factors driving the diversification of *E. coli* in the human gut and what it takes to make a strain of *E. coli* successful in this habitat. In addition, this work has added significantly to the number of commensal *E. coli* genomes, which are underrepresented in public repositories.

# 6 CHAPTER SIX: CONCLUSIONS

This work investigated the genomic within-host diversity of *E. coli* in three interconnected hosts in the Gambia: non-human primates dispersed across the Gambia, with varying degrees of contacts with humans, backyard poultry sharing proximity with humans, particularly children and healthy children from a rural setting. All three hosts exist in a life-long mutualistic relationship with a familiar ally and foe: *E. coli*, the all-rounder.

As expected, humans and non-human primates have specific *E. coli* lineages in common (ST1204, ST1727, ST8826, ST226 and ST38) (Figure 6.1), probably representing resident strains that may have existed in the guts of the most recent common ancestors of humans and non-human primates.

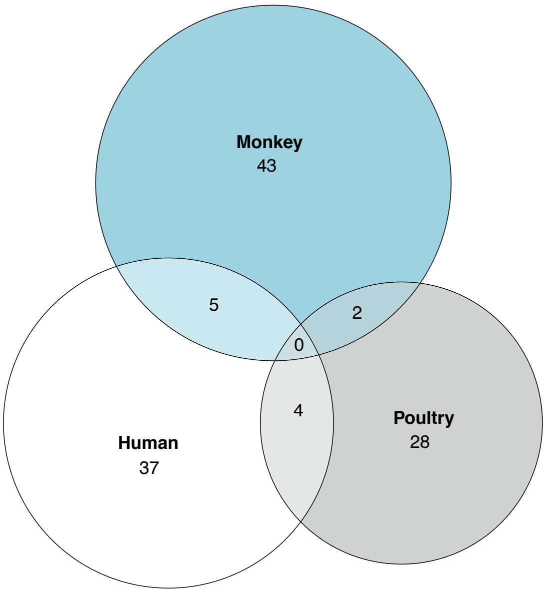


***Figure 6.1:*** *Overlap of STs among humans, non-human primates and poultry.*
The numbers in each circle represent the total number of STs that were observed in the respective populations. The areas of overlap highlight the number of shared STs between the respective host species.

Of note among these is the ST38 lineage, which is linked with the dissemination of carbapenemase resistance, particularly, the OXA-48 carbapenemase gene [637]. Although the human ST38 strain did not appear to harbour any antibiotic resistance genes, the simian ST38 strain encoded a beta-lactamase gene. It was interesting that hierarchical clustering found that the nearest relative to the human commensal ST38 strain was a pathogenic strain that caused diarrhoea in the same locality where the resident strain was sourced from; the differences in core-genomic loci indicating that both isolates originate from a common source. As discussed, comparative analyses with all publicly available strains worldwide showed that several notable lineages that have long been associated with extraintestinal infections or antimicrobial resistance in humans (such as ST73 and ST681) commonly colonise non-human primates innocuously. The hyperconnectivity of strains, irrespective of the species barrier was further highlighted by finding within ST349 both a simian isolate and a human bloodstream isolate from Canada.

Similarly, four lineages were found to share overlap between humans and backyard poultry within this study: ST155, ST58, ST540 and ST337. Also, two STs were detected in common with both poultry and monkeys (ST212 and ST196). The results show that there are potential interactions between strains from poultry and livestock within the context of sub-Saharan Africa and that, worryingly, multi-drug resistance genotypes are widespread in poultry. I also reported close relationships between human commensal strains and isolates from livestock.

Overall, the results presented in this thesis highlight the potential role of the resident *E. coli* population in the evolution of disease and antimicrobial resistance, while blurring the lines of demarcation between harmless commensals and pathogens. Colonisation with multiple strains in all three host species studied exemplifies the role of the gut *E. coli* population in facilitating the exchange of potential virulence and AMR determinants, thus influencing the inter- and intra-strain dissemination of these traits between within and between host species. The existence of multi-drug resistance genotypes in backyard poultry and human habituated non-human primates is of concern, given the proximity to humans and the opportunities for intra- and inter-specific exchange of antimicrobial resistance determinants between these host species via horizontal gene transfer. As has been stated: "Any *E. coli* can probably cause invasive disease given the right opportunities"[18]. My results confirm this statement, as I have shown that the commensal *E. coli* population possess the armamentarium to cause extraintestinal infection: factors which coincidentally promote intestinal fitness facilitate survival and competitiveness within the gastrointestinal niche.

## 6.1 Novel insights

This study has unearthed previously unknown diversity within the resident *E. coli* population, particularly in non-human primates and poultry and added significantly to our appreciation of the diversity of *E. coli* in the healthy vertebrate gut. Before my work, we had a broad understanding of the phylogroups of *E. coli* that prevail in the non-human primate gut and caused disease in captive or laboratory animals; however, the lineages existing in wild monkeys and how these interact genetically with those found in humans were largely unknown. Similarly, we appreciated the potential burden of antimicrobial resistance in backyard poultry from sub-

Saharan Africa. However, due to the lack of whole-genome sequence-based data, our understanding of the genomic diversity of *E. coli* in this population was limited. Moreover, we lacked data on the lineages of *E. coli* circulating in these backyard birds in the Gambia. Despite *E. coli* having an established role in diarrhoea among children from rural Gambia, it was unclear what role independent exposure to this organism played in the evolution of disease among this population before my study.

Several novel strains were discovered in this thesis, seven of which have been completely sequenced and deposited in public archives, adding significantly to the number of commensal genomes available for future comparative studies.

## 6.2 Prospects

Unsurprisingly, this thesis has raised some new questions while failing to answer some persistent ones. Questions that remain outstanding include:

- What factors govern co-existence or competition between the different *E. coli* strains co-colonising the vertebrate gut?
- Do they exploit different micro-geographical or nutrient niches in the gut?
- What is the importance, if any, of flagellar motility?
- How far do within-species antibacterial factors (e.g., colicins and type VI secretions systems) play a role in competitive exclusion?
- Can differences in the distribution of *E. coli* strains in vivo be accounted for by behaviours *in vitro* (e.g., competitive growth under different nutrient conditions)

Biolog assays (Biolog Inc., Hayward, California, USA) are one way to screen for the key metabolic differences between different strains that may account for their co-existence or competitive exclusion within the gut. Candidate strains could then be selected for onward in-vitro competition assays via individual and competitive fitness experiments and indirect antagonism assays as have been used by Durso *et al.* to assess fitness differences and competition between commensal strains and O157:H7 strains [638]. Applikon Biotechnology (https://www.applikon-biotechnology.com/en/products/cultivation-systems/micro-matrix/) have designed mini-bioreactors which can be used for multiple competition assays in parallel, using 24-well microtiter plates with built-in controls for pH, temperature and dissolved oxygen. Such experiments may shed light on the physiological and biochemical properties underlying the co-colonisation of different strains in the vertebrate gut.

It satisfies the curiosity to see some close genetic relationships between strains from non-human primates, backyard poultry, livestock and humans from the Gambia and sub-region. However, it remains unknown the direct contribution of human-associated animals to *E. coli*'s transmission. To address this would require future studies incorporating samples from whole households and animals living in proximity. The Gambia and Africa need such genomics-based studies to address the burden of infectious diseases confronting our societies.

Armed with the skills I have gained from this PhD, I aim to contribute to the body of research in this area in the medium to long-term, elucidating the genomic epidemiology of pathogens to address the global disease burden via dry and wet-lab approaches.

## REFERENCES

1. **Downie JA, Young JP**. Genome sequencing. The ABC of symbiosis. *Nature* 2001;412(6847):597-598.
2. **Tenaillon O, Skurnik D, Picard B, Denamur E**. The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol* 2010;8(3):207-217.
3. **Friedmann H**. Escherich and *Escherichia*. *EcoSal Plus* 2014. Reprinted from *Adv Appl Microbiol* 2014;60:133-196.
4. **Escherich T**. The intestinal bacteria of the neonate and breast-fed infant. 1884. *Rev Infect Dis* 1988;10(6):1220-1225.
5. **Shulman ST, Friedmann HC, Sims RH**. Theodor Escherich: the first pediatric infectious diseases physician? *Clin Infect Dis* 2007;45(8):1025-1029.
6. **Escherich T, Pfaundler M**. *Bacterium coli commune*. In: **Kolle W, Wassermann A** (eds). Handbuch der pathogenen mikroorganismen, vol 2. Verlag von Gustav Fischer, Jena 1903; pp. 334–474.
7. **Bettelheim KA**. Commemoration of the publication 100 years ago of the papers by Dr. Th. Escherich in which are described for the first time the organisms that bear his name. *Zentralbl Bakteriol Mikrobiol Hyg A* 1986;261(3):255-265.
8. **Escherich T**. Ueber die bacterien des milchkotes. *Æerztliches Intelligenz-Blatt, Münchner Medicinische Wochenschrift* 1888;32:243.
9. **Castellani A, Chalmers A**. **J**. Manual of tropical medicine (3rd ed.). London: Baillière, Tindall and Cox; 1919; p. 941.
10. **Judicial Commission of the International Committee on Bacterial Nomenclature**. Conservation of the family name *Enterobacteriaceae,* of the name of the type genus, and designation of the type species. Opinion No. 15. *Int Bull Bacteriol Nomencl Taxon* 1958;8:73–74.
11. **Skerman V, McGowan, V., Sneath, P**. Approved lists of bacterial names. *Int J Syst Bacteriol* 1980;30:225–420.
12. **Blount ZD**. The unexhausted potential of *E. coli*. *Elife* 2015;4.
13. **van Elsas JD, Semenov AV, Costa R, Trevors JT**. Survival of *Escherichia coli* in the environment: fundamental and public health aspects. *ISME J* 2011;5(2):173-183.
14. **Alm EW, Walk, ST and Gordon, DM**. The niche of *Escherichia coli.* In: **Walk ST and Feng, PCH** (eds). Population Genetics of Bacteria: A tribute to Thomas S. Whittam. ASM Press; Washington, DC. 2011; Chapter 6 pp. 67-89.
15. **Macfarlane GT, Macfarlane S**. Human colonic microbiota: ecology, physiology and metabolic potential of intestinal bacteria. *Scand J Gastroenterol Suppl* 1997;222:3-9.

16. **Leimbach A, Hacker J, Dobrindt U**. *E. coli* as an all-rounder: The thin line between commensalism and pathogenicity. In: **Dobrindt U, Hacker JH, Svanborg C** (eds). *Between Pathogenicity and Commensalism*. Springer Berlin Heidelberg; 2013. pp. 3-32.

17. **Kosek M, Bern C, Guerrant RL**. The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. *Bull World Health Organ* 2003;81(3):197-204.

18. **Orskov F, Orskov I**. *Escherichia coli* serotyping and disease in man and animals. *Can J Microbiol* 1992;38(7):699-704.

19. **Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI**. Host-bacterial mutualism in the human intestine. *Science* 2005;307(5717):1915-1920.

20. **Claesson MJ, O'Sullivan O, Wang Q, Nikkilä J, Marchesi JR *et al.*** Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLos One* 2009;4(8):e6669.

21. **Crick FH, Barnett L, Brenner S, Watts-Tobin RJ**. General nature of the genetic code for proteins. *Nature* 1961;192:1227-1232.

22. **Lehman IR, Bessman MJ, Simms ES, Kornberg A**. Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from *Escherichia coli*. *J Biol Chem* 1958;233(1):163-170.

23. **Stevens A**. Incorporation of the adenine ribonucleotide into RNA by cell fractions from *E. coli*. *Biochem Biophys Res Commun* 1*960*;3:92–96.

24. **Ellis EL, Delbrück M**. The growth of bacteriophage. *J Gen Physiol* 1939;22(3):365-384.

25. **Lwoff A**. Lysogeny. *Bacteriol Rev* 1953;17(4):269-337.

26. **Hu Y, Coates AR**. Transposon mutagenesis identifies genes which control antimicrobial drug tolerance in stationary-phase *Escherichia coli*. *FEMS Microbiol Lett* 2005;243(1):117-124.

27. **Linn S, Arber W**. Host specificity of DNA produced by *Escherichia coli*, X. *In vitro* restriction of phage fd replicative form. *Proc Natl Acad Sci USA* 1968;59(4):1300-1306.

28. **Meselson M, Yuan R**. DNA restriction enzyme from *E. coli*. *Nature* 1968;217(5134):1110-1114.

29. **Harshey RM, Matsuyama T**. Dimorphic transition in *Escherichia coli* and *Salmonella typhimurium*: surface-induced differentiation into hyperflagellate swarmer cells. *Proc Natl Acad Sci USA* 1994;91(18):8631-8635.

30. **Inoue T, Shingaki R, Hirose S, Waki K, Mori H *et al.*** Genome-wide screening of genes required for swarming motility in *Escherichia coli* K-12. *J Bacteriol* 2007;189(3):950-957.

31. **Jacob F, Perrin D, Sanchez C, Monod J**. [Operon: a group of genes with the expression coordinated by an operator]. *C R Hebd Seances Acad Sci* 1960;250:1727-1729.

32. **Capaldi RA, Schulenberg B, Murray J, Aggeler R**. Cross-linking and electron microscopy studies of the structure and functioning of the *Escherichia coli* ATP synthase. *J Exp Biol* 2000;203(Pt 1):29-33.

33. **Huang CJ, Lin H, Yang X**. Industrial production of recombinant therapeutics in *Escherichia coli* and its recent advancements. *J Ind Microbiol Biotechnol* 2012;39(3):383-399.

34. **Baeshen MN, Al-Hejin AM, Bora RS, Ahmed MM, Ramadan HA *et al.*** Production of biopharmaceuticals in *E. coli*: Current scenario and future perspectives. *J Microbiol Biotechnol* 2015;25(7):953-962.

35. **Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM**. The dynamics of molecular evolution over 60,000 generations. *Nature* 2017;551(7678):45-50.

36. **Luria SE, Delbrück M**. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 1943;28(6):491-511.

37. **Lederberg J, Lederberg EM**. Replica plating and indirect selection of bacterial mutants. *J Bacteriol* 1952;63(3):399-406.

38. **Wiser MJ, Ribeck N, Lenski RE**. Long-term dynamics of adaptation in asexual populations. *Science* 2013;342(6164):1364-1367.

39. **Cooper TF**. Recombination speeds adaptation by reducing competition between beneficial mutations in populations of *Escherichia coli*. *PLoS Biol* 2007;5(9):e225.

40. **Lenski RE**. Experimental studies of pleiotropy and epistasis in *Escherichia coli*. I. Variation in competitive fitness among mutants resistant to virus T4. *Evolution* 1988;42(3):425-432.

41. **Blount ZD, Barrick JE, Davidson CJ, Lenski RE**. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* 2012;489(7417):513-518.

42. **Kim B, Park H, Na D, Lee SY**. Metabolic engineering of *Escherichia coli* for the production of phenol from glucose. *Biotechnol J* 2014;9(5):621-629.

43. **Kaup B, Bringer-Meyer S, Sahm H**. Metabolic engineering of *Escherichia coli*: construction of an efficient biocatalyst for D-mannitol formation in a whole-cell biotransformation. *Appl Microbiol Biotechnol* 2004;64(3):333-339.

44. **Hildebrand A, Schlacta T, Warmack R, Kasuga T, Fan Z**. Engineering *Escherichia coli* for improved ethanol production from gluconate. *J Biotechnol* 2013;168(1):101-106.

45. **Liu T, Khosla C**. Genetic engineering of *Escherichia coli* for biofuel production. *Annu Rev Genet* 2010;44:53-69.

46. **Kaper JB, Nataro JP, Mobley HL**. Pathogenic *Escherichia coli*. *Nat Rev Microbiol* 2004;2(2):123-140.

47. **Croxen MA, Finlay BB**. Molecular mechanisms of *Escherichia coli* pathogenicity. *Nat Rev Microbiol* 2010;8(1):26-38.

48. **Denamur E, Clermont O, Bonacorsi S, Gordon D**. The population genetics of pathogenic *Escherichia coli*. *Nat Rev Microbiol* 2020.

49. **Pitout JD**. Extraintestinal pathogenic *Escherichia coli*: A combination of virulence with antibiotic resistance. *Front Microbiol* 2012;3:9.

50. **Escobar-Páramo P, Giudicelli C, Parsot C, Denamur E**. The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised. *J Mol Evol* 2003;57(2):140-148.

51. **Pupo GM, Lan R, Reeves PR**. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci USA* 2000;97(19):10567-10572.

52. **Lan R, Alles MC, Donohoe K, Martinez MB, Reeves PR**. Molecular evolutionary relationships of enteroinvasive *Escherichia coli* and *Shigella* spp. *Infect Immun* 2004;72(9):5080-5088.

53. **The HC, Thanh DP, Holt KE, Thomson NR, Baker S**. The genomic signatures of *Shigella* evolution, adaptation and geographical spread. *Nat Rev Microbiol* 2016;14(4):235-250.

54. **Chaudhuri RR, Henderson IR**. The evolution of the *Escherichia coli* phylogeny. *Infect Genet Evol* 2012;12(2):214-226.

55. **Robins-Browne RM, Holt KE, Ingle DJ, Hocking DM, Yang J *et al.*** Are *Escherichia coli* pathotypes still relevant in the era of whole-genome sequencing? *Frontiers in Cellular and Infection Microbiology* 2016;6:141.

56. **Qin J, Cui Y, Zhao X, Rohde H, Liang T *et al.*** Identification of the Shiga toxin-producing *Escherichia coli* O104:H4 strain responsible for a food poisoning outbreak in Germany by PCR. *J Clin Microbiol* 2011;49(9):3439-3440.

57. **Sheldon IM, Rycroft AN, Dogan B, Craven M, Bromfield JJ *et al.*** Specific strains of *Escherichia coli* are pathogenic for the endometrium of cattle and cause pelvic inflammatory disease in cattle and mice. *PLoS One* 2010;5(2):e9192.

58. **Mainil J**. *Escherichia coli* virulence factors. *Vet Immunol Immunopathol* 2013;152(1):2-12.

59. **Nicholson BA, West AC, Mangiamele P, Barbieri N, Wannemuehler Y *et al.*** Genetic characterization of ExPEC-like virulence plasmids among a subset of NMEC. *PLoS One* 2016;11(1):e0147757.

60. **Blum SE, Heller ED, Sela S, Elad D, Edery N *et al.*** Genomic and phenomic study of mammary pathogenic *Escherichia coli*. *PLoS One* 2015;10(9):e0136387.

61. **Dogan B, Rishniw M, Bruant G, Harel J, Schukken YH *et al.*** Phylogroup and *lpfA* influence epithelial invasion by mastitis associated *Escherichia coli*. *Veterinary Microbiology* 2012;159(1-2):163-170.

62. **Frank C, Werber D, Cramer JP, Askar M, Faber M *et al.*** Epidemic profile of Shiga-toxin-producing *Escherichia coli* O104:H4 outbreak in Germany. *N Engl J Med* 2011;365(19):1771-1780.

63. **Karch H, Denamur E, Dobrindt U, Finlay BB, Hengge R *et al.*** The enemy within us: lessons from the 2011 European *Escherichia coli* O104:H4 outbreak. *EMBO Mol Med* 2012;4(9):841-848.

64. **Soysal N, Mariani-Kurkdjian P, Smail Y, Liguori S, Gouali M *et al.*** Enterohemorrhagic *Escherichia coli* hybrid pathotype O80:H2 as a new therapeutic challenge. *Emerg Infect Dis* 2016;22(9):1604-1612.

65. **Mariani-Kurkdjian P, Lemaître C, Bidet P, Perez D, Boggini L *et al.*** Haemolytic-uraemic syndrome with bacteraemia caused by a new hybrid *Escherichia coli* pathotype. *New Microbes New Infect* 2014;2(4):127-131.

66. **Thiry D, Saulmont M, Takaki S, De Rauw K, Duprez JN *et al.*** Enteropathogenic *Escherichia coli* O80:H2 in young calves with diarrhea, Belgium. *Emerg Infect Dis* 2017;23(12):2093-2095.

67. **De Rauw K, Thiry D, Caljon B, Saulmont M, Mainil J *et al.*** Characteristics of Shiga toxin producing- and enteropathogenic *Escherichia coli* of the emerging serotype O80:H2 isolated from humans and diarrhoeic calves in Belgium. *Clin Microbiol Infect* 2019;25(1):111.e115-111.e118.

68. **Nüesch-Inderbinen M, Cernela N, Wüthrich D, Egli A, Stephan R**. Genetic characterization of Shiga toxin producing *Escherichia coli* belonging to the emerging hybrid pathotype O80:H2 isolated from humans 2010-2017 in Switzerland. *Int J Med Microbiol* 2018;308(5):534-538.

69. **Bruyand M, Mariani-Kurkdjian P, Le Hello S, King LA, Van Cauteren D *et al.*** Paediatric haemolytic uraemic syndrome related to Shiga toxin-producing *Escherichia coli*, an overview of 10 years of surveillance in France, 2007 to 2016. *Euro Surveill* 2019;24(8).

70. **Cointe A, Birgy A, Bridier-Nahmias A, Mariani-Kurkdjian P, Walewski V *et al.*** *Escherichia coli* O80 hybrid pathotype strains producing Shiga toxin and ESBL: molecular characterization and potential therapeutic options. *J Antimicrob Chemother* 2020;75(3):537-542.

71. **Peigne C, Bidet P, Mahjoub-Messai F, Plainvert C, Barbe V *et al.*** The plasmid of *Escherichia coli* strain S88 (O45:K1:H7) that causes neonatal meningitis is closely related

to avian pathogenic *E. coli* plasmids and is associated with high-level bacteremia in a neonatal rat meningitis model. *Infect Immun* 2009;77(6):2272-2284.

72. **Díaz-Jiménez D, García-Meniño I, Herrera A, García V, López-Beceiro AM *et al.*** Genomic characterization of *Escherichia coli* isolates belonging to a new hybrid aEPEC/ExPEC pathotype O153:H10-A-ST10 *eae*-beta1 occurred in meat, poultry, wildlife and human diarrheagenic samples. *Antibiotics (Basel)* 2020;9(4).

73. **Gioia-Di Chiacchio RM, Cunha MPV, de Sá LRM, Davies YM, Pereira CBP *et al.*** Novel hybrid of typical enteropathogenic *Escherichia coli* and Shiga-toxin-producing *E. coli* (tEPEC/STEC) emerging from pet birds. *Front Microbiol* 2018;9:2975.

74. **Gati NS, Middendorf-Bauchart B, Bletz S, Dobrindt U, Mellmann A**. Origin and evolution of hybrid Shiga toxin-producing and uropathogenic *Escherichia coli* strains of sequence type 141. *J Clin Microbiol* 2019;58(1).

75. **Bielaszewska M, Schiller R, Lammers L, Bauwens A, Fruth A *et al.*** Heteropathogenic virulence and phylogeny reveal phased pathogenic metamorphosis in *Escherichia coli* O2:H6. *EMBO Mol Med* 2014;6(3):347-357.

76. **Kessler R, Nisa S, Hazen TH, Horneman A, Amoroso A *et al.*** Diarrhea, bacteremia and multiorgan dysfunction due to an extraintestinal pathogenic *Escherichia coli* strain with enteropathogenic *E. coli* genes. *Pathog Dis* 2015;73(8):ftv076.

77. **Dutta S, Pazhani GP, Nataro JP, Ramamurthy T**. Heterogenic virulence in a diarrheagenic *Escherichia coli*: evidence for an EPEC expressing heat-labile toxin of ETEC. *Int J Med Microbiol* 2015;305(1):47-54.

78. **Totsuka, K**. Studien über *Bacterium coli*. *Z Hyg Infektionskrankh* 1902;45:115-124.

79. **Wallick H, Stuart CA**. Antigenic relationships of *Escherichia coli* isolated from one Individual. *J Bacteriol* 1943;45(2):121-126.

80. **Smith HL**. Zur kenntnis der colibacillen des säuglingsstuhles. *Zentr Bakt Parasitenk I Orig* 1899;25:689-693.

81. **Robinet HG**. Relationship of host antibody to fluctuations of *Escherichia coli* serotypes in the human intestine. *J Bacteriol* 1962;84:896-901.

82. **Fratamico PM, DebRoy C, Liu Y, Needleman DS, Baranzoni GM *et al.*** Advances in molecular serotyping and subtyping of *Escherichia coli*. *Front Microbiol* 2016;7:644.

83. **Kauffmann F**. The serology of the coli group. *J Immunol* 1947;57(1):71-100.

84. **Kauffmann F, Vahlne, G**. Über die bedeutung des serologischen formenwechsels für die bakteriophagen-wirking in der Coli-Gruppe *Acta Pathologica et Microbiologica Scandinavica* 1945;22:119.

85. **Robins-Browne RM**. Traditional enteropathogenic *Escherichia coli* of infantile diarrhea. *Rev Infect Dis* 1987;9(1):28-53.

86. **Kauffmann F, Dupont A**. *Escherichia* strains from infantile epidemic gastro enteritis. *Acta Pathol Microbiol Scand* 1950;27(4):552-564.

87. **Neter E, Shumway CN**. *E. coli* serotype D433: occurrence in intestinal and respiratory tracts, cultural characteristics, pathogenicity, sensitivity to antibiotics. *Proc Soc Exp Biol Med* 1950;75(2):504-507.

88. **Kirby AC, Hall EG, Coackley W**. Neonatal diarrhoea and vomiting; outbreaks in the same maternity unit. *Lancet* 1950;2(6623):201-207.

89. **Ferguson WW, June RC**. Experiments on feeding adult volunteers with *Escherichia coli* 111, B4, a coliform organism associated with infant diarrhea. *Am J Hyg* 1952;55(2):155-169.

90. **Wentworth FH, Brock DW, Stulberg CS, Page RH**. Clinical, bacteriological, and serological observations of two human volunteers following ingestion of *Escherichia coli* O127:B8. *Proc Soc Exp Biol Med* 1956;91(4):586-588.

91. **Achtman M, Heuzenroeder M, Kusecek B, Ochman H, Caugant D *et al.*** Clonal analysis of *Escherichia coli* O2:K1 isolated from diseased humans and animals. *Infect Immun* 1986;51(1):268-276.

92. **Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F**. Rapid and easy *in silico* serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. *J Clin Microbiol* 2015;53(8):2410-2426.

93. **Lynch T, Petkau A, Knox N, Graham M, Van Domselaar G**. A Primer on infectious disease bacterial genomics. *Clin Microbiol Rev* 2016;29(4):881-913.

94. **Lacher DW, Gangiredla J, Jackson SA, Elkins CA, Feng PC**. Novel microarray design for molecular serotyping of shiga toxin- producing *Escherichia coli* strains isolated from fresh produce. *Appl Environ Microbiol* 2014;80(15):4677-4682.

95. **Caugant DA, Levin BR, Orskov I, Orskov F, Svanborg Eden C *et al.*** Genetic diversity in relation to serotype in *Escherichia coli*. *Infect Immun* 1985;49(2):407-413.

96. **Selander RK, Levin BR**. Genetic diversity and structure in *Escherichia coli* populations. *Science* 1980;210(4469):545-547.

97. **Milkman R**. Electrophoretic variation in *Escherichia coli* from natural sources. *Science* 1973;182(4116):1024-1026.

98. **Ochman H,** Evolution of bacterial pathogens. In: **Eduardo A**. **Groisman** (eds). Principles of Bacterial Pathogenesis. Academic Press; California, USA. 2001; Chapter 1 pp. 1-41.

99. **Levin BR**. Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. *Genetics* 1981;99(1):1-23.

100. **Whittam TS, Ochman H, Selander RK**. Geographic components of linkage disequilibrium in natural populations of *Escherichia coli*. *Mol Biol Evol* 1983;1(1):67-83.

101. **Dykhuizen DE, de Framond J, Hartl DL**. Selective neutrality of glucose-6-phosphate dehydrogenase allozymes in *Escherichia coli*. *Mol Biol Evol* 1984;1(2):162-170.

102. **Selander RK, Caugant DA, Ochman H, Musser JM, Gilmour MN *et al.*** Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Appl Environ Microbiol* 1986;51(5):873-884.

103. **Whittam TS, Ochman H, Selander RK**. Multilocus genetic structure in natural populations of *Escherichia coli*. *Proc Natl Acad Sci USA* 1983;80(6):1751-1755.

104. **Caugant DA, Levin BR, Selander RK**. Genetic diversity and temporal variation in the *E. coli* population of a human host. *Genetics* 1981;98(3):467-490.

105. **Ochman H, Selander RK**. Standard reference strains of *Escherichia coli* from natural populations. *J Bacteriol* 1984;157(2):690-693.

106. **Ochman H, Selander RK**. Evidence for clonal population structure in *Escherichia coli*. *Proc Natl Acad Sci USA* 1984;81(1):198-201.

107. **Selander RK, Caugant, DA, Whittam, TS**. Genetic structure and variation in natural populations of *Escherichia coli*. In: **Neidhardt FC, EIngraham JL, Low KB, Magasanik B, Schaechter M & Umbarger HE** (eds). *Escherichia coli* and *Salmonella typhimurium:* Cellular and Molecular Biology. ASM Press; Washington, DC. 1987; pp. 1625–1648.

108. **Desjardins P, Picard B, Kaltenböck B, Elion J, Denamur E**. Sex in *Escherichia coli* does not disrupt the clonal structure of the population: evidence from random amplified polymorphic DNA and restriction-fragment-length polymorphism. *J Mol Evol* 1995;41(4):440-448.

109. **Escobar-Páramo P, Clermont O, Blanc-Potard AB, Bui H, Le Bouguénec C *et al.*** A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. *Mol Biol Evol* 2004;21(6):1085-1094.

110. **Wirth T, Falush D, Lan R, Colles F, Mensa P *et al.*** Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* 2006;60(5):1136-1151.

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

112. **Sneath, PHA, Sokal, RR**. Numerical Taxonomy: The Principles and Practice of Numerical Classification. Freeman, San Francisco. 1973.

113. **Herzer PJ, Inouye S, Inouye M, Whittam TS**. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. *J Bacteriol* 1990;172(11):6175-6181.

114. **Saitou N, Nei M**. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;4(4):406-425.

115. **Vangchhia B, Abraham S, Bell JM, Collignon P, Gibson JS *et al.*** Phylogenetic diversity, antimicrobial susceptibility and virulence characteristics of phylogroup F *Escherichia coli* in Australia. *Microbiology (Reading)* 2016;162(11):1904-1912.

116. **Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O**. ClermonTyping: an easy-to-use and accurate *in silico* method for *Escherichia* genus strain phylotyping. *Microbial genomics* 2018;4(7):e000192.

117. **Clermont O, Dixit OVA, Vangchhia B, Condamine B, Bridier-Nahmias A *et al.*** Characterisation and rapid identification of phylogroup G in *Escherichia coli*, a lineage with high virulence and antibiotic resistance potential. *Environ Microbiol* 2019.

118. **Walk ST, Alm EW, Gordon DM, Ram JL, Toranzos GA *et al.*** Cryptic lineages of the genus *Escherichia*. *Appl Environ Microbiol* 2009;75(20):6534-6544.

119. **Picard B, Garcia JS, Gouriou S, Duriez P, Brahimi N *et al.*** The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. *Infect Immun* 1999;67(2):546-553.

120. **Clermont O, Bonacorsi S, Bingen E**. Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Appl Environ Microbiol* 2000;66(10):4555-4558.

121. **Clermont O, Christenson JK, Denamur E, Gordon DM**. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ Microbiol Rep* 2013;5(1):58-65.

122. **Clermont O, Gordon DM, Brisse S, Walk ST, Denamur E**. Characterization of the cryptic *Escherichia l*ineages: rapid identification and prevalence. *Environ Microbiol* 2011;13(9):2468-2477.

123. **Vignaroli C, Di Sante L, Magi G, Luna GM, Di Cesare A *et al.*** Adhesion of marine cryptic *Escherichia* isolates to human intestinal epithelial cells. *ISME J* 2015;9(2):508-515.

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

125. **Waters NR, Abram F, Brennan F, Holmes A, Pritchard L**. Easily phylotyping *E. coli* via the EzClermont web app and command-line tool. *Access Microbiol* 2020;2(9):acmi000143.

126. **Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE *et al.*** Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA* 1998;95(6):3140-3145.

127. **Ibarz Pavón AB, Maiden MC**. Multilocus sequence typing. *Methods Mol Biol* 2009;551:129-140.

128. **Smith JM, Smith NH, O'Rourke M, Spratt BG**. How clonal are bacteria? *Proc Natl Acad Sci USA* 1993;90(10):4384-4388.

129. **Smith JM, Dowson CG, Spratt BG**. Localized sex in bacteria. *Nature* 1991;349(6304):29-31.

130. **Smith JM, Feil EJ, Smith NH**. Population structure and evolutionary dynamics of pathogenic bacteria. *Bioessays* 2000;22(12):1115-1122.

131. **Maiden MC**. Multilocus sequence typing of bacteria. *Annu Rev Microbiol* 2006;60:561-588.

132. **Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG**. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol* 2004;186(5):1518-1530.

133. **Clermont O, Gordon D, Denamur E**. Guide to the various phylogenetic classification schemes for *Escherichia coli* and the correspondence among schemes. *Microbiology* 2015;161(Pt 5):980-988.

134. **Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS**. Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* 2000;406(6791):64-67.

135. **Escobar-Páramo P, Sabbagh A, Darlu P, Pradillon O, Vaury C *et al.*** Decreasing the effects of horizontal gene transfer on bacterial phylogeny: the *Escherichia coli* case study. *Mol Phylogenet Evol* 2004;30(1):243-250.

136. **Kaas RS, Friis C, Ussery DW, Aarestrup FM**. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics* 2012;13:577.

137. **Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ**. Within-host evolution of bacterial pathogens. *Nat Rev Microbiol* 2016;14(3):150-162.

138. **Page AJ, Alikhan NF, Carleton HA, Seemann T, Keane JA *et al.*** Comparison of classical multi-locus sequence typing software for next-generation sequencing data. *Microb Genom* 2017;3(8):e000124.

139. **Zhou Z, Alikhan NF, Mohamed K, Fan Y, Achtman M *et al.*** The EnteroBase user's guide, with case studies on *Salmonella* transmission, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res* 2020;30(1):138-152.

140. **Frentrup M, Zhou Z, Steglich M, Meier-Kolthoff JP, Göker M *et al.*** A publicly accessible database for *Clostridiodes difficile* genome sequences supports tracing of transmission chains and epidemics. *Microb Genom* 2020;6(8).

141. **Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C *et al.*** GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res* 2018;28(9):1395-1404.

142. **Burgess NR, McDermott SN, Whiting J**. Aerobic bacteria occurring in the hind-gut of the cockroach, *Blatta orientalis*. *J Hyg (Lond)* 1973;71(1):1-7.

143. **Farmer JJ, Fanning GR, Davis BR, O'Hara CM, Riddle C *et al.*** *Escherichia fergusonii* and *Enterobacter taylorae*, two new species of *Enterobacteriaceae* isolated from clinical specimens. *J Clin Microbiol* 1985;21(1):77-81.

144. **Brenner DJ, Davis BR, Steigerwalt AG, Riddle CF, McWhorter AC *et al.*** Atypical biogroups of *Escherichia coli* found in clinical specimens and description of *Escherichia hermannii* sp. nov. *J Clin Microbiol* 1982;15(4):703-713.

145. **Brenner DJ, McWhorter AC, Knutson JK, Steigerwalt AG**. *Escherichia vulneris*: a new species of *Enterobacteriaceae* associated with human wounds. *J Clin Microbiol* 1982;15(6):1133-1140.

146. **Lawrence JG, Ochman H, Hartl DL**. Molecular and evolutionary relationships among enteric bacteria. *J Gen Microbiol* 1991;137(8):1911-1921.

147. **Albert MJ, Alam K, Islam M, Montanaro J, Rahaman AS *et al.*** *Hafnia alvei*, a probable cause of diarrhea in humans. *Infect Immun* 1991;59(4):1507-1513.

148. **Huys G, Cnockaert M, Janda JM, Swings J**. *Escherichia albertii* sp. nov., a diarrhoeagenic species isolated from stool specimens of Bangladeshi children. *Int J Syst Evol Microbiol* 2003;53(Pt 3):807-810.

149. **Hyma KE, Lacher DW, Nelson AM, Bumbaugh AC, Janda JM *et al.*** Evolutionary genetics of a new pathogenic *Escherichia* species: *Escherichia albertii* and related *Shigella boydii* strains. *J Bacteriol* 2005;187(2):619-628.

150. **Priest FG, Barker M**. Gram-negative bacteria associated with brewery yeasts: reclassification of *Obesumbacterium proteus* biogroup 2 as *Shimwellia pseudoproteus* gen. nov., sp. nov., and transfer of *Escherichia blattae* to *Shimwellia blattae* comb. nov. *Int J Syst Evol Microbiol* 2010;60(Pt 4):828-833.

151. **Hata H, Natori T, Mizuno T, Kanazawa I, Eldesouky I *et al.*** Phylogenetics of family *Enterobacteriacea*e and proposal to reclassify *Escherichia hermannii* and *Salmonella subterranea* as *Atlantibacter hermannii* and *Atlantibacter subterranea* gen. nov., comb. nov. *Microbiol Immunol* 2016;60(5):303-311.

152. **Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S**. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;9(1):5114.

153. **Liu S, Jin D, Lan R, Wang Y, Meng Q *et al.*** *Escherichia marmotae* sp. nov., isolated from faeces of *Marmota himalayana*. *Int J Syst Evol Microbiol* 2015;65(7):2130-2134.

154. **Liu S, Feng J, Pu J, Xu X, Lu S *et al.*** Genomic and molecular characterisation of *Escherichia marmotae* from wild rodents in Qinghai-Tibet plateau as a potential pathogen. *Sci Rep* 2019;9(1):10619.

155. **Ocejo M, Tello M, Oporto B, Lavín JL, Hurtado A**. Draft genome sequence of *Escherichia marmotae* E690, isolated from beef cattle. *Microbiol Resour Announc* 2020;9(32).

156. **Gilroy R RA, Getino M, Pursley I, Horton D, Alikhan N, Baker D, Gharbi K, Hall N, Watson M, Adriaenssens EM, Foster-Nyarko E, Jarju S, Secka A, Antonio M, Oren A, Chaudhuri R, Hildebrand F, Pallen M**. A Genomic blueprint of the chicken gut microbiome. *Research Square* (Preprint) 2020.

157. **Richter M, Rosselló-Móra R**. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci USA* 2009;106(45):19126-19131.

158. **Steinsland H, Lacher DW, Sommerfelt H, Whittam TS**. Ancestral lineages of human enterotoxigenic *Escherichia coli*. *J Clin Microbiol* 2010;48(8):2916-2924.

159. **Ingle DJ, Clermont O, Skurnik D, Denamur E, Walk ST *et al.*** Biofilm formation by and thermal niche and virulence characteristics of *Escherichia* spp. *Appl Environ Microbiol* 2011;77(8):2695-2700.

160. **Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH**. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* 2019.

161. **Pallen MJ, Telatin A, Oren A**. The next million names for Archaea and Bacteria. *Trends Microbiol* 2020.

162. **Teuber M**. Spread of antibiotic resistance with food-borne pathogens. *Cell Mol Life Sci* 1999;56(9-10):755-763.

163. **Bongers JH, Franssen F, Elbers AR, Tielen MJ**. Antimicrobial resistance of *Escherichia coli* isolates from the faecal flora of veterinarians with different professional specialties. *Vet Q* 1995;17(4):146-149.

164. **Datta N**. Drug resistance and R factors in the bowel bacteria of London patients before and after admission to hospital. *Br Med J* 1969;2(5654):407-411.

165. **Guinée P, Ugueto N, van Leeuwen N**. *Escherichia coli* with resistance factors in vegetarians, babies, and nonvegetarians. *Appl Microbiol* 1970;20(4):531-535.

166. **Linton KB, Lee PA, Richmond MH, Gillespie WA, Rowland AJ et al.** Antibiotic resistance and transmissible R-factors in the intestinal coliform flora of healthy adults and children in an urban and a rural community. *J Hyg (Lond)* 1972;70(1):99-104.

167. **Gyles CL, Palchaudhuri S, Maas WK**. Naturally occurring plasmid carrying genes for enterotoxin production and drug resistance. *Science* 1977;198(4313):198-199.

168. **Davies M, Stewart PR**. Transferable drug resistance in man and animals: genetic relationship between R-plasmids in enteric bacteria from man and domestic pets. *Aust Vet J* 1978;54(11):507-512.

169. **Bourque R, Lallier R, Larivière S**. Influence of oral antibiotics on resistance and enterotoxigenicity of *Escherichia coli*. *Can J Comp Med* 1980;44(1):101-108.

170. **Nijsten R, London N, van den Bogaard A, Stobberingh E**. In-vitro transfer of antibiotic resistance between faecal *Escherichia coli* strains isolated from pig farmers and pigs. *J Antimicrob Chemother* 1996;37(6):1141-1154.

171. **Brown L, Langelier C, Reid MJ, Rutishauser RL, Strnad L.** Antimicrobial resistance: A call to action! Clin Infect Dis 2017;64(1):106-107.

172. **Riley LW**. Distinguishing pathovars from nonpathovars: *Escherichia coli*. *Microbiol Spectr* 2020;8(4).

173. **de Kraker ME, Jarlier V, Monen JC, Heuer OE, van de Sande N *et al.*** The changing epidemiology of bacteraemias in Europe: trends from the European antimicrobial resistance surveillance system. *Clin Microbiol Infect* 2013;19(9):860-868.

174. **Foxman B**. The epidemiology of urinary tract infection. *Nat Rev Urol* 2010;7(12):653-660.

175. **Webber MA, Piddock, LJ**. Antibiotic resistance in *Escherichia coli*. *Frontiers in Antimicrobial Resistance: a Tribute to Stuart B Levy*. ASM Press; Washington, DC. 2005;pp. 374-386.

176. **Ventola CL**. The antibiotic resistance crisis: part 1: causes and threats. *P T* 2015;40(4):277-283.

177. **Cheng G, Dai M, Ahmed S, Hao H, Wang X *et al.*** Antimicrobial drugs in fighting against antimicrobial resistance. *Front Microbiol* 2016;7:470.

178. **de la Cruz F, Davies J**. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol* 2000;8(3):128-133.

179. **Lee JW**. Protocol measuring horizontal gene transfer from algae to non-photosynthetic organisms. *MethodsX* 2019;6:1564-1574.

180. **Burmeister AR**. Horizontal gene transfer. *Evol Med Public Health* 2015;2015(1):193-194.

181. **Croxall G, Hale J, Weston V, Manning G, Cheetham P *et al.*** 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. *J Antimicrob Chemother* 2011;66(11):2501-2508.

182. **Alhashash F, Weston V, Diggle M, McNally A**. Multidrug-resistant *Escherichia coli* bacteremia. *Emerg Infect Dis* 2013;19(10):1699-1701.

183. **Dunn SJ, Connor C, McNally A**. The evolution and transmission of multi-drug resistant *Escherichia coli* and *Klebsiella pneumoniae*: the complexity of clones and plasmids. *Curr Opin Microbiol* 2019;51:51-56.

184. **McNally A, Kallonen T, Connor C, Abudahab K, Aanensen DM *et al.*** Diversification of colonization factors in a multidrug-resistant *Escherichia coli* lineage evolving under negative frequency-dependent selection. *mBio* 2019;10(2).

185. **Banerjee R, Johnson JR**. A new clone sweeps clean: the enigmatic emergence of *Escherichia coli* sequence type 131. *Antimicrob Agents Chemother* 2014;58(9):4997-5004.

186. **Stoesser N, Sheppard AE, Pankhurst L, De Maio N, Moore CE *et al.*** Evolutionary history of the global emergence of the *Escherichia coli* epidemic clone ST131. *mBio* 2016;7(2):e02162.

187. **Ben Zakour NL, Alsheikh-Hussain AS, Ashcroft MM, Khanh Nhu NT, Roberts LW *et al.*** Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of *Escherichia coli* ST131. *mBio* 2016;7(2):e00347-00316.

188. **Zogg AL, Zurfluh K, Schmitt S, Nuesch-Inderbinen M, Stephan R**. Antimicrobial resistance, multilocus sequence types and virulence profiles of ESBL producing and non-ESBL producing uropathogenic *Escherichia coli* isolated from cats and dogs in Switzerland. *Vet Microbiol* 2018;216:79-84.

189. **Johnson JR, Clabots C, Kuskowski MA**. Multiple-host sharing, long-term persistence, and virulence of *Escherichia coli* clones from human and animal household members. *J Clin Microbiol* 2008;46(12):4078-4082.

190. **Johnson JR, Owens K, Gajewski A, Clabots C**. *Escherichia coli* colonization patterns among human household members and pets, with attention to acute urinary tract infection. *J Infect Dis* 2008;197(2):218-224.

191. **Johnson JR, Miller S, Johnston B, Clabots C, Debroy C**. Sharing of *Escherichia coli* sequence type ST131 and other multidrug-resistant and urovirulent *E. coli* strains among dogs and cats within a household. *J Clin Microbiol* 2009;47(11):3721-3725.

192. **Ewers C, Grobbel M, Stamm I, Kopp PA, Diehl I *et al.*** Emergence of human pandemic O25:H4-ST131 CTX-M-15 extended-spectrum-β-lactamase-producing *Escherichia coli* among companion animals. *Antimicrob Agents Chemother* 2010;65(4):651-660.

193. **Bennani H, Mateus A, Mays N, Eastmure E, Stärk KDC *et al.*** Overview of evidence of antimicrobial use and antimicrobial resistance in the food chain. *Antibiotics (Basel)* 2020;9(2).

194. **Castanon JIR**. History of the use of antibiotic as growth promoters in European poultry feeds. *Poult Sci J* 2007;86(11):2466-2471.

195. **Van Boeckel TP, Brower C, Gilbert M, Grenfell BT, Levin SA *et al.*** Global trends in antimicrobial use in food animals. *Proceedings of the National Academy of Sciences* 2015;112(18):5649.

196. **Lazarus B, Paterson DL, Mollinger JL, Rogers BA**. Do human extraintestinal *Escherichia coli* infections resistant to expanded-spectrum cephalosporins originate from food-producing animals? A systematic review. *Clin Infect Dis* 2015;60(3):439-452.

197. **Borges CA, Tarlton NJ, Riley LW**. *Escherichia coli* from commercial broiler and backyard chickens share sequence types, antimicrobial resistance profiles, and resistance genes with human extraintestinal pathogenic *Escherichia coli*. *Foodborne Pathog Dis* 2019;16(12):813-822.

198. **Thakuria B, Lahon K**. The beta lactam antibiotics as an empirical therapy in a developing country: An update on their current status and recommendations to counter the resistance against them. *J Clin Diagn Res* 2013;7(6):1207-1214.

199. **Lacey RW**. Antibiotic resistance in *Staphylococcus aureus* and streptococci. *Br Med Bull* 1984;40(1):77-83.

200. **Bush K, Jacoby GA**. Updated functional classification of beta-lactamases. *Antimicrob Agents Chemother* 2010;54(3):969-976.

201. **Ur Rahman S, Ali T, Ali I, Khan NA, Han B *et al.*** The growing genetic and functional diversity of extended spectrum beta-lactamases. *Biomed Res Int* 2018;2018:9519718.

202. **Jochum JM, Redweik GAJ, Ott LC, Mellata M**. Bacteria broadly-resistant to last resort antibiotics detected in commercial chicken farms. *Microorganisms* 2021;9(1).

203. **Vounba P, Arsenault J, Bada-Alambédji R, Fairbrother JM**. Prevalence of antimicrobial resistance and potential pathogenicity, and possible spread of third generation cephalosporin resistance, in *Escherichia coli* isolated from healthy chicken farms in the region of Dakar, Senegal. *PLos One* 2019;14(3):e0214304.

204. **Moawad AA, Hotzel H, Neubauer H, Ehricht R, Monecke S *et al.*** Antimicrobial resistance in *Enterobacteriaceae* from healthy broilers in Egypt: emergence of colistin-resistant and extended-spectrum β-lactamase-producing *Escherichia coli*. *Gut Pathog* 2018;10:39.

205. **Tufa TB, Fuchs A, Stötter L, Kaasch AJ, Feldt T *et al.*** High rate of extended-spectrum beta-lactamase-producing gram-negative infections and associated mortality in Ethiopia: a systematic review and meta-analysis. *Antimicrob Resist Infect Control* 2020;9(1):128.

206. **Al Fadhli AH, Jamal WY, Rotimi VO**. Prevalence of carbapenem-resistant *Enterobacteriaceae* and emergence of high rectal colonization rates of *bla*$_{OXA-181}$-positive

isolates in patients admitted to two major hospital intensive care units in Kuwait. *PLos One* 2020;15(11):e0241971.

207. **Hubbard ATM, Mason J, Roberts P, Parry CM, Corless C *et al.*** Piperacillin/tazobactam resistance in a clinical isolate of *Escherichia coli* due to IS*26*-mediated amplification of *bla*$_{TEM-1B}$. *Nat Commun* 2020;11(1):4915.

208. **Shnaiderman-Torban A, Navon-Venezia S, Dahan R, Dor Z, Taulescu M *et al.*** CTX-M-15 Producing *Escherichia coli* sequence type 361 and sequence type 38 causing bacteremia and umbilical infection in a neonate foal. *J Equine Vet Sci* 2020;85:102881.

209. **Kremer K, Kramer R, Neumann B, Haller S, Pfennigwerth N *et al.*** Rapid spread of OXA-244-producing *Escherichia coli* ST38 in Germany: insights from an integrated molecular surveillance approach; 2017 to January 2020. *Euro Surveill* 2020;25(25).

210. **Kot B**. Antibiotic resistance among uropathogenic *Escherichia coli*. *Pol J Microbiol* 2019;68(4):403-415.

211. **Pfeifer Y, Cullik A, Witte W**. Resistance to cephalosporins and carbapenems in Gram-negative bacterial pathogens. *Int J Med Microbiol* 2010;300(6):371-379.

212. **Jacoby GA**. AmpC beta-lactamases. *Clin Microbiol Rev* 2009;22(1):161-182.

213. **Medeiros AA**. Evolution and dissemination of beta-lactamases accelerated by generations of beta-lactam antibiotics. *Clin Infect Dis* 1997;24 Suppl 1:S19-45.

214. **Roberts MC**. Tetracycline resistance determinants: mechanisms of action, regulation of expression, genetic mobility, and distribution. *FEMS Microbiol Rev* 1996;19(1):1-24.

215. **van den Bogaard AE, Stobberingh EE**. Epidemiology of resistance to antibiotics. Links between animals and humans. *Int J Antimicrob Agents* 2000;14(4):327-335.

216. **Ingle DJ, Levine MM, Kotloff KL, Holt KE, Robins-Browne RM**. Dynamics of antimicrobial resistance in intestinal *Escherichia coli* from children in community settings in South Asia and sub-Saharan Africa. *Nat Microbiol* 2018;3(9):1063-1073.

217. **Adesoji AT, Liadi AM**. Antibiogram studies of *Escherichia coli* and *Salmonella* species isolated from diarrheal patients attending Malam Mande General Hospital Dutsin-Ma, Katsina State, Nigeria. *Pan Afr Med J* 2020;37:110.

218. **Talavera-Gonzalez JM, Talavera-Rojas M, Soriano-Vargas E, Vazquez-Navarrete J, Salgado-Miranda C**. *In vitro* transduction of antimicrobial resistance genes into *Escherichia coli* isolates from backyard poultry in Mexico. *Can J Microbiol* 2021.

219. **Boriollo MFG, Moreira BS, Oliveira MC, Santos TO, Rufino LRA *et al.*** Incidence of Shiga toxin-producing *Escherichia coli* in diarrheic calves and its susceptibility profile to antimicrobials and *Eugenia uniflora* L. *Can J Vet Res* 2021;85(1):18-26.

220. **Rahman MM, Husna A, Elshabrawy HA, Alam J, Runa NY *et al.*** Isolation and molecular characterization of multidrug-resistant *Escherichia coli* from chicken meat. *Sci Rep* 2020;10(1):21999.

221. **Mesa-Varona O, Kaspar H, Grobbel M, Tenhagen BA**. Phenotypical antimicrobial resistance data of clinical and non-clinical *Escherichia coli* from poultry in Germany between 2014 and 2017. *PLos One* 2020;15(12):e0243772.

222. **Madoshi BP, Kudirkiene E, Mtambo MM, Muhairwa AP, Lupindu AM *et al.*** Characterisation of commensal *Escherichia coli* isolated from apparently healthy cattle and their attendants in Tanzania. *PLos One* 2016;11(12):e0168160.

223. **Karami N, Nowrouzian F, Adlerberth I, Wold AE**. Tetracycline resistance in *Escherichia coli* and persistence in the infantile colonic microbiota. *Antimicrob Agents Chemother* 2006;50(1):156-161.

224. **Chopra I, Roberts M**. Tetracycline antibiotics: mode of action, applications, molecular biology, and epidemiology of bacterial resistance. *Microbiol Mol Biol Rev* 2001;65(2):232-260.

225. **Levy SB, McMurry LM, Barbosa TM, Burdett V, Courvalin P *et al.*** Nomenclature for new tetracycline resistance determinants. *Antimicrob Agents Chemother* 1999;43(6):1523-1524.

226. **Roberts MC, Schwarz S**. Tetracycline and phenicol resistance genes and mechanisms: Importance for agriculture, the environment, and humans. *J Environ Qual* 2016;45(2):576-592.

227. **Schnappinger D, Hillen W**. Tetracyclines: antibiotic action, uptake, and resistance mechanisms. *Arch Microbiol* 1996;165(6):359-369.

228. **Roberts MC**. Update on acquired tetracycline resistance genes. *FEMS Microbiol Lett* 2005;245(2):195-203.

229. **Koo HJ, Woo GJ**. Distribution and transferability of tetracycline resistance determinants in *Escherichia coli* isolated from meat and meat products. *Int J Food Microbiol* 2011;145(2-3):407-413.

230. **Bailey JK, Pinyon JL, Anantham S, Hall RM**. Commensal *Escherichia coli* of healthy humans: a reservoir for antibiotic-resistance determinants. *J Med Microbiol* 2010;59(Pt 11):1331-1339.

231. **Bryan A, Shapir N, Sadowsky MJ**. Frequency and distribution of tetracycline resistance genes in genetically diverse, nonselected, and nonclinical *Escherichia coli* strains isolated from diverse human and animal sources. *Appl Environ Microbiol* 2004;70(4):2503-2507.

232. **Pong CH, Moran RA, Hall RM**. Evolution of IS*26*-bounded pseudo-compound transposons carrying the *tet(C)* tetracycline resistance determinant. *Plasmid* 2020;112:102541.

233. **Turner PE, Williams ES, Okeke C, Cooper VS, Duffy S *et al.*** Antibiotic resistance correlates with transmission in plasmid evolution. *Evolution* 2014;68(12):3368-3380.

234. **Anantham S, Hall RM**. pCERC1, a small, globally disseminated plasmid carrying the *dfrA14* cassette in the *strA* gene of the *sul2-strA-strB* gene cluster. *Microb Drug Resist* 2012;18(4):364-371.

235. **Moodley A, Guardabassi L**. Transmission of IncN plasmids carrying *bla*$_{CTX-M-1}$ between commensal *Escherichia coli* in pigs and farm workers. *Antimicrob Agents Chemother* 2009;53(4):1709-1711.

236. **Hellweger FL**. *Escherichia coli* adapts to tetracycline resistance plasmid (pBR322) by mutating endogenous potassium transport: *in silico* hypothesis testing. *FEMS Microbiol Ecol* 2013;83(3):622-631.

237. **Serio AW, Keepers T, Andrews L, Krause KM**. Aminoglycoside revival: Review of a historically important class of antimicrobials undergoing rejuvenation. *EcoSal Plus* 2018;8(1).

238. **Ramirez MS, Tolmasky ME**. Aminoglycoside modifying enzymes. *Drug Resist Updat* 2010;13(6):151-171.

239. **Kettner M, Macicková T**. [Bacterial resistance to aminoglycoside antibiotics]. *Bratisl Lek Listy* 1992;93(4):175-178.

240. **Cirit OS, Fernández-Martínez M, Yayla B, Martínez-Martínez L**. Aminoglycoside resistance determinants in multiresistant *Escherichia coli* and *Klebsiella pneumoniae* clinical isolates from Turkish and Syrian patients. *Acta Microbiol Immunol Hung* 2019;66(3):327-335.

241. **Alyamani EJ, Khiyami AM, Booq RY, Majrashi MA, Bahwerth FS *et al.*** The occurrence of ESBL-producing *Escherichia coli* carrying aminoglycoside resistance genes in urinary tract infections in Saudi Arabia. *Ann Clin Microbiol Antimicrob* 2017;16(1):1.

242. **Lindemann PC, Risberg K, Wiker HG, Mylvaganam H**. Aminoglycoside resistance in clinical *Escherichia coli* and *Klebsiella pneumoniae* isolates from Western Norway. *APMIS* 2012;120(6):495-502.

243. **Davies J, Wright GD**. Bacterial resistance to aminoglycoside antibiotics. *Trends Microbiol* 1997;5(6):234-240.

244. **Karczmarczyk M, Martins M, Quinn T, Leonard N, Fanning S**. Mechanisms of fluoroquinolone resistance in *Escherichia coli* isolates from food-producing animals. *Appl Environ Microbiol* 2011;77(20):7113-7120.

245. **Jacoby GA, Strahilevitz J, Hooper DC**. Plasmid-mediated quinolone resistance. *Microbiol Spectr* 2014;2(5).

246. **Caron F, Etienne M**. Emergence of fluoroquinolone resistance in outpatient urinary *Escherichia coli* isolates. *Am J Med* 2010;123(3):e13.

247. **Johnson L, Sabel A, Burman WJ, Everhart RM, Rome M *et al.*** Emergence of fluoroquinolone resistance in outpatient urinary *Escherichia coli* isolates. *Am J Med* 2008;121(10):876-884.

248. **Hooton TM**. Fluoroquinolones and resistance in the treatment of uncomplicated urinary tract infection. *Int J Antimicrob Agents* 2003;22 Suppl 2:65-72.

249. **Chen SL, Wu M, Henderson JP, Hooton TM, Hibbing ME *et al.*** Genomic diversity and fitness of *E. coli* strains recovered from the intestinal and urinary tracts of women with recurrent urinary tract infection. *Sci Transl Med* 2013;5(184):184ra160.

250. **Urbánek K, Kolár M, Strojil J, Koukalová D, Cekanová L *et al.*** Utilization of fluoroquinolones and *Escherichia coli* resistance in urinary tract infection: inpatients and outpatients. *Pharmacoepidemiol Drug Saf* 2005;14(10):741-745.

251. **Zhanel GG, Hisanaga TL, Laing NM, DeCorby MR, Nichol KA *et al.*** Antibiotic resistance in outpatient urinary isolates: final results from the North American Urinary Tract Infection Collaborative Alliance (NAUTICA). *Int J Antimicrob Agents* 2005;26(5):380-388.

252. **Karlowsky JA, Hoban DJ, Decorby MR, Laing NM, Zhanel GG**. Fluoroquinolone-resistant urinary isolates of *Escherichia coli* from outpatients are frequently multidrug resistant: results from the North American Urinary Tract Infection Collaborative Alliance-Quinolone Resistance study. *Antimicrob Agents Chemother* 2006;50(6):2251-2254.

253. **Fedler KA, Biedenbach DJ, Jones RN**. Assessment of pathogen frequency and resistance patterns among pediatric patient isolates: report from the 2004 SENTRY Antimicrobial Surveillance Program on 3 continents. *Diagn Microbiol Infect Dis* 2006;56(4):427-436.

254. **Karlowsky JA, Kelly LJ, Thornsberry C, Jones ME, Sahm DF**. Trends in antimicrobial resistance among urinary tract infection isolates of *Escherichia coli* from female outpatients in the United States. *Antimicrob Agents Chemother* 2002;46(8):2540-2545.

255. **Sader HS, Castanheira M, Flamm RK, Jones RN**. Antimicrobial activities of ceftazidime-avibactam and comparator agents against Gram-negative organisms isolated from patients with urinary tract infections in U.S. medical centers, 2012 to 2014. *Antimicrob Agents Chemother* 2016;60(7):4355-4360.

256. **Sköld O**. Resistance to trimethoprim and sulfonamides. *Vet Res* 2001;32(3-4):261-273.

257. **Roberts MC**. Resistance to tetracycline, macrolide-lincosamide-streptogramin, trimethoprim, and sulfonamide drug classes. *Mol Biotechnol* 2002;20(3):261-283.

258. **Huovinen P, Sundström L, Swedberg G, Sköld O**. Trimethoprim and sulfonamide resistance. *Antimicrob Agents Chemother* 1995;39(2):279-289.

259. **Sköld O**. Sulfonamides and trimethoprim. *Expert Rev Anti Infect Ther* 2010;8(1):1-6.

260. **Sundström L, Rådström P, Swedberg G, Sköld O**. Site-specific recombination promotes linkage between trimethoprim- and sulfonamide resistance genes. Sequence characterization of *dhfrV* and *sulI* and a recombination active locus of Tn*21*. *Mol Gen Genet* 1988;213(2-3):191-201.

261. **Morrill HJ, Morton JB, Caffrey AR, Jiang L, Dosa D *et al.*** Antimicrobial resistance of *Escherichia coli* urinary isolates in the Veterans Affairs Health Care System. *Antimicrob Agents Chemother* 2017;61(5).

262. **Flensburg J, Sköld O**. Massive overproduction of dihydrofolate reductase in bacteria as a response to the use of trimethoprim. *Eur J Biochem* 1987;162(3):473-476.

263. **King CH, Shlaes DM, Dul MJ**. Infection caused by thymidine-requiring, trimethoprim-resistant bacteria. *J Clin Microbiol* 1983;18(1):79-83.

264. **Crofts TS, Sontha P, King AO, Wang B, Biddy BA *et al.*** Discovery and characterization of a nitroreductase capable of conferring bacterial resistance to chloramphenicol. *Cell Chem Biol* 2019;26(4):559-570.e556.

265. **Lim C, Takahashi E, Hongsuwan M, Wuthiekanun V, Thamlikitkul V *et al.*** Epidemiology and burden of multidrug-resistant bacterial infection in a developing country. *Elife* 2016;5.

266. **Sood S**. Chloramphenicol - A potent armament against multi-drug resistant (MDR) Gram negative bacilli? *J Clin Diagn Res* 2016;10(2):DC01-03.

267. **Dinos GP, Athanassopoulos CM, Missiri DA, Giannopoulou PC, Vlachogiannis IA *et al.*** Chloramphenicol derivatives as antibacterial and anticancer agents: Historic problems and current solutions. *Antibiotics (Basel)* 2016;5(2).

268. **Long KS, Poehlsgaard J, Kehrenberg C, Schwarz S, Vester B**. The Cfr rRNA methyltransferase confers resistance to phenicols, lincosamides, oxazolidinones, pleuromutilins, and streptogramin A antibiotics. *Antimicrob Agents Chemother* 2006;50(7):2500-2505.

269. **Schwarz S, Kehrenberg C, Doublet B, Cloeckaert A**. Molecular basis of bacterial resistance to chloramphenicol and florfenicol. *FEMS Microbiol Rev* 2004;28(5):519-542.

270. **Abo-Amer AE, Shobrak MY, Altalhi AD**. Isolation and antimicrobial resistance of *Escherichia coli* isolated from farm chickens in Taif, Saudi Arabia. *J Glob Antimicrob Resist* 2018;15:65-68.

271. **Jaja IF, Oguttu J, Jaja CI, Green E**. Prevalence and distribution of antimicrobial resistance determinants of *Escherichia coli* isolates obtained from meat in South Africa. *PLoS One* 2020;15(5):e0216914.

272. **Du Z, Wang M, Cui G, Zu X, Zhao Z *et al.*** The prevalence of amphenicol resistance in *Escherichia coli* isolated from pigs in mainland China from 2000 to 2018: A systematic review and meta-analysis. *PLos One* 2020;15(2):e0228388.

273. **Hoang PH, Awasthi SP, DO Nguyen P, Nguyen NL, Nguyen DT *et al.*** Antimicrobial resistance profiles and molecular characterization of *Escherichia coli* strains isolated from healthy adults in Ho Chi Minh City, Vietnam. *J Vet Med Sci* 2017;79(3):479-485.

274. **Naber KG, Bonkat G, Wagenlehner FME**. The EAU and AUA/CUA/SUFU guidelines on recurrent urinary tract infections: What is the difference? *Eur Urol* 2020;78(5):645-646.

275. **Sanger F**. Sequences, sequences, and sequences. *Annu Rev Biochem* 1988;57:1-28.

276. **Maxam AM, Gilbert W**. A new method for sequencing DNA. *Proc Natl Acad Sci USA* 1977;74(2):560-564.

277. **Staden R**. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* 1979;6(7):2601-2610.

278. **Sanger F**. Determination of nucleotide sequences in DNA. *Science* 1981;214(4526):1205-1210.

279. **Sanger F, Nicklen S, Coulson AR**. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977;74(12):5463-5467.

280. **Fraser CM, Fleischmann RD**. Strategies for whole microbial genome sequencing and analysis. *Electrophoresis* 1997;18(8):1207-1216.

281. **Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF *et al.*** Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995;269(5223):496-512.

282. **Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA *et al.*** The minimal gene complement of *Mycoplasma genitalium*. *Science* 1995;270(5235):397-403.

283. **Loman NJ, Pallen MJ**. Twenty years of bacterial genome sequencing. *Nat Rev Microbiol* 2015;13(12):787-794.

284. **Land M, Hauser L, Jun SR, Nookaew I, Leuze MR *et al.*** Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics* 2015;15(2):141-161.

285. **Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M *et al.*** High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol* 2012;10(9):599-606.

286. **Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P *et al.*** Artemis: sequence visualization and annotation. *Bioinformatics* 2000;16(10):944-945.

287. **Delcher AL, Harmon D, Kasif S, White O, Salzberg SL**. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 1999;27(23):4636-4641.

288. **Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C *et al.*** Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 1998;393(6685):537-544.

289. **Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V *et al.*** The complete genome sequence of *Escherichia coli* K-12. *Science* 1997;277(5331):1453-1462.

290. **Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G *et al.*** The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 1997;390(6657):249-256.

291. **Eiglmeier K, Parkhill J, Honoré N, Garnier T, Tekaia F *et al.*** The decaying genome of *Mycobacterium leprae*. *Lepr Rev* 2001;72(4):387-398.

292. **Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C *et al.*** The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* 2000;403(6770):665-668.

293. **Achtman M**. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu Rev Microbiol* 2008;62:53-70.

294. **Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K *et al.*** Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* 2001;8(1):11-22.

295. **Welch RA, Burland V, Plunkett G, Redford P, Roesch P *et al.*** Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci USA* 2002;99(26):17020-17024.

296. **Arnold KE, Williams NJ, Bennett M**. 'Disperse abroad in the land': the role of wildlife in the dissemination of antimicrobial resistance. *Biol Lett* 2016;12(8).

297. **Metzker ML**. Sequencing technologies - the next generation. *Nat Rev Genet* 2010;11(1):31-46.

298. **Paterson GK, Harrison EM, Murray GGR, Welch JJ, Warland JH *et al.*** Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission. *Nat Commun* 2015;6:6560.

299. **Bahram M, Netherway T, Frioux C, Ferretti P, Coelho LP *et al.*** Metagenomic assessment of the global distribution of bacteria and fungi. *Environ Microbiol* 2021;23(1):316-326.

300. **Human Microbiome Project Consortium**. Structure, function and diversity of the healthy human microbiome. *Nature* 2012;486(7402):207-214.

301. **Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME *et al.*** Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 2020;21(1):30.

302. **Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R *et al.*** Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat Biotechnol* 2019;37(7):783-792.

303. **Quick J, Ashton P, Calus S, Chatt C, Gossain S *et al.*** Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol* 2015;16:114.

304. **Behrmann O, Spiegel M**. COVID-19: from rapid genome sequencing to fast decisions. *Lancet Infect Dis* 2020;20(11):1218.

305. **Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z *et al.*** Whole-genome haplotyping using long reads and statistical methods. *Nature biotechnology* 2014;32(3):261-266.

306. **Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP**. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* 2014;15(2):121-132.

307. **Earl D, Bradnam K, St John J, Darling A, Lin D *et al.*** Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res* 2011;21(12):2224-2241.

308. **Nagarajan N, Pop M**. Sequence assembly demystified. *Nature Reviews Genetics*, Review Article 2013;14:157.

309. **Liu W, Wu S, Lin Q, Gao S, Ding F *et al.*** RGAAT: A Reference-based genome assembly and annotation tool for new genomes and upgrade of known genomes. *Genomics Proteomics Bioinformatics* 2018;16(5):373-381.

310. **Li H**. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34(18):3094-3100.

311. **Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX *et al.*** High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat Genet* 2008;40(8):987-993.

312. **Wick RR, Judd LM, Gorrie CL, Holt KE**. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;13(6):e1005595.

313. **Lischer HEL, Shimizu KK**. Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics* 2017;18(1):474.

314. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M *et al.*** SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19(5):455-477.

315. **Li D, Liu CM, Luo R, Sadakane K, Lam TW**. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics (Oxford, England)* 2015;31(10):1674-1676.

316. **Boetzer M, Pirovano W**. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* 2014;15:211.

317. **Zerbino DR, Birney E**. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 2008;18(5):821-829.

318. **Li H**. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* 2016;32(14):2103-2110.

319. **Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH *et al.*** Canu: scalable and accurate long-read assembly via adaptive. *Genome Res* 2017;27(5):722-736.

320. **Langmead B, Salzberg SL**. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357.

321. **Li H, Durbin R**. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* 2009;25(14):1754-1760.

322. **Seemann T**. Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)* 2014;30(14):2068-2069.

323. **Aziz RK, Bartels D, Best AA, DeJongh M, Disz T *et al.*** The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 2008;9:75.

324. **Lomsadze A, Gemayel K, Tang S, Borodovsky M**. Improved prokaryotic gene prediction yields insights into transcription and translation mechanisms on whole genome scale. *bioRxiv* 2017.

325. **Zhu W, Lomsadze A, Borodovsky M**. *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Res* 2010;38(12):e132-e132.

326. **Zolfo M, Tett A, Jousson O, Donati C, Segata N**. MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples. *Nucleic Acids Res* 2017;45(2):e7-e7.

327. **Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M *et al.*** Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat Methods* 2016;13:435.

328. **Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S *et al.*** Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics (Oxford, England)* 2015;31(22):3691-3693.

329. **Stamatakis A**. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 2006;22(21):2688-2690.

330. **Price MN, Dehal PS, Arkin AP**. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5(3):e9490.

331. **Han MV, Zmasek CM**. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* 2009;10:356.

332. **Page RD**. Visualizing phylogenetic trees using TreeView. *Curr Protoc Bioinformatics* 2002;Chapter 6:Unit 6.2.

333. **Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C *et al.*** GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res* 2018;28(9):1395-1404.

334. **Li H, Ruan J, Durbin R**. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* 2008;18(11):1851-1858.

335. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J *et al.*** The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25(16):2078-2079.

336. **Garrison E, Marth, G**. Haplotype-based variant detection from short-read sequencing. *arXiv* (Preprint) 2012 arXiv:1207.3907 [q-bio.GN].

337. **He J**. Phylogenetic Methods. In: **Schmidt TM** (ed). Encyclopedia of Microbiology (4th ed.). Academic Press; Oxford. 2019; pp. 538-550.

338. **Ajawatanawong P**. Molecular phylogenetics: Concepts for a newcomer. *Adv Biochem Eng Biotechnol* 2017;160:185-196.

339. **Gregory TR**. Understanding evolutionary trees. *Evolution: Education and Outreach* 2008; 1:121–137.

340. **Zielezinski A, Girgis HZ, Bernard G, Leimeister CA, Tang K *et al.*** Benchmarking of alignment-free sequence comparison methods. *Genome Biol* 2019;20(1):144.

341. **Didelot X**. Computational methods in microbial population genomics. Springer; Cham. 2017; pp. 1-27.

342. **De Silva D, Peters J, Cole K, Cole MJ, Cresswell F *et al.*** Whole-genome sequencing to determine transmission of *Neisseria gonorrhoeae*: an observational study. *The Lancet Infect Dis* 2016;16(11):1295-1303.

343. **Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA *et al.*** Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43(3):e15.

344. **Didelot X, Falush D**. Inference of bacterial microevolution using multilocus sequence data. *Genetics* 2007;175(3):1251-1266.

345. **Munjal G HM, Srivastava S**. Phylogenetics algorithms and applications. In: **Hu, YC, Tiwari, S, Mishra, KK, Trivedi, MC** (eds.) Ambient Communications and Computer Systems. Springer; Heidelberg, Berlin. 2018;10:904:187–994.

346. **Wheeler TJ**. Large-scale neighbor-joining with NINJA. in Algorithms in Bioinformatics. Springer; Heidelberg, Berlin. 2009.

347. **Darling AE, Mau B, Perna NT**. ProgressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLos One* 2010;5(6):e11147.

348. **Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S *et al.*** Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLOS Genetics* 2009;5(1):e1000344.

349. **Blattner FR**. Biological frontiers. *Science* 1983;222(4625):719-720.

350. **Pennisi E**. Laboratory workhorse decoded. *Science* 1997;277(5331):1432-1434.

351. **Hayashi K, Morooka N, Yamamoto Y, Fujita K, Isono K *et al.*** Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. *Mol Syst Biol* 2006;2:2006.0007.

352. **Lawrence JG, Ochman H**. Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci USA* 1998;95(16):9413-9417.

353. **Muto A, Osawa S**. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci U S A* 1987;84(1):166-169.

354. **Lawrence JG, Ochman H**. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* 1997;44(4):383-397.

355. **Lawrence JG, Ochman H**. Reconciling the many faces of lateral gene transfer. *Trends Microbiol* 2002;10(1):1-4.

356. **Perna NT, Plunkett G, Burland V, Mau B, Glasner JD *et al.*** Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 2001;409(6819):529-533.

357. **Watanabe H, Wada A, Inagaki Y, Itoh K, Tamura K**. Outbreaks of enterohaemorrhagic *Escherichia coli* O157:H7 infection by two different genotype strains in Japan, 1996. *Lancet* 1996;348(9030):831-832.

358. **Chaudhuri RR, Sebaihia M, Hobman JL, Webber MA, Leyton DL *et al.*** Complete genome sequence and comparative metabolic profiling of the prototypical enteroaggregative *Escherichia coli* strain 042. *PLoS One* 2010;5(1):e8801.

359. **Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF *et al.*** The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol* 2008;190(20):6881-6893.

360. **Mau B, Glasner JD, Darling AE, Perna NT**. Genome-wide detection and analysis of homologous recombination among sequenced strains of *Escherichia coli*. *Genome Biol* 2006;7(5):R44.

361. **Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D *et al.*** Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial “pan-genome”. *Proc Natl Acad Sci USA* 2005;102(39):13950-13955.

362. **Tettelin H, Medini D**. The pangenome: diversity, dynamics and evolution of genomes. 2020. Springer; Cham. 2020.

363. **Lukjancenko O, Wassenaar TM, Ussery DW**. Comparison of 61 sequenced *Escherichia coli* genomes. *Microb Ecol* 2010;60(4):708-720.

364. **Fukiya S, Mizoguchi H, Tobe T, Mori H**. Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* strains revealed by comparative genomic hybridization microarray. *J Bacteriol* 2004;186(12):3911-3921.

365. **Ochman H, Jones IB**. Evolutionary dynamics of full genome content in *Escherichia coli*. *EMBO J* 2000;19(24):6637-6643.

366. **Decano AG, Downing T**. An *Escherichia coli* ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. *Sci Rep* 2019;9(1):17394.

367. **McNally A, Oren Y, Kelly D, Pascoe B, Dunn S *et al.*** Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLoS Genet* 2016;12(9):e1006280.

368. **Xu J**. Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Mol Ecol* 2006;15(7):1713-1731.

369. **Nicholson JK, Holmes E, Kinross J, Burcelin R, Gibson G *et al.*** Host-gut microbiota metabolic interactions. *Science* 2012;336(6086):1262-1267.

370. **Round JL, Mazmanian SK**. The gut microbiota shapes intestinal immune responses during health and disease. *Nat Rev Immunol* 2009;9(5):313-323.

371. **Muegge BD, Kuczynski J, Knights D, Clemente JC, González A *et al.*** Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* 2011;332(6032):970-974.

372. **Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR *et al.*** Evolution of mammals and their gut microbes. *Science* 2008;320(5883):1647-1651.

373. **Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY *et al.*** Linking long-term dietary patterns with gut microbial enterotypes. *Science* 2011;334(6052):105-108.

374. **David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE *et al.*** Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 2014;505(7484):559-563.

375. **Flint HJ**. The impact of nutrition on the human microbiome. *Nutr Rev* 2012;70 Suppl 1:S10-13.

376. **Hooper LV, Macpherson AJ**. Immune adaptations that maintain homeostasis with the intestinal microbiota. *Nat Rev Immunol* 2010;10(3):159-169.

377. **Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R *et al.*** The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci Transl Med* 2009;1(6):6ra14.

378. **Claus SP, Guillou H, Ellero-Simatos S**. The gut microbiota: a major player in the toxicity of environmental pollutants? *NPJ Biofilms Microbiomes* 2016;2:16003.

379. **Kamada N, Chen GY, Inohara N, Núñez G**. Control of pathogens and pathobionts by the gut microbiota. *Nat Immunol* 2013;14(7):685-690.

380. **Brown AC, Valiere A**. Probiotics and medical nutrition therapy. *Nutr Clin Care* 2004;7(2):56-68.

381. **Tanaka M, Nakayama J**. Development of the gut microbiota in infancy and its impact on health in later life. *Allergol Int* 2017;66(4):515-522.

382. **Quince C, Ijaz UZ, Loman N, Eren AM, Saulnier D *et al.*** Extensive modulation of the fecal metagenome in children with Crohn's disease during exclusive enteral nutrition. *Am J Gastroenterol* 2015;110(12):1718-1729; quiz 1730.

383. **Codling C, O'Mahony L, Shanahan F, Quigley EM, Marchesi JR**. A molecular analysis of fecal and mucosal bacterial communities in irritable bowel syndrome. *Dig Dis Sci* 2010;55(2):392-397.

384. **Miyoshi M, Ogawa A, Higurashi S, Kadooka Y**. Anti-obesity effect of *Lactobacillus gasseri* SBT2055 accompanied by inhibition of pro-inflammatory gene expression in the visceral adipose tissue in diet-induced obese mice. *Eur J Nutr* 2014;53(2):599-606.

385. **Kang JH, Yun SI, Park MH, Park JH, Jeong SY *et al.*** Anti-obesity effect of *Lactobacillus gasseri* BNR17 in high-sucrose diet-induced obese mice. *PLoS One* 2013;8(1):e54617.

386. **Kadooka Y, Sato M, Imaizumi K, Ogawa A, Ikuyama K *et al.*** Regulation of abdominal adiposity by probiotics (*Lactobacillus gasseri* SBT2055) in adults with obese tendencies in a randomized controlled trial. *Eur J Clin Nutr* 2010;64(6):636-643.

387. **Lynch SV, Boushey HA**. The microbiome and development of allergic disease. *Curr Opin Allergy Clin Immunol* 2016;16(2):165-171.

388. **Berg RD**. The indigenous gastrointestinal microflora. *Trends Microbiol* 1996;4(11):430-435.

389. **Gordon DM, Cowling A**. The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. *Microbiology (Reading)* 2003;149(Pt 12):3575-3586.

390. **Mitsuoka T, Ohno K, Benno Y, Suzuki K, Namba K**. [The fecal flora of man. IV. Communication: Comparison of the newly developed method with the old conventional method for the analysis of intestinal flora (author's transl)]. *Zentralbl Bakteriol Orig A* 1976;234(2):219-233.

391. **Penders J, Thijs C, Vink C, Stelma FF, Snijders B *et al.*** Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics* 2006;118(2):511-521.

392. **Lescat M, Clermont O, Woerther PL, Glodt J, Dion S *et al.*** Commensal *Escherichia coli* strains in Guiana reveal a high genetic diversity with host-dependant population structure. *Environ Microbiol Rep* 2013;5(1):49-57.

393. **Poulsen LK, Lan F, Kristensen CS, Hobolth P, Molin S *et al.*** Spatial distribution of *Escherichia coli* in the mouse large intestine inferred from rRNA in situ hybridization. *Infect Immun* 1994;62(11):5191-5194.

394. **Poulsen LK, Licht TR, Rang C, Krogfelt KA, Molin S**. Physiological state of *Escherichia coli* BJ4 growing in the large intestines of streptomycin-treated mice. *J Bacteriol* 1995;177(20):5840-5845.

395. **Bettelheim KA, Faiers M, Shooter RA**. Serotypes of *Escherichia coli* in normal stools. *Lancet* 1972;2(7789):1223-1224.

396. **Slanetz LW, Bartley CH**. Numbers of enterococci in water, sewage, and feces determined by the membrane filter technique with an improved medium. *J Bacteriol* 1957;74(5):591-595.

397. **Escherich T**. Die darmbakterien des däuglings und ihre beziehungen zur physiologie der verdauung. Verlag von Ferdinand Enke, Stuttgart 1886.

398. **Mueller NT, Bakacs E, Combellick J, Grigoryan Z, Dominguez-Bello MG**. The infant microbiome development: Mom matters. *Trends Mol Med* 2015;21(2):109-117.

399. **Syed SA, Abrams GD, Freter R**. Efficiency of various intestinal bacteria in assuming normal functions of enteric flora after association with germ-free mice. *Infect Immun* 1970;2(4):376-386.

400. **Richter TKS, Michalski JM, Zanetti L, Tennant SM, Chen WH *et al.*** Responses of the human gut. *mSystems* 2018;3(4).

401. **Suvarna K, Stevenson D, Meganathan R, Hudspeth ME**. Menaquinone (vitamin K2) biosynthesis: localization and characterization of the *menA* gene from *Escherichia coli*. *J Bacteriol* 1998;180(10):2782-2787.

402. **Martinson JNV, Walk ST**. Residency in the gut of healthy human adults. *EcoSal Plus* 2020;9(1).

403. **Abia ALK, Ubomba-Jaswa E**. Dirty money on holy ground: Isolation of potentially pathogenic bacteria and fungi on money collected from church offerings. *Iran J Public Health* 2019;48(5):849-857.

404. **Akoachere JF, Gaelle N, Dilonga HM, Nkuo-Akenji TK**. Public health implications of contamination of Franc CFA (XAF) circulating in Buea (Cameroon) with drug resistant pathogens. *BMC Res Notes* 2014;7:16.

405. **Nwankwo EO, Ekwunife N, Mofolorunsho KC**. Nosocomial pathogens associated with the mobile phones of healthcare workers in a hospital in Anyigba, Kogi state, Nigeria. *J Epidemiol Glob Health* 2014;4(2):135-140.

406. **Pal S, Juyal D, Adekhandi S, Sharma M, Prakash R *et al.*** Mobile phones: Reservoirs for the transmission of nosocomial pathogens. *Adv Biomed Res* 2015;4:144.

407. **Reeves PR, Liu B, Zhou Z, Li D, Guo D *et al.*** Rates of mutation and host transmission for an *Escherichia coli* clone over 3 years. *PLos One* 2011;6(10):e26907-e26907.

408. **Derakhshandeh A, Eraghi V, Boroojeni AM, Niaki MA, Zare S *et al.*** Virulence factors, antibiotic resistance genes and genetic relatedness of commensal *Escherichia coli* isolates from dogs and their owners. *Microb Pathog* 2018;116:241-245.

409. **Toombs-Ruane LJ, Benschop J, French NP, Biggs PJ, Midwinter AC *et al.*** Carriage of extended-spectrum-beta-lactamase- and AmpC beta-lactamase-producing *Escherichia coli* strains from humans and pets in the same households. *Appl Environ Microbiol* 2020;86(24).

410. **Ercumen A, Pickering AJ, Kwong LH, Arnold BF, Parvez SM *et al.*** Animal feces contribute to domestic fecal contamination: Evidence from *E. coli* measured in water, hands, food, flies, and soil in Bangladesh. *Environ Sci Technol* 2017;51(15):8725-8734.

411. **Aijuka M, Santiago AE, Girón JA, Nataro JP, Buys EM**. Enteroaggregative *Escherichia coli* is the predominant diarrheagenic *E. coli* pathotype among irrigation water and food sources in South Africa. *Int J Food Microbiol* 2018;278:44-51.

412. **Yar DD**. Bacterial contaminants and antibiogram of Ghana paper currency notes in circulation and their associated health risks in Asante-Mampong, Ghana. *Int J Microbiol* 2020;2020:8833757.

413. **Dixit OVA, O'Brien CL, Pavli P, Gordon DM**. Within-host evolution versus immigration as a determinant of *Escherichia coli* diversity in the human gastrointestinal tract. *Environ Microbiol* 2018;20(3):993-1001.

414. **Richter TKS, Hazen TH, Lam D, Coles CL, Seidman JC *et al.*** Temporal variability of *Escherichia coli* diversity in the gastrointestinal tracts of Tanzanian children with and without exposure to antibiotics. mSphere 2018;3(6).

415. **Shooter RA, Bettleheim KA, Lennox-King SM, O'Farrell S**. *Escherichia coli* serotypes in the faeces of healthy adults over a period of several months. *J Hyg (Lond)* 1977;78(1):95-98.

416. **Wiedemann B, Habermann R, Knothe H, Ihrig L**. [Stablility of *Escherichia coli*-flora in healthy men. 3. Occurrence of permanent and transitory strains in infants]. *Arch Hyg Bakteriol* 1971;154(6):581-589.

417. **Sears HJ, Brownlee I, Uchiyama JK**. Persistence of individual strains of *Escherichia coli* in the intestinal tract of man. *J Bacteriol* 1950;59(2):293-301.

418. **Sears HJ, Brownlee I**. Further observations on the persistence of individual strains of *Escherichia coli* in the intestinal tract of man. *J Bacteriol* 1952;63(1):47-57.

419. **Tannock GW, Tiong IS, Priest P, Munro K, Taylor C *et al.*** Testing probiotic strain *Escherichia coli* Nissle 1917 (Mutaflor) for its ability to reduce carriage of multidrug-resistant *E. coli* by elderly residents in long-term care facilities. *J Med Microbiol* 2011;60(Pt 3):366-370.

420. **Johnson JR, Clabots C**. Sharing of virulent *Escherichia coli* clones among household members of a woman with acute cystitis. *Clin Infect Dis* 2006;43(10):e101-108.

421. **Freter R, Brickner H, Botney M, Cleven D, Aranki A**. Mechanisms that control bacterial populations in continuous-flow culture models of mouse large intestinal flora. *Infect Immun* 1983;39(2):676-685.

422. **Freter R, Brickner H, Fekete J, Vickerman MM, Carey KE**. Survival and implantation of *Escherichia coli* in the intestinal tract. *Infect Immun* 1983;39(2):686-703.

423. **Pereira FC, Berry D**. Microbial nutrient niches in the gut. *Environ Microbiol* 2017;19(4):1366-1378.

424. **Conway T, Cohen PS**. Commensal and pathogenic *Escherichia coli* metabolism in the gut. *Microbiol Spectr* 2015;3(3).

425. **Maltby R, Leatham-Jensen MP, Gibson T, Cohen PS, Conway T**. Nutritional basis for colonization resistance by human commensal *Escherichia coli* strains HS and Nissle 1917 against *E. coli* O157:H7 in the mouse intestine. *PLoS One* 2013;8(1):e53957.

426. **Hardin G**. The competitive exclusion principle. *Science* 1960;131(3409):1292-1297.

427. **Nowrouzian F, Adlerberth I, Wold AE**. P fimbriae, capsule and aerobactin characterize colonic resident *Escherichia coli*. *Epidemiol Infect* 2001;126(1):11-18.

428. **Granato ET, Meiller-Legrand TA, Foster KR**. The evolution and ecology of bacterial warfare. *Curr Biol* 2019;29(11):R521-R537.

429. **Ghoul M, Mitri S**. The ecology and evolution of microbial competition. *Trends Microbiol* 2016;24(10):833-845.

430. **Fabich AJ, Leatham MP, Grissom JE, Wiley G, Lai H *et al.*** Genotype and phenotypes of an intestine-adapted *Escherichia coli* K-12 mutant selected by animal passage for superior colonization. *Infect Immun* 2011;79(6):2430-2439.

431. **Leatham MP, Stevenson SJ, Gauger EJ, Krogfelt KA, Lins JJ *et al.*** Mouse intestine selects nonmotile *flhDC* mutants of *Escherichia coli* MG1655 with increased colonizing ability and better utilization of carbon sources. *Infect Immun* 2005;73(12):8039-8049.

432. **Gauger EJ, Leatham MP, Mercado-Lubo R, Laux DC, Conway T *et al.*** Role of motility and the *flhDC* operon in *Escherichia coli* MG1655 colonization of the mouse intestine. *Infect Immun* 2007;75(7):3315-3324.

433. **Leatham-Jensen MP, Frimodt-Moller J, Adediran J, Mokszycki ME, Banner ME *et al.*** The streptomycin-treated mouse intestine selects *Escherichia coli envZ* missense mutants that interact with dense and diverse intestinal microbiota. *Infect Immun* 2012;80(5):1716-1727.

434. **Ng KM, Ferreyra JA, Higginbottom SK, Lynch JB, Kashyap PC *et al.*** Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens. *Nature* 2013;502(7469):96-99.

435. **Dolfing J**. Syntrophy in microbial fuel cells. *ISME J* 2014;8(1):4-5.

436. **Tuncil YE, Xiao Y, Porter NT, Reuhs BL, Martens EC *et al.*** Reciprocal prioritization to dietary glycans by gut bacteria in a competitive environment promotes stable coexistence. *mBio* 2017;8(5).

437. **Fabich AJ, Jones SA, Chowdhury FZ, Cernosek A, Anderson A *et al.*** Comparison of carbon nutrition for pathogenic and commensal *Escherichia coli* strains in the mouse intestine. *Infect Immun* 2008;76(3):1143-1152.

438. **Leatham MP, Banerjee S, Autieri SM, Mercado-Lubo R, Conway T *et al.*** Precolonized human commensal *Escherichia coli* strains serve as a barrier to *E. coli* O157:H7 growth in the streptomycin-treated mouse intestine. *Infect Immun* 2009;77(7):2876-2886.

439. **Naylor SW, Low JC, Besser TE, Mahajan A, Gunn GJ et al.** Lymphoid follicle-dense mucosa at the terminal rectum is the principal site of colonization of enterohemorrhagic *Escherichia coli* O157:H7 in the bovine host. *Infect Immun* 2003;71(3):1505-1512.

440. **Pamer EG**. Resurrecting the intestinal microbiota to combat antibiotic-resistant pathogens. *Science* 2016;352(6285):535-538.

441. **Buffie CG, Pamer EG**. Microbiota-mediated colonization resistance against intestinal pathogens. *Nat Rev Immunol* 2013;13(11):790-801.

442. **Ravi A, Halstead FD, Bamford A, Casey A, Thomson NM *et al.*** Loss of microbial diversity and pathogen domination of the gut microbiota in critically ill patients. *Microb Genom* 2019;5(9).

443. **Buffie CG, Jarchum I, Equinda M, Lipuma L, Gobourne A *et al.*** Profound alterations of intestinal microbiota following a single dose of clindamycin results in sustained susceptibility to *Clostridium difficile*-induced colitis. *Infect Immun* 2012;80(1):62-73.

444. **Dethlefsen L, Huse S, Sogin ML, Relman DA**. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol* 2008;6(11):e280.

445. **Ribeiro CFA, Silveira GGOS, Cândido ES, Cardoso MH, Espínola Carvalho CM *et al.*** Effects of antibiotic treatment on gut microbiota and how to overcome its negative impacts on human health. *ACS Infect Dis* 2020;6(10):2544-2559.

446. **Morrow LE, Wischmeyer P**. Blurred lines: Dysbiosis and probiotics in the ICU. *Chest* 2017;151(2):492-499.

447. **Gough E, Shaikh H, Manges AR**. Systematic review of intestinal microbiota transplantation (fecal bacteriotherapy) for recurrent *Clostridium difficile* infection. *Clin Infect Dis* 2011;53(10):994-1002.

448. **Manges AR, Steiner TS, Wright AJ**. Fecal microbiota transplantation for the intestinal decolonization of extensively antimicrobial-resistant opportunistic pathogens: a review. *Infect Dis (Lond)* 2016;48(8):587-592.

449. **Youngster I, Russell GH, Pindar C, Ziv-Baran T, Sauk J *et al.*** Oral, capsulized, frozen fecal microbiota transplantation for relapsing *Clostridium difficile* infection. *JAMA* 2014;312(17):1772-1778.

450. **Hill C, Guarner F, Reid G, Gibson GR, Merenstein DJ *et al.*** Expert consensus document. The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic. *Nat Rev Gastroenterol Hepatol* 2014;11(8):506-514.

451. **Gorbach SL**. Probiotics and gastrointestinal health. *Am J Gastroenterol* 2000;95(1 Suppl):S2-4.

452. **Sivan A, Corrales L, Hubert N, Williams JB, Aquino-Michaels K *et al.*** Commensal *Bifidobacterium* promotes antitumor immunity and facilitates anti-PD-L1 efficacy. *Science* 2015;350(6264):1084-1089.

453. **Jacobi CA, Malfertheiner P**. *Escherichia coli* Nissle 1917 (Mutaflor): new insights into an old probiotic bacterium. *Dig Dis* 2011;29(6):600-607.

454. **Nissle A**. [Mutaflor and its medical significance]. *Z Klin Med* 1951;2(3-4):68.

455. **Nissle A**. [Explanations of the significance of colonic dysbacteria & the mechanism of action of *E. coli* therapy (mutaflor)]. *Medizinische* 1959;4(21):1017-1022.

456. **Nissle A**. [Old and new experiences on therapeutic successes by restoration of the colonic flora with mutaflor in gastrointestinal diseases]. *Med Welt* 1961;29-30:1519-1523.

457. **Hancock V, Dahl M, Klemm P**. Probiotic *Escherichia coli* strain Nissle 1917 outcompetes intestinal pathogens during biofilm formation. *J Med Microbiol* 2010;59(Pt 4):392-399.

458. **Lodinová-Zádniková R, Sonnenborn U**. Effect of preventive administration of a nonpathogenic *Escherichia coli* strain on the colonization of the intestine with microbial pathogens in newborn infants. *Biol Neonate* 1997;71(4):224-232.

459. **Altenhoefer A, Oswald S, Sonnenborn U, Enders C, Schulze J *et al.*** The probiotic *Escherichia coli* strain Nissle 1917 interferes with invasion of human intestinal epithelial cells by different enteroinvasive bacterial pathogens. *FEMS Immunol Med Microbiol* 2004;40(3):223-229.

460. **Boudeau J, Glasser AL, Julien S, Colombel JF, Darfeuille-Michaud A**. Inhibitory effect of probiotic *Escherichia coli* strain Nissle 1917 on adhesion to and invasion of intestinal epithelial cells by adherent-invasive *E. coli* strains isolated from patients with Crohn's disease. *Aliment Pharmacol Ther* 2003;18(1):45-56.

461. **Zyrek AA, Cichon C, Helms S, Enders C, Sonnenborn U *et al.*** Molecular mechanisms underlying the probiotic effects of *Escherichia coli* Nissle 1917 involve ZO-2 and PKCzeta redistribution resulting in tight junction and epithelial barrier repair. *Cell Microbiol* 2007;9(3):804-816.

462. **Wehkamp J, Harder J, Wehkamp K, Wehkamp-von Meissner B, Schlee M *et al.*** NF-kappaB- and AP-1-mediated induction of human beta defensin-2 in intestinal epithelial cells by *Escherichia coli* Nissle 1917: a novel effect of a probiotic bacterium. *Infect Immun* 2004;72(10):5750-5758.

463. **Timmerman HM, Niers LE, Ridwan BU, Koning CJ, Mulder L *et al.*** Design of a multispecies probiotic mixture to prevent infectious complications in critically ill patients. *Clin Nutr* 2007;26(4):450-459.

464. **Venturi A, Gionchetti P, Rizzello F, Johansson R, Zucconi E *et al.*** Impact on the composition of the faecal flora by a new probiotic preparation: preliminary data on maintenance treatment of patients with ulcerative colitis. *Aliment Pharmacol Ther* 1999;13(8):1103-1108.

465. **Meador JP, Caldwell ME, Cohen PS, Conway T**. *Escherichia coli* pathotypes occupy distinct niches in the mouse intestine. *Infection and immunity* 2014;82(5):1931-1938.

466. **Bryant J, Chewapreecha C, Bentley SD**. Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. *Future Microbiol* 2012;7(11):1283-1296.

467. **Ochman H, Lawrence JG, Groisman EA**. Lateral gene transfer and the nature of bacterial innovation. *Nature* 2000;405(6784):299-304.

468. **Hacker J, Carniel E**. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep* 2001;2(5):376-381.

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

470. **Cao Q, Didelot X, Wu Z, Li Z, He L *et al.*** Progressive genomic convergence of two *Helicobacter pylori* strains during mixed infection of a patient with chronic gastritis. *Gut* 2015;64(4):554-561.

471. **Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B *et al.*** *Helicobacter pylori* genome evolution during human infection. *Proc Natl Acad Sci USA* 2011;108(12):5033-5038.

472. **Didelot X, Nell S, Yang I, Woltemate S, van der Merwe S *et al.*** Genomic evolution and transmission of *Helicobacter pylori* in two South African families. *Proc Natl Acad Sci USA* 2013;110(34):13880-13885.

473. **Andersson JO, Andersson SG**. Insights into the evolutionary process of genome degradation. *Curr Opin Genet Dev* 1999;9(6):664-671.

474. **Rodgers K, McVey M**. Error-prone repair of DNA double-strand breaks. *J Cell Physiol* 2016;231(1):15-24.

475. **Periwal V, Scaria V**. Insights into structural variations and genome rearrangements in prokaryotic genomes. *Bioinformatics (Oxford, England)* 2015;31(1):1-9.

476. **Charlesworth B**. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 2009;10(3):195-205.

477. **Kuo CH, Moran NA, Ochman H**. The consequences of genetic drift for bacterial genome complexity. *Genome Res* 2009;19(8):1450-1454.

478. **Mather AE, Harris SR**. Playing fast and loose with mutation. *Nat Rev Microbiol* 2013;11(12):822.

479. **Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM *et al.*** Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *NEJM* 2012;366(24):2267-2275.

480. **Wang S, Wu C, Shen J, Wu Y, Wang Y**. Hypermutable *Staphylococcus aureus* strains present at high frequency in subclinical bovine mastitis isolates are associated with the development of antibiotic resistance. *Vet Microbiol* 2013;165(3-4):410-415.

481. **Marvig RL, Johansen HK, Molin S, Jelsbak L**. Genome analysis of a transmissible lineage of *Pseudomonas aeruginosa* reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators. *PLoS Genet* 2013;9(9):e1003741.

482. **Golubchik T, Batty EM, Miller RR, Farr H, Young BC *et al.*** Within-host evolution of *Staphylococcus aureus* during asymptomatic carriage. *PLoS One* 2013;8(5):e61319.

483. **Hall MD, Holden MT, Srisomang P, Mahavanakul W, Wuthiekanun V *et al.*** Improved characterisation of MRSA transmission using within-host bacterial sequence diversity. *Elife* 2019;8.

484. **Worby CJ, Lipsitch M, Hanage WP**. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS Comput Biol* 2014;10(3):e1003549.

485. **Ley SD, de Vos M, Van Rie A, Warren RM**. Deciphering within-host microevolution of *Mycobacterium tuberculosis* through whole-genome sequencing: the phenotypic impact and way forward. *Microbiol Mol Biol Rev* 2019;83(2).

486. **Ssengooba W, de Jong BC, Joloba ML, Cobelens FG, Meehan CJ**. Whole genome sequencing reveals mycobacterial microevolution among concurrent isolates from sputum and blood in HIV infected TB patients. *BMC Infect Dis* 2016;16:371.

487. **Trauner A, Liu Q, Via LE, Liu X, Ruan X *et al.*** The within-host population dynamics of *Mycobacterium tuberculosis* vary with treatment efficacy. *Genome Biol* 2017;18(1):71.

488. **Herranz M, Pole I, Ozere I, Chiner-Oms Á, Martínez-Lirola M *et al.*** *Mycobacterium tuberculosis* acquires limited genetic diversity in prolonged infections, reactivations and transmissions involving multiple hosts. *Front Microbiol* 2017;8:2661.

489. **Kay GL, Sergeant MJ, Zhou Z, Chan JZ, Millard A *et al.*** Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat Commun* 2015;6:6717.

490. **Halachev MR, Chan JZM, Constantinidou CI, Cumley N, Bradley C *et al.*** Genomic epidemiology of a protracted hospital outbreak caused by multidrug-resistant *Acinetobacter baumannii* in Birmingham, England. *Genome Medicine* 2014;6(11):70.

491. **Wymant C, Hall M, Ratmann O, Bonsall D, Golubchik T *et al.*** PHYLOSCANNER: Inferring transmission from within- and between-host pathogen genetic diversity. *Mol Biol Evol* 2018;35(3):719-733.

492. **Cornelissen M, Pasternak AO, Grijsen ML, Zorgdrager F, Bakker M *et al.*** HIV-1 dual infection is associated with faster CD4+ T-cell decline in a cohort of men with primary HIV infection. *Clin Infect Dis* 2012;54(4):539-547.

493. **Branche WC, Young VM, Robinet HG, Massey ED**. Effect of colicine production on *Escherichia coli* in the normal human intestine. *Proc Soc Exp Biol Med* 1963;114:198-201.

494. **Tzabar Y, Pennington TH**. The population structure and transmission of *Escherichia coli* in an isolated human community; studies on an Antarctic base. *Epidemiol Infect* 1991;107(3):537-542.

495. **Damborg P, Nielsen SS, Guardabassi L**. *Escherichia coli* shedding patterns in humans and dogs: insights into within-household transmission of phylotypes associated with urinary tract infections. *Epidemiol Infect* 2009;137(10):1457-1464.

496. **Martinson JNV, Pinkham NV, Peters GW, Cho H, Heng J *et al.*** Rethinking gut microbiome residency and the *Enterobacteriaceae* in healthy human adults. *ISME J* 2019; 13(9):2306-2318.

497. **Blyton MD, Banks SC, Peakall R, Gordon DM**. High temporal variability in commensal *Escherichia coli* strain communities of a herbivorous marsupial. *Environ Microbiol* 2013;15(8):2162-2172.

498. **Stoesser N, Sheppard AE, Moore CE, Golubchik T, Parry CM *et al.*** Extensive within-host diversity in fecally carried extended-spectrum-beta-lactamase-producing *Escherichia coli* isolates: Implications for transmission analyses. *J Clin Microbiol* 2015;53(7):2122-2131.

499. **Li SP, Tan J, Yang X, Ma C, Jiang L**. Niche and fitness differences determine invasion success and impact in laboratory bacterial communities. *ISME J* 2019;13(2):402-412.

500. **Knudsen PK, Brandtzaeg P, Høiby EA, Bohlin J, Samuelsen Ø et al.** Impact of extensive antibiotic treatment on faecal carriage of antibiotic-resistant enterobacteria in children in a low resistance prevalence setting. *PLoS One* 2017;12(11):e0187618.

501. **Knudsen PK, Gammelsrud KW, Alfsnes K, Steinbakk M, Abrahamsen TG *et al.*** Transfer of a *$bla_{CTX-M-1}$*-carrying plasmid between different *Escherichia coli* strains within the human gut explored by whole genome sequencing analyses. *Sci Rep* 2018;8(1):280.

502. **Stegger M, Leihof RF, Baig S, Sieber RN, Thingholm KR *et al.*** A snapshot of diversity: Intraclonal variation of *Escherichia coli* clones as commensals and pathogens. *Int J Med Microbiol* 2020;310(3):151401.

503. **Nielsen KL, Dynesen P, Larsen P, Frimodt-Møller N**. Faecal *Escherichia coli* from patients with *E. coli* urinary tract infection and healthy controls who have never had a urinary tract infection. *J Med Microbiol* 2014;63(Pt 4):582-589.

504. **Neidhardt FC, Eingraham JL, Low KB, Magasanik B, Schaechter M & Umbarger HE** (eds). *Escherichia coli* and *Salmonella typhimurium: Cellular and Molecular Biology.* ASM Press; Washington, DC. 1996.

505. **Schlager TA, Hendley JO, Bell AL, Whittam TS**. Clonal diversity of *Escherichia coli* colonizing stools and urinary tracts of young girls. *Infect Immun* 2002;70(3):1225-1229.

506. **Lidin-Janson G, Kaijser B, Lincoln K, Olling S, Wedel H**. The homogeneity of the faecal coliform flora of normal school-girls, characterized by serological and biochemical properties. *Med Microbiol Immunol* 1978;164(4):247-253.

507. **Connor TR, Loman NJ, Thompson S, Smith A, Southgate J *et al.*** CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *Microb Genom* 2016;2(9):e000086.

508. **Wingett SW, Andrews S**. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res* 2018;7:1338-1338.

509. **Bolger AM, Lohse M, Usadel B**. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* 2014;30(15):2114-2120.

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

511. **Wirth T, Falush D, Lan R, Colles F, Mensa P *et al.*** Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* 2006;60(5):1136-1151.

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

513. **Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA *et al.*** Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43(3):e15-e15.

514. **Tamura K, Stecher G, Peterson D, Filipski A, Kumar S**. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* 2013;30(12):2725-2729.

515. **Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH *et al.*** Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *The Lancet* 2013;382(9888):209-222.

516. **Liu J, Platts-Mills JA, Juma J, Kabir F, Nkeze J *et al.*** Use of quantitative molecular diagnostic methods to identify causes of diarrhoea in children: a reanalysis of the GEMS case-control study. *Lancet* 2016;388(10051):1291-1301.

517. **Hunt M, Mather AE, Sanchez-Buso L, Page AJ, Parkhill J *et al.*** ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genom* 2017;3(10):e000131.

518. **Liu B, Zheng D, Jin Q, Chen L, Yang J**. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res* 2019;47(D1):D687-D692.

519. **Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S *et al.*** Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* 2012;67(11):2640-2644.

520. **Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O *et al.*** *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 2014;58(7):3895-3903.

521. **Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA**. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* 2012;28(4):464-469.

522. **Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW**. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;25(7):1043-1055.

523. **Arndt D, Grant JR, Marcu A, Sajed T, Pon A *et al.*** PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;44(W1):W16-21.

524. **Wiegand I, Hilpert K, Hancock RE**. Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat Protoc* 2008;3(2):163-175.

525. **Sousa CP**. The versatile strategies of *Escherichia coli* pathotypes: a mini review. *J Venom Anim Toxins incl Trop Dis* 2006; 12(3):363-373.

526. **Clayton JB, Danzeisen JL, Trent AM, Murphy T, Johnson TJ**. Longitudinal characterization of *Escherichia coli* in healthy captive non-human primates. *Front Vet Sci* 2014;1:24.

527. **Feng Y, Mannion A, Madden CM, Swennes AG, Townes C *et al.*** Cytotoxic *Escherichia coli* strains encoding colibactin and cytotoxic necrotizing factor (CNF) colonize laboratory macaques. *Gut Pathog* 2017;9:71.

528. **Thomson JA, Scheffler JJ**. Hemorrhagic typhlocolitis associated with attaching and effacing *Escherichia coli* in common marmosets. *Lab Anim Sci* 1996;46(3):275-279.

529. **Mansfield KG, Lin KC, Newman J, Schauer D, MacKey J *et al.*** Identification of enteropathogenic *Escherichia coli* in simian immunodeficiency virus-infected infant and adult rhesus macaques. *J Clin Microbiol* 2001;39(3):971-976.

530. **Mansfield KG, Lin KC, Xia D, Newman JV, Schauer DB *et al.*** Enteropathogenic *Escherichia coli* and ulcerative colitis in cotton-top tamarins (*Saguinus oedipus*). *J Infect Dis* 2001;184(6):803-807.

531. **Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI**. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol* 2008;6(10):776-788.

532. **Moeller AH, Caro-Quintero A, Mjungu D, Georgiev AV, Lonsdorf EV *et al.*** Cospeciation of gut microbiota with hominids. *Science* 2016;353(6297):380-382.

533. **Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D *et al.*** Meta-analyses of studies of the human microbiota. *Genome Res* 2013;23(10):1704-1714.

534. **Iovine ReO, Dejuste C, Miranda F, Filoni C, Bueno MG *et al.*** Isolation of *Escherichia coli* and *Salmonella* spp. from free-ranging wild animals. *Braz J Microbiol* 2015;46(4):1257-1263.

535. **Weiss D, Wallace RM, Rwego IB, Gillespie TR, Chapman CA *et al.*** Antibiotic-resistant *Escherichia coli* and class 1 integrons in humans, domestic animals, and wild primates in rural Uganda. *Appl Environ Microbiol* 2018;84(21).

536. **Bublitz DC, Wright PC, Rasambainarivo FT, Arrigo-Nelson SJ, Bodager JR *et al.*** Pathogenic enterobacteria in lemurs associated with anthropogenic disturbance. *Am J Primatol* 2015;77(3):330-337.

537. **Senghore M, Bayliss SC, Kwambana-Adams BA, Foster-Nyarko E, Manneh J *et al.*** Transmission of *Staphylococcus aureus* from humans to green monkeys in the Gambia as revealed by whole-genome sequencing. *Appl Environ Microbiol* 2016;82(19):5910-7.

538. **Goldberg TL, Gillespie TR, Rwego IB, Estoff EL, Chapman CA**. Forest fragmentation as cause of bacterial transmission among nonhuman primates, humans, and livestock, Uganda. *Emerg Infect Dis* 2008;14(9):1375-1382.

539. **Rwego IB, Isabirye-Basuta G, Gillespie TR, Goldberg TL**. Gastrointestinal bacterial transmission among humans, mountain gorillas, and livestock in Bwindi Impenetrable National Park, Uganda. *Conserv Biol* 2008;22(6):1600-1607.

540. **Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O**. ClermonTyping: an easy-to-use and accurate *in silico* method for *Escherichia* genus strain phylotyping. *Microb Genom* 2018;4(7).

541. **Picard B, Garcia JS, Gouriou S, Duriez P, Brahimi N *et al.*** The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. *Infect Immun* 1999;67(2):546-553.

542. **Matsui Y, Hu Y, Rubin J, de Assis RS, Suh J *et al.*** Multilocus sequence typing of *Escherichia coli* isolates from urinary tract infection patients and from fecal samples of healthy subjects in a college community. *Microbiologyopen* 2020;9(6):1225-1233.

543. **Ho W-S, Gan H-M, Yap K-P, Balan G, Yeo CC *et al.*** Genome sequence of multidrug-resistant *Escherichia coli* EC302/04, isolated from a human tracheal aspirate. *J Bacteriol* 2012;194(23):6691-6692.

544. **Manges AR, Johnson JR**. Reservoirs of extraintestinal pathogenic *Escherichia coli*. *Microbiology Spectrum* 2015;3(5).

545. **Kamjumphol W, Wongboot W, Suebwongsa N, Kluabwang P, Chantaroj S *et al.*** Draft genome sequence of a colistin-resistant *Escherichia coli* ST226: A clinical strain harbouring an mcr-1 variant. *J Glob Antimicrob Resist* 2019;16:168-169.

546. **Markovska R, Stoeva T, Boyanova L, Stankova P, Schneider I *et al.*** Multicentre investigation of carbapenemase-producing *Klebsiella pneumoniae* and *Escherichia coli* in Bulgarian hospitals – Interregional spread of ST11 NDM-1-producing *K. pneumoniae*. *Infect Genet Evol* 2019;69:61-67.

547. **Salinas L, Cárdenas P, Johnson TJ, Vasco K, Graham J *et al.*** Diverse commensal *Escherichia coli* clones and plasmids disseminate antimicrobial resistance genes in domestic animals and children in a semirural community in Ecuador. *mSphere* 2019;4(3):e00316-00319.

548. **Manges AR, Geum HM, Guo A, Edens TJ, Fibke CD *et al.*** Global Extraintestinal Pathogenic *Escherichia coli* (ExPEC) lineages. *Clin Microbiology Rev* 2019;32(3):e00135-00118.

549. **Krieger JN, Dobrindt U, Riley DE, Oswald E**. Acute *Escherichia coli* prostatitis in previously healthy young men: bacterial virulence factors, antimicrobial resistance, and clinical outcomes. *Urology* 2011;77(6):1420-1425.

550. **Faïs T, Delmas J, Barnich N, Bonnet R, Dalmasso G**. Colibactin: more than a new bacterial toxin. *Toxins (Basel)* 2018;10(4).

551. **Johnson JR, Johnston B, Kuskowski MA, Nougayrede JP, Oswald E**. Molecular epidemiology and phylogenetic distribution of the *Escherichia coli pks* genomic island. *J Clin Microbiol* 2008;46(12):3906-3911.

552. **Zogg AL, Zurfluh K, Schmitt S, Nuesch-Inderbinen M, Stephan R**. Antimicrobial resistance, multilocus sequence types and virulence profiles of ESBL producing and non-ESBL producing uropathogenic *Escherichia coli* isolated from cats and dogs in Switzerland. *Veterinary Microbiology* 2018;216:79-84.

553. **Keusch GT, Pappaioanou M, Gonzalez MC, National Research Council (US) Committee on Achieving Sustainable Global Capacity for Surveillance and Response to Emerging Diseases of Zoonotic Origin**. *Sustaining Global Surveillance and Response to Emerging Zoonotic Diseases,* 3. Washington, DC: National Academies Press (US); 2009.

554. **Ewers C, Grobbel M, Stamm I, Kopp PA, Diehl I *et al.*** Emergence of human pandemic O25:H4-ST131 CTX-M-15 extended-spectrum-beta-lactamase-producing *Escherichia coli* among companion animals. *J Antimicrob Chemother* 2010;65(4):651-660.

555. **Martinson JNV, Pinkham NV, Peters GW, Cho H, Heng J *et al.*** Rethinking gut microbiome residency and the *Enterobacteriaceae* in healthy human adults. *ISME J* 2019;13(9):2306-2318.

556. **Wang J, Ma ZB, Zeng ZL, Yang XW, Huang Y *et al.*** The role of wildlife (wild birds) in the global transmission of antimicrobial resistance genes. *Zool Res* 2017;38(2):55-80.

557. **Carroll D, Wang J, Fanning S, McMahon BJ**. Antimicrobial resistance in wildlife: Implications for public health. *Zoonoses Public Health* 2015;62(7):534-542.

558. **Ramey AM, Ahlstrom CA**. Antibiotic resistant bacteria in wildlife: perspectives on trends, acquisition and dissemination, data gaps, and future directions. *J Wildl Dis* 2020;56(1):1-15.

559. **Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC**. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom* 2017;3(10):e000128.

560. **Carattoli A**. Resistance plasmid families in *Enterobacteriaceae*. *Antimicrob Agents Chemother* 2009;53(6):2227-2238.

561. **Johnson TJ, Nolan LK**. Pathogenomics of the virulence plasmids of *Escherichia coli*. *Microbiol Mol Biol Rev* 2009;73(4):750-774.

562. **Bennett CE, Thomas R, Williams M, Zalasiewicz J, Edgeworth M *et al.*** The broiler chicken as a signal of a human reconfigured biosphere. *R Soc Open Sci*;5(12):180325.

563. **Storey AA, Athens JS, Bryant D, Carson M, Emery K *et al.*** Investigating the global dispersal of chickens in prehistory using ancient mitochondrial DNA signatures. *PloS One* 2012;7(7):e39171.

564. **Miao YW, Peng MS, Wu GS, Ouyang YN, Yang ZY *et al.*** Chicken domestication: an updated perspective based on mitochondrial genomes. *Heredity* 2013;110(3):277-282.

565. **Alders RG, Pym RAE**. Village poultry: Still important to millions, eight thousand years after domestication. *Worlds Poult Sci J* 2009;65:181-190.

566. **Alders RG, Dumas SE, Rukambile E, Magoke G, Maulaga W *et al.*** Family poultry: Multiple roles, systems, challenges, and options for sustainable contributions to household nutrition security through a planetary health lens. *Matern Child Nutr* 2018;14 Suppl 3:e12668.

567. **Food and Agriculture Organization of the United Nations (FAO)**. *Recommendations on the Prevention, Control and Eradication of Highly Pathogenic Avian Influenza in Asia*. Rome, Italy: FAO Position Paper. September; 2004.

568. **Olaniyan OF, Camara S**. Rural household chicken management and challenges in the Upper River Region of the Gambia. *Trop Anim Health Prod* 2018;50:1921–1928.

569. **Alders R, Costa R, Gallardo RA, Sparks N, Zhou H**. Smallholder poultry: contributions to food and nutrition security. In**: Ferranti P, Berry EM, Anderson JR** (editors). *Encyclopedia of Food Security and Sustainability. Elsevier; Oxford.* 2019:292–298.

570. **Kilonzo-Nthenge A, Nahashon SN, Chen F, Adefope N**. Prevalence and antimicrobial resistance of pathogenic bacteria in chicken and guinea fowl. *Poult Sci* 2008;87(9):1841-1848.

571. **Samanta I, Joardar SN, Das PK**. Biosecurity strategies for backyard poultry: A controlled way for safe food production. *Food Control and Biosecurity*. 2018;481-517.

572. **Food and Agriculture Organization of the United Nations (FAO)**. *Small-scale Poultry Production Technical Guide*. Rome, Italy. 2004.

573. **Alam J**. Impact of smallholder livestock development project in some selected areas of rural Bangladesh. *Livestock Research for Rural Development*. 1997;9(3); Article no. 25. Retrieved May 4, 2020, from http://www.lrrd.org/lrrd9/3/bang932.htm.

574. **Branckaert RDS, Guèye EF**. FAO's programme for support to family poultry production. In: **Dolberg F, Petersen PH** (eds). *Poultry as a Tool in Poverty Eradication and Promotion of Gender Equality*. Tune Landboskole, Denmark: Proceedings workshop; 1999. pp. 244–256. Retrieved May 4, 2020, from http://www.fao.org/3/AC154E/AC154E00.htm.

575. **Burns TE, Kelton D, Ribble C, Stephen C**. Preliminary investigation of bird and human movements and disease-management practices in noncommercial poultry flocks in southwestern British Columbia. *Avian Dis* 2011;55(3):350-357.

576. **Karabozhilova I, Wieland B, Alonso S, Salonen L, Hasler B**. Backyard chicken keeping in the Greater London Urban Area: welfare status, biosecurity and disease control issues. *Br Poultry Sci* 2012;53(4):421-430.

577. **Beam A, Garber L, Sakugawa J, Kopral C**. *Salmonella* awareness and related management practices in U.S. urban backyard chicken flocks. *Prev Vet Med* 2013;110(3-4):481-488.

578. **Rodriguez-Siek KE, Giddings CW, Doetkott C, Johnson TJ, Nolan LK**. Characterizing the APEC pathotype. *Vet Res* 2005;36(2):241-256.

579. **Barnes HJ NL, Vaillancourt J**. Colibacillosis. In: **Saif YM, Fadly AM, Glisson JR, Mcdougald LR, Nolan LK *et al.*** (eds). *Diseases of Poultry*, 12th. Iowa: Iowa State University Press; 2008. pp. 691–738.

580. **Hedman HD, Eisenberg JNS, Trueba G, Rivera DLV, Herrera RAZ *et al.*** Impacts of small-scale chicken farming activity on antimicrobial-resistant *Escherichia coli* carriage in backyard chickens and children in rural Ecuador. *One Health* 2019;8:100112.

581. **Sarba EJ, Kelbesa KA, Bayu MD, Gebremedhin EZ, Borena BM *et al.*** Identification and antimicrobial susceptibility profile of *Escherichia coli* isolated from backyard chicken in and around Ambo, Central Ethiopia. *BMC Vet Res* 2019;15:85.

582. **Langata LM, Maingi JM, Musonye HA, Kiiru J, Nyamache AK**. Antimicrobial resistance genes in *Salmonella* and *Escherichia coli* isolates from chicken droppings in Nairobi, Kenya. *BMC Res Notes* 2019;12:22.

583. **Borzi MM, Cardozo MV, Oliveira ES, Pollo AS, Guastalli EAL *et al.*** Characterization of avian pathogenic *Escherichia coli* isolated from free-range helmeted guineafowl. *Braz J Microbiol* 2018;49 Suppl 1:107-112.

584. **Adzitey F, Assoah-Peprah P, Teye GA**. Whole-genome sequencing of *Escherichia coli* isolated from contaminated meat samples collected from the Northern Region of Ghana reveals the presence of multiple antimicrobial resistance genes. *J Glob Antimicrob Resist* 2019;18:179-182.

585. **Vounba P, Kane Y, Ndiaye C, Arsenault J, Fairbrother JM *et al.*** Molecular characterization of *Escherichia coli* isolated from chickens with colibacillosis in Senegal. *Foodborne Pathog Dis* 2018;15(8):517-525.

586. **Falgenhauer L, Imirzalioglu C, Oppong K, Akenten CW, Hogan B *et al.*** Detection and characterization of ESBL-producing *Escherichia coli* from humans and poultry in Ghana. *Front Microbiol* 2018;9:3358.

587. **Chah KF, Ugwu IC, Okpala A, Adamu KY, Alonso CA *et al.*** Detection and molecular characterisation of extended-spectrum β-lactamase-producing enteric bacteria from pigs and chickens in Nsukka, Nigeria. *J Glob Antimicrob Resist* 2018;15:36-40.

588. **Xing-Ping L, Sun R-Y, Song J-Q, Fang L-X, Zhang R-M *et al.*** Within-host heterogeneity and flexibility of mcr-1 transmission in chicken gut. *Int J Antimicrob Agents* 2020;55:105806.

589. **Roca A, Hill PC, Townend J, Egere U, Antonio M *et al.*** Effects of community-wide vaccination with PCV-7 on pneumococcal nasopharyngeal carriage in the Gambia: A cluster-randomized trial. *PLoS Med* 2011;8(10):e1001107.

590. **Foster-Nyarko E, Alikhan NF, Ravi A, Thomson NM, Jarju S *et al.*** Genomic diversity of *Escherichia coli* from chickens and guinea fowl in the Gambia. *Microb Genom* 2020.

591. **Clermont O, Dixit OVA, Vangchhia B, Condamine B, Dion S *et al.*** Characterization and rapid identification of phylogroup G in *Escherichia coli*, a lineage with high virulence and antibiotic resistance potential. *Environ Microbiol* 2019;21(8):3107-3117.

592. **Mellata M**. Human and avian extraintestinal pathogenic *Escherichia coli*: Infections, zoonotic risks, and antibiotic resistance trends. *Foodborne Pathog Dis* 2013;10(11):916-932.

593. **Coura FM, Diniz Sde A, Silva MX, Mussi JM, Barbosa SM *et al.*** Phylogenetic group determination of *Escherichia coli* isolated from animals samples. *ScientificWorldJournal* 2015;2015:258424.

594. **Asadi A, Zahraei Salehi T, Jamshidian M, Ghanbarpour R**. ECOR phylotyping and determination of virulence genes in *Escherichia coli* isolates from pathological conditions of broiler chickens in poultry slaughter-houses of southeast of Iran. *Vet Res Forum* 2018;9(3):211-216.

595. **Messaili C, Messai Y, Bakour R**. Virulence gene profiles, antimicrobial resistance and phylogenetic groups of fecal *Escherichia coli* strains isolated from broiler chickens in Algeria. *Vet Ital* 2019;55(1):35-46.

596. **Patyk KA, Helm J, Martin MK, Forde-Folle KN, Olea-Popelka FJ *et al.*** An epidemiologic simulation model of the spread and control of highly pathogenic avian influenza (H5N1) among commercial and backyard poultry flocks in South Carolina, United States. *Prev Vet Med* 2013;110(3):510-524.

597. **Dione MM, Ikumapayi UN, Saha D, Mohammed NI, Geerts S *et al.*** Clonal differences between Non-Typhoidal Salmonella (NTS) recovered from children and animals living in close contact in the Gambia. *PLoS Negl Trop Dis* 2011;5(5):e1148.

598. **Castellanos LR, Donado-Godoy P, León M, Clavijo V, Arevalo A *et al.*** High heterogeneity of *Escherichia coli* sequence types harbouring ESBL/AmpC genes on IncI1 plasmids in the Colombian poultry chain. *PloS One* 2017;12(1):e0170777-e0170777.

599. **Mathers AJ, Peirano G, Pitout JD**. The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant *Enterobacteriaceae*. *Clin Microbiol Rev* 2015;28(3):565-591.

600. **Reich F, Atanassova V, Klein G**. Extended-spectrum beta-lactamase- and AmpC-producing enterobacteria in healthy broiler chickens, Germany. *Emerg Infect Dis* 2013;19(8):1253-1259.

601. **Dominguez JE, Faccone D, Tijet N, Gomez S, Corso A *et al.*** Characterization of *Escherichia coli* carrying mcr-1-plasmids recovered from food animals from Argentina. *Front Cell Infect Microbiol* 2019;9:41.

602. **van Hoek A, Veenman C, Florijn A, Huijbers PMC, Graat EAM *et al.*** Longitudinal study of ESBL *Escherichia coli* carriage on an organic broiler farm. *J Antimicrob Chemother* 2018;73(12):3298-3304.

603. **Samanta I, Joardar SN, Das PK, Sar TK**. Comparative possession of Shiga toxin, intimin, enterohaemolysin and major extended spectrum beta lactamase (ESBL) genes in *Escherichia coli* isolated from backyard and farmed poultry. *Iran J Vet Res* 2015;16(1):90-93.

604. **Pohjola L, Nykäsenoja S, Kivistö R, Soveri T, Huovilainen A *et al.*** Zoonotic public health hazards in backyard chickens. *Zoonoses Public Health* 2016;63(5):420-430.

605. **Fairchild AS, Smith JL, Idris U, Lu J, Sanchez S *et al.*** Effects of orally administered tetracycline on the intestinal community structure of chickens and on tet determinant carriage by commensal bacteria and *Campylobacter jejuni*. *Appl Environ Microbiol* 2005;71(10):5865-5872.

606. **Alonso CA, Zarazaga M, Ben Sallem R, Jouini A, Ben Slama K *et al.*** Antibiotic resistance in *Escherichia coli* in husbandry animals: the African perspective. *Lett Appl Microbiol* 2017;64(5):318-334.

607. **Maron DF, Smith TJS, Nachman KE**. Restrictions on antimicrobial use in food animal production: an international regulatory and economic survey. *Global Health* 2013;9:48-48.

608. **Schouler C, Schaeffer B, Bree A, Mora A, Dahbi G *et al.*** Diagnostic strategy for identifying avian pathogenic *Escherichia coli* based on four patterns of virulence genes. *J Clin Microbiol* 2012;50(5):1673-1678.

609. **Laxminarayan R, Duse A, Wattal C, Zaidi AK, Wertheim HF *et al.*** Antibiotic resistance-the need for global solutions. *Lancet Infect Dis* 2013;13(12):1057-1098.

610. **Foster-Nyarko E, Alikhan NF, Ravi A, Thilliez G, Thomson NM *et al.*** Genomic diversity of *Escherichia coli* from non-human primates in the Gambia. *Microb Genom* 2020;6(9).

611. **Doyle RM, O'Sullivan DM, Aller SD, Bruchmann S, Clark T *et al.*** Discordant bioinformatic predictions of antimicrobial resistance from whole-genome sequencing data of bacterial isolates: An inter-laboratory study. *Microb Genom* 2020:6(2).

612. **Camins BC, Marschall J, DeVader SR, Maker DE, Hoffman MW *et al.*** The clinical impact of fluoroquinolone resistance in patients with *E. coli* bacteremia. *J Hosp Med* 2011;6(6):344-349.

613. **Russo TA, Johnson JR**. Medical and economic impact of extraintestinal infections due to *Escherichia coli*: focus on an increasingly important endemic problem. *Microbes Infect* 2003;5(5):449-456.

614. **Rodríguez-Baño J, Picón E, Gijón P, Hernández JR, Cisneros JM *et al.*** Risk factors and prognosis of nosocomial bloodstream infections caused by extended-spectrum-beta-lactamase-producing *Escherichia coli. J Clin Microbiol* 2010;48(5):1726-1731.

615. **Hobman JL, Penn CW, Pallen MJ**. Laboratory strains of *Escherichia coli*: model citizens or deceitful delinquents growing old disgracefully? *Mol Microbiol* 2007;64(4):881-885.

616. **Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S *et al.*** Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* 2009;5(1):e1000344.

617. **Oshima K, Toh H, Ogura Y, Sasamoto H, Morita H *et al.*** Complete genome sequence and comparative analysis of the wild-type commensal *Escherichia coli* strain SE11 isolated from a healthy adult. *DNA Res* 2008;15(6):375-386.

618. **Ferjani S, Saidani M, Hamzaoui Z, Alonso CA, Torres C *et al.*** Community fecal carriage of broad-spectrum cephalosporin-resistant *Escherichia coli* in Tunisian children. *Diagn Microbiol Infect Dis* 2017;87(2):188-192.

619. **Moremi N, Claus H, Vogel U, Mshana SE**. Faecal carriage of CTX-M extended-spectrum beta-lactamase-producing *Enterobacteriaceae* among street children dwelling in Mwanza city, Tanzania. *PLoS One* 2017;12(9):e0184592.

620. **Ahmed SF, Ali MM, Mohamed ZK, Moussa TA, Klena JD**. Fecal carriage of extended-spectrum β-lactamases and AmpC-producing *Escherichia coli* in a Libyan community. *Ann Clin Microbiol Antimicrob* 2014;13:22.

621. **Stoppe NC, Silva JS, Carlos C, Sato MIZ, Saraiva AM *et al.*** Worldwide phylogenetic group patterns of *Escherichia coli* from commensal human and wastewater treatment plant isolates. *Front Microbiol* 2017;8:2512.

622. **McNally A, Alhashash F, Collins M, Alqasim A, Paszckiewicz K *et al.*** Genomic analysis of extra-intestinal pathogenic *Escherichia coli* urosepsis. *Clin Microbiol Infect* 2013;19(8):E328-334.

623. **Nielsen KL, Stegger M, Godfrey PA, Feldgarden M, Andersen PS *et al.*** Adaptation of *Escherichia coli* traversing from the faecal environment to the urinary tract. *Int J Med Microbiol* 2016;306(8):595-603.

624. **Hartl DL, Dykhuizen DE**. The population genetics of *Escherichia coli*. *Annu Rev Genet* 1984;18:31-68.

625. **Kotloff KL, Blackwelder WC, Nasrin D, Nataro JP, Farag TH *et al.*** The Global Enteric Multicenter Study (GEMS) of diarrheal disease in infants and young children in developing countries: epidemiologic and clinical methods of the case/control study. *Clin Infect Dis* 2012;55 Suppl 4:S232-245.

626. **Foster-Nyarko E, Alikhan NF, Ikumapayi UN, Sarwar G, Okoi C *et al.*** Genomic diversity of *Escherichia coli* from healthy children in rural Gambia. *PeerJ* 2021;9:e10572.

627. **Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L *et al.*** Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform* 2013;14(2):193-202.

628. **Pokharel P, Habouria H, Bessaiah H, Dozois CM**. Serine Protease Autotransporters of the *Enterobacteriaceae* (SPATEs): Out and about and chopping it up. *Microorganisms* 2019;7(12).

629. **Escobar-Páramo P, Grenet K, Le Menac'h A, Rode L, Salgado E *et al.*** Large-scale population structure of human commensal *Escherichia coli* isolates. *Appl Environ Microbiol* 2004;70(9):5698-5700.

630. **Duriez P, Clermont O, Bonacorsi S, Bingen E, Chaventré A *et al.*** Commensal *Escherichia coli* isolates are phylogenetically distributed among geographically distinct human populations. *Microbiology* 2001;147(Pt 6):1671-1676.

631. **Gordon DM, O'Brien CL, Pavli P**. *Escherichia coli* diversity in the lower intestinal tract of humans. *Environ Microbiol Rep* 2015;7(4):642-648.

632. **Skurnik D, Bonnet D, Bernède-Bauduin C, Michel R, Guette C *et al.*** Characteristics of human intestinal *Escherichia coli* with changing environments. *Environ Microbiol* 2008;10(8):2132-2137.

633. **Massot M, Daubié AS, Clermont O, Jauréguy F, Couffignal C *et al.*** Phylogenetic, virulence and antibiotic resistance characteristics of commensal strain populations of *Escherichia coli* from community subjects in the Paris area in 2010 and evolution over 30 years. *Microbiology* 2016;162(4):642-650.

634. **Wold AE, Caugant DA, Lidin-Janson G, de Man P, Svanborg C**. Resident colonic *Escherichia col*i strains frequently display uropathogenic characteristics. *J Infect Dis* 1992;165(1):46-52.

635. **González-González A, Sánchez-Reyes LL, Delgado Sapien G, Eguiarte LE, Souza V**. Hierarchical clustering of genetic diversity associated to different levels of mutation and recombination in *Escherichia coli*: a study based on Mexican isolates. *Infect Genet Evol* 2013;13:187-197.

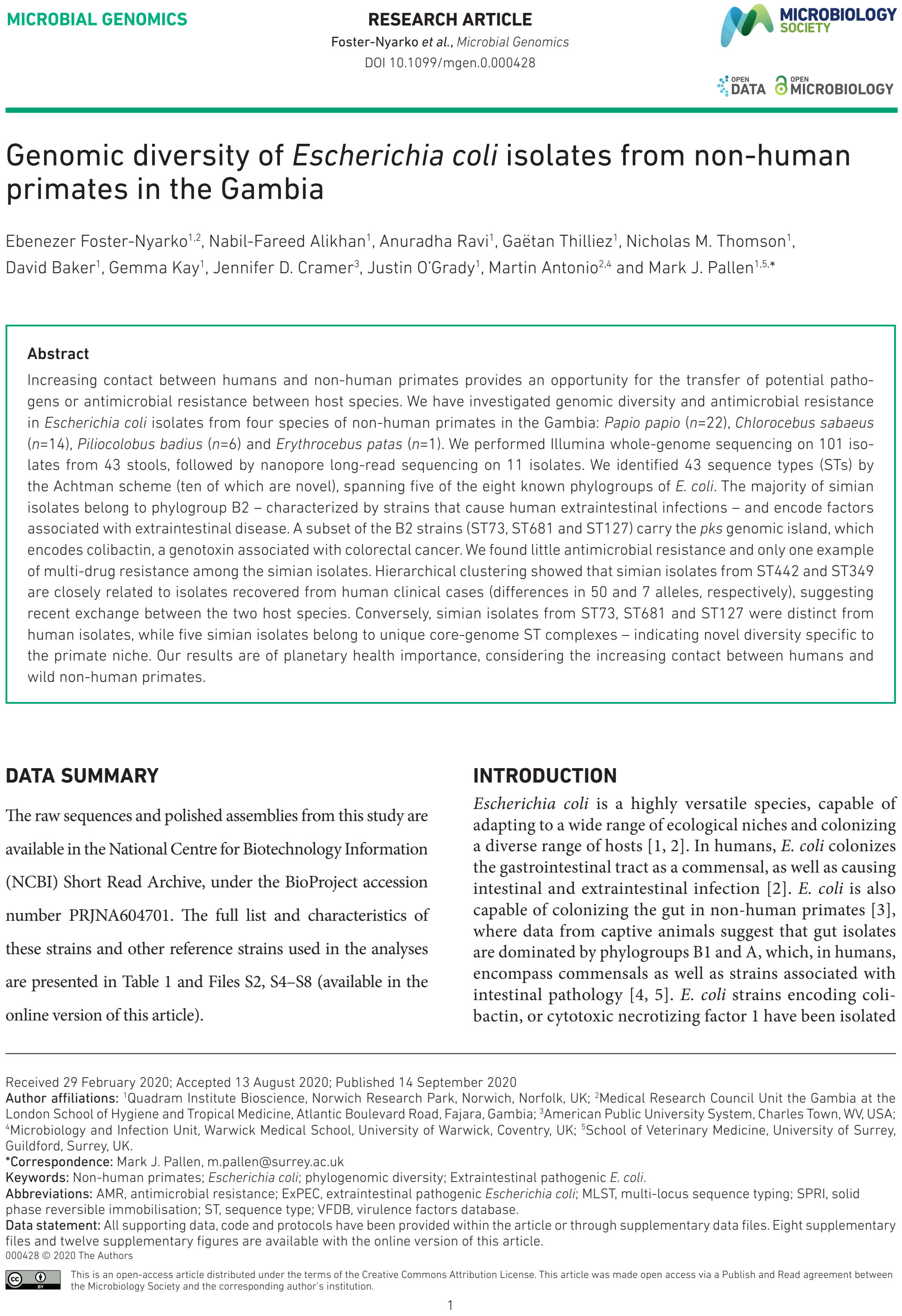
636. **Rodríguez JM, Murphy K, Stanton C, Ross RP, Kober OI *et al.*** The composition of the gut microbiota throughout life, with an emphasis on early life. *Microb Ecol Health Dis* 2015;26:26050.

637. **Potron A, Poirel L, Rondinaud E, Nordmann P**. Intercontinental spread of OXA-48 beta-lactamase-producing *Enterobacteriaceae* over a 11-year period, 2001 to 2011. *Euro Surveill* 2013;18(31).

638. **Durso LM, Smith D, Hutkins RW**. Measurements of fitness and competition in commensal *Escherichia coli* and *E. coli* O157:H7 strains. *Appl Environ Microbiol* 2004;70(11):6466-6472.

# APPENDICES

## APPENDIX 1: Genomic diversity of Escherichia coli from non-human primates in the Gambia



MICROBIAL GENOMICS

RESEARCH ARTICLE
Foster-Nyarko *et al.*, *Microbial Genomics*
DOI 10.1099/mgen.0.000428

MICROBIOLOGY SOCIETY
OPEN DATA OPEN MICROBIOLOGY

# Genomic diversity of *Escherichia coli* isolates from non-human primates in the Gambia

Ebenezer Foster-Nyarko[1,2], Nabil-Fareed Alikhan[1], Anuradha Ravi[1], Gaëtan Thilliez[1], Nicholas M. Thomson[1], David Baker[1], Gemma Kay[1], Jennifer D. Cramer[3], Justin O'Grady[1], Martin Antonio[2,4] and Mark J. Pallen[1,5,*]

**Abstract**

Increasing contact between humans and non-human primates provides an opportunity for the transfer of potential pathogens or antimicrobial resistance between host species. We have investigated genomic diversity and antimicrobial resistance in *Escherichia coli* isolates from four species of non-human primates in the Gambia: *Papio papio* (*n*=22), *Chlorocebus sabaeus* (*n*=14), *Piliocolobus badius* (*n*=6) and *Erythrocebus patas* (*n*=1). We performed Illumina whole-genome sequencing on 101 isolates from 43 stools, followed by nanopore long-read sequencing on 11 isolates. We identified 43 sequence types (STs) by the Achtman scheme (ten of which are novel), spanning five of the eight known phylogroups of *E. coli*. The majority of simian isolates belong to phylogroup B2 – characterized by strains that cause human extraintestinal infections – and encode factors associated with extraintestinal disease. A subset of the B2 strains (ST73, ST681 and ST127) carry the *pks* genomic island, which encodes colibactin, a genotoxin associated with colorectal cancer. We found little antimicrobial resistance and only one example of multi-drug resistance among the simian isolates. Hierarchical clustering showed that simian isolates from ST442 and ST349 are closely related to isolates recovered from human clinical cases (differences in 50 and 7 alleles, respectively), suggesting recent exchange between the two host species. Conversely, simian isolates from ST73, ST681 and ST127 were distinct from human isolates, while five simian isolates belong to unique core-genome ST complexes – indicating novel diversity specific to the primate niche. Our results are of planetary health importance, considering the increasing contact between humans and wild non-human primates.

## DATA SUMMARY

The raw sequences and polished assemblies from this study are available in the National Centre for Biotechnology Information (NCBI) Short Read Archive, under the BioProject accession number PRJNA604701. The full list and characteristics of these strains and other reference strains used in the analyses are presented in Table 1 and Files S2, S4–S8 (available in the online version of this article).

## INTRODUCTION

*Escherichia coli* is a highly versatile species, capable of adapting to a wide range of ecological niches and colonizing a diverse range of hosts [1, 2]. In humans, *E. coli* colonizes the gastrointestinal tract as a commensal, as well as causing intestinal and extraintestinal infection [2]. *E. coli* is also capable of colonizing the gut in non-human primates [3], where data from captive animals suggest that gut isolates are dominated by phylogroups B1 and A, which, in humans, encompass commensals as well as strains associated with intestinal pathology [4, 5]. *E. coli* strains encoding colibactin, or cytotoxic necrotizing factor 1 have been isolated

Received 29 February 2020; Accepted 13 August 2020; Published 14 September 2020

**Author affiliations:** [1]Quadram Institute Bioscience, Norwich Research Park, Norwich, Norfolk, UK; [2]Medical Research Council Unit the Gambia at the London School of Hygiene and Tropical Medicine, Atlantic Boulevard Road, Fajara, Gambia; [3]American Public University System, Charles Town, WV, USA; [4]Microbiology and Infection Unit, Warwick Medical School, University of Warwick, Coventry, UK; [5]School of Veterinary Medicine, University of Surrey, Guildford, Surrey, UK.
***Correspondence:** Mark J. Pallen, m.pallen@surrey.ac.uk

**Keywords:** Non-human primates; *Escherichia coli*; phylogenomic diversity; Extraintestinal pathogenic *E. coli*.

**Abbreviations:** AMR, antimicrobial resistance; ExPEC, extraintestinal pathogenic *Escherichia coli*; MLST, multi-locus sequence typing; SPRI, solid phase reversible immobilisation; ST, sequence type; VFDB, virulence factors database.

**Data statement:** All supporting data, code and protocols have been provided within the article or through supplementary data files. Eight supplementary files and twelve supplementary figures are available with the online version of this article.

from healthy laboratory rhesus macaques [6], while enteropathogenic *E. coli* strains can – in the laboratory – cause colitis in marmosets [7], rhesus macaques infected with simian immunodeficiency virus [8] and cotton-top tamarins [9].

There are two potential explanations for the occurrence of *E. coli* in humans and non-human primates. Some bacterial lineages may have been passed on through vertical transmission within the same host species for long periods, perhaps even arising from ancestral bacteria that colonized the guts of the most recent common ancestors of humans and non-human primate species [10, 11]. In such a scenario, isolates from non-human primates would be expected to be novel and distinct from the diversity seen in humans [11]. However, there is also clearly potential for horizontal transfer of strains from one host species to another [12].

The exchange of bacteria between humans and human-habituated animals, particularly non-human primates, is of interest in light of the fragmentation of natural habitats globally [13–15]. We have seen that wild non-human primates in the Gambia are frequently exposed to humans through tourism, deforestation and urbanization [16]. In Uganda, PCR-based studies have suggested transmission of *E. coli* between humans, non-human primates and livestock [17, 18]. These studies are complicated by the low resolution of PCR-based methods; nonetheless, their findings highlight the possibility that wild non-human primates may constitute a reservoir for the zoonotic spread of *E. coli* strains associated with virulence and antimicrobial resistance to humans. Alternatively, humans might provide a reservoir of strains with the potential for anthroponotic spread to animals – or transmission might occur in both directions [19].

We do not know how many different lineages can co-exist within the same non-human primate host. Such information may help us contextualize the potential risks associated with transmission of bacterial strains between humans and non-human primates. In humans, up to 11 serotypes could be sampled from picking colonies from individual stool samples [20–22].

To address these issues, we have exploited whole-genome sequencing to explore the population structure and phylogenomic diversity of *E. coli* in wild non-human primates from rural and urban Gambia.

## METHODS

### Study population and sample collection

In June 2017, wild non-human primates were sampled from six sampling sites in the Gambia: Abuko Nature Reserve (riparian forest), Bijilo Forest Park (coastal fenced woodland), Kartong village (mangrove swamp), Kiang West National park (dry-broad-leaf forest), Makasutu Cultural Forest (ecotourism woodland) and River Gambia National park (riparian forest) (Fig. 1). The sampling was opportunistic and throughout the range of the primates in the country (all four of the diurnal non-human primate species indigenous to the Gambia), where primates overlap with human communities to varying degrees.

**Impact Statement**

Little is known about the population structure, virulence potential and the burden of antimicrobial resistance among *Escherichia coli* from wild non-human primates, despite increased exposure to humans through the fragmentation of natural habitats. Previous studies, primarily involving captive animals, have highlighted the potential for bacterial exchange between non-human primates and humans living nearby, including strains associated with intestinal pathology. Using multiple-colony sampling and whole-genome sequencing, we investigated the strain distribution and population structure of *E. coli* from wild non-human primates from the Gambia. Our results indicate that these monkeys harbour strains that can cause extraintestinal infections in humans. We document the transmission of virulent *E. coli* strains between monkeys of the same species sharing a common habitat and evidence of recent interaction between strains from humans and wild non-human primates. Also, we present complete genome assemblies for five novel sequence types of *E. coli*.

Monkeys in Abuko and Bijilo are frequently hand-fed by visiting tourists, despite guidelines prohibiting this practice [16].

Troops of monkeys were observed and followed. We collected single freshly passed formed stool specimens from 43 visibly healthy individuals (38 adults, 5 juveniles; 24 females, 11 males, 8 of undetermined sex), drawn from four species: *Erythrocebus patas* (patas monkey), *Papio papio* (Guinea baboon), *Chlorocebus sabaeus* (green monkey) and *Piliocolobus badius* (Western colobus monkey). Stool samples were immediately placed into sterile falcon tubes, taking care to collect portions of stool material that had not touched the ground, then placed on dry ice and stored at 80 °C within 6 h (Fig. 2).

### Microbiological processing

For the growth and isolation of *E. coli*, 0.1–0.2 g aliquots were taken from each stool sample into 1.5 ml microcentrifuge tubes under aseptic conditions. To each tube, 1 ml of physiological saline (0.85%) was added and the saline-stool samples were vortexed for 2 min at 4, 200 r.p.m. The homogenized samples were taken through four tenfold serial dilutions and a 100 µl aliquot from each dilution was spread on a plate of tryptone-bile-X-glucoronide agar using the cross-hatching method. Plates were incubated at 37 °C for 18–24 h in air. Colony counts were performed for each serial dilution, counting translucent colonies with blue-green pigmentation and entire margins as *E. coli*. Up to five colonies from each sample were sub-cultured on MacConkey agar at 37 °C for 18–24 h and then stored in 20% glycerol broth at −80 °C. Previous studies have shown that sampling five colonies provides a 99.3% chance of recovering at least one of the dominant genotypes present in a single stool specimen [23, 24].

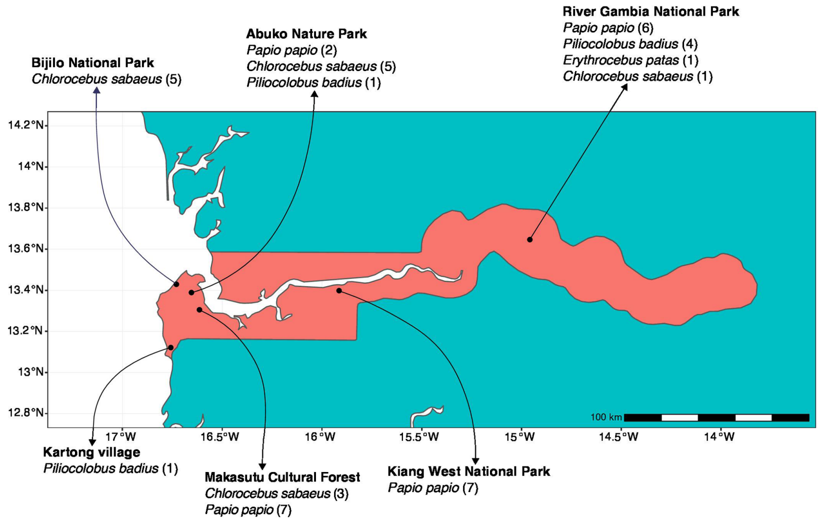


**Fig. 1.** Study sites and distribution of study subjects.

## Genomic DNA extraction

A single colony from each subculture was picked into 1 ml Luria–Bertani broth and incubated overnight at 37 °C. Broth cultures were spun at 3, 500 r.p.m. for 2 min and lysed using lysozyme, proteinase K, 10% SDS and RNase A in Tris EDTA buffer (pH 8.0). Suspensions were placed on a thermomixer with vigorous shaking at 1600 r.p.m., first at 37 °C for 25 min and subsequently at 65 °C for 15 min. DNA was extracted using solid-phase reversible immobilisation magnetic beads (Becter Coulter, Brea, CA, USA), precipitated with ethanol, eluted in

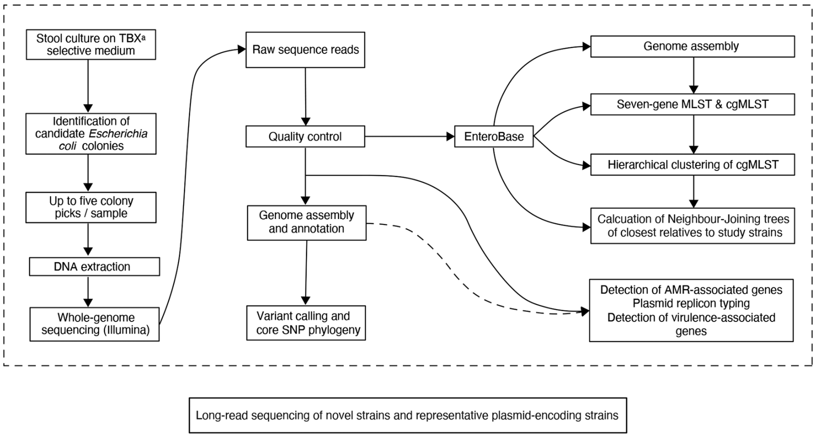

Stool culture on TBX^a selective medium
Identification of candidate Escherichia coli colonies
Up to five colony picks / sample
DNA extraction
Whole-genome sequencing (Illumina)
Raw sequence reads
Quality control
Genome assembly and annotation
Variant calling and core SNP phylogeny
EnteroBase
Genome assembly
Seven-gene MLST & cgMLST
Hierarchical clustering of cgMLST
Calcuation of Neighbour-Joining trees of closest relatives to study strains
Detection of AMR-associated genes
Plasmid replicon typing
Detection of virulence-associated genes
Long-read sequencing of novel strains and representative plasmid-encoding strains

**Fig. 2.** Study sample-processing flow diagram.

Tris-Cl and evaluated for protein and RNA contamination using $A_{260}/A_{280}$ and $A_{260}/A_{230}$ ratios on the NanoDrop 2000 Spectrophotometer (Fisher Scientific, Loughborough, UK). DNA concentrations were measured using the Qubit HS DNA assay (Invitrogen, MA, USA). DNA was stored at −20 °C.

### Illumina sequencing

Whole-genome sequencing was carried out on the Illumina NextSeq 500 platform (Illumina, San Diego, CA, USA). We used a modified Nextera XT DNA protocol for the library preparation (File S1). The genomic DNA was normalized to 0.5 ng $\mu l^{-1}$ with 10 mM Tris-HCl prior to the library preparation. The pooled library was run at a final concentration of 1.8 pM on a mid-output flow cell (NSQ 500 Mid Output KT v2 300 cycles; Illumina Catalogue No. FC-404–2003) following the Illumina recommended denaturation and loading parameters, which included a 1% PhiX spike (PhiX Control v3; Illumina Catalogue FC-110–3001). The data was uploaded to BaseSpace (http://www.basespace.illumina.com) and then converted to FASTQ files.

### Oxford nanopore sequencing

We used the rapid barcoding kit (Oxford Nanopore Catalogue No. SQK-RBK004) to prepare libraries according to the manufacturer's instructions. We used 400 ng DNA for library preparation and loaded 75 µl of the prepared library on an R9.4 MinION flow cell. The size of the DNA fragments was assessed using the Agilent 2200 TapeStation (Agilent Catalogue No. 5067–5579) before sequencing. The concentration of the final library pool was measured using the Qubit high-sensitivity DNA assay (Invitrogen, MA, USA).

### Genome assembly and phylogenetic analysis

Sequences were analysed on the Cloud Infrastructure for Microbial Bioinformatics [25]. Paired-end short-read sequences were concatenated, then quality-checked using FastQC v0.11.7 [26]. Reads were assembled using Shovill (https://github.com/tseemann/shovill) and assemblies assessed using QUAST v 5.0.0, de6973bb [27]. Draft bacterial genomes were annotated using Prokka v 1.13 [28]. Multilocus sequence types were called from assemblies according to the Achtman scheme using the mlst software (https://github.com/tseemann/mlst) to scan alleles in PubMLST (https://pubmlst.org/) [29]. Novel STs were assigned by EnteroBase – an online integrated software environment, which routinely retrieves short-read *E. coli* sequences from the public domain, or using user-uploaded short reads, *de novo* assembles these and assigns seven-allele MLST (ST) and phylogroups from genome assemblies using standardized pipelines [30]. EnteroBase assigns new allele IDs or STs in the event of a locus being discovered with a novel allele. Snippy v4.3.2 (https://github.com/tseemann/snippy) was used for variant calling and core genome alignment, including reference genome sequences representing the major phylogroups of *E. coli* and *Escherichia fergusonii* as an outgroup (File S2b). We used Gubbins (Genealogies Unbiased By recomBinations In Nucleotide Sequences) to detect and remove recombinant regions of the core genome alignment [31]. RAxML v 8.2.4 [32] was used for maximum-likelihood phylogenetic inference from this masked alignment based on a general time-reversible nucleotide substitution model with 1, 000 bootstrap replicates. The phylogenetic tree was visualized using Mega v. 7.2 [33] and annotated using Adobe Illustrator v 23.0.3 (Adobe, San Jose, CA, USA). Pair-wise SNP distances between genomes were computed from the core-gene alignment using snp-dists v0.6 (https://github.com/tseemann/snp-dists).

### Population structure and analysis of gene content

Merged short reads were uploaded to EnteroBase [30], where we used the Hierarchical Clustering (HierCC) algorithm to assign our genomes from non-human primates to HC1100 clusters, which in *E. coli* correspond roughly to the clonal complexes seen in seven-allele MLST. Core genome MLST (cgMLST) profiles based on the typing of 2, 512 core loci for *E. coli* facilitates single-linkage hierarchical clustering according to fixed core genome MLST (cgMLST) allelic distances, based on cgMLST allelic differences. Thus, cgST HierCC provides a robust approach to analyse population structures at multiple levels of resolution. The identification of closely related genomes using HierCC has been shown to be 89% consistent between cgMLST and SNPs [34]. Neighbour-joining trees were reconstructed with NINJA – a hierarchical clustering algorithm for inferring phylogenies that is capable of scaling to inputs larger than 100, 000 sequences [35].

ARIBA v2.12.1 [36] was used to search short reads against the Virulence Factors Database [37] (VFDB-core) (virulence-associated genes), ResFinder (AMR) [38] and PlasmidFinder (plasmid-associated genes) [39] databases (both ResFinder and PlasmidFinder databases downloaded 29 October 2018). Percentage identity of ≥90% and coverage of ≥70% of the respective gene length were taken as a positive result. Analyses were performed on assemblies using ABRicate v 0.8.7 (https://github.com/tseemann/abricate). A heat map of detected virulence- and AMR-associated genes was plotted on the phylogenetic tree using ggtree and phangorn in RStudio v 3.5.1.

We searched EnteroBase for all *E. coli* strains isolated from humans in the Gambia (*n*=128), downloaded the genomes and screened them for resistance genes using ABRicate v 0.9.8. Assembled genomes for isolates that clustered with our colibactin-encoding ST73, ST127 and ST681 isolates were downloaded and screened for the colibactin operon using ABRicate's VFDB database (accessed 28 July 2019). Assemblies reported to contain colibactin genes were aligned against the colibactin-encoding *E. coli* IHE3034 reference genome (NCBI accession: GCA_000025745.1) using minimap2 2.13-r850. BAM files were visualized in Artemis Release 17.0.1 [40] to confirm the presence of the *pks* genomic island, which encodes the colibactin operon (*clbA-S*).

### Hybrid assembly and analysis of plasmids and phages

Base-called FASTQ files were concatenated into a single file and demultiplexed into individual FASTQ files based on

**Table 1.** Study isolates

| Name | Source | Individual sampling no. | Colony-pick | Sampling site | ST |
|---|---|---|---|---|---|
| PapRG-03–1 | *Papio papio* | 3 | 1 | River Gambia national park | 336 |
| PapRG-03–2 | *Papio papio* | 3 | 2 | River Gambia national park | 336 |
| PapRG-03–3 | *Papio papio* | 3 | 3 | River Gambia national park | 336 |
| PapRG-03–4 | *Papio papio* | 3 | 4 | River Gambia national park | 336 |
| PapRG-03–5 | *Papio papio* | 3 | 5 | River Gambia national park | 336 |
| PapRG-04–1 | *Papio papio* | 4 | 1 | River Gambia national park | 1665 |
| PapRG-04–2 | *Papio papio* | 4 | 2 | River Gambia national park | 1204 |
| PapRG-04–4 | *Papio papio* | 4 | 3 | Makasutu cultural forest | 8826 |
| PapRG-04–5 | *Papio papio* | 4 | 4 | Makasutu cultural forest | 1204 |
| PapRG-05–2 | *Papio papio* | 5 | 1 | Makasutu cultural forest | 1431 |
| PapRG-05–3 | *Papio papio* | 5 | 2 | Makasutu cultural forest | 99 |
| PapRG-05–4 | *Papio papio* | 5 | 3 | Makasutu cultural forest | 6316 |
| PapRG-05–5 | *Papio papio* | 5 | 4 | Makasutu cultural forest | 1431 |
| PapRG-06–1 | *Papio papio* | 6 | 1 | Makasutu cultural forest | 4080 |
| PapRG-06–2 | *Papio papio* | 6 | 2 | Makasutu cultural forest | 2521 |
| PapRG-06–3 | *Papio papio* | 6 | 3 | Makasutu cultural forest | 8827 |
| PapRG-06–4 | *Papio papio* | 6 | 4 | Makasutu cultural forest | 1204 |
| PapRG-06–5 | *Papio papio* | 6 | 5 | River Gambia national park | 8525 |
| ProbRG-07–1 | *Piliocolobus badius* | 7 | 1 | River Gambia national park | 73 |
| ProbRG-07–2 | *Piliocolobus badius* | 7 | 2 | River Gambia national park | 73 |
| ProbRG-07–3 | *Piliocolobus badius* | 7 | 3 | River Gambia national park | 73 |
| ProbRG-07–4 | *Piliocolobus badius* | 7 | 4 | River Gambia national park | 73 |
| ProbRG-07–5 | *Piliocolobus badius* | 7 | 5 | River Gambia national park | 73 |
| ChlosRG-12–1 | *Chlorocebus sabaeus* | 12 | 1 | River Gambia national park | 8824 |
| ChlosRG-12–2 | *Chlorocebus sabaeus* | 12 | 2 | River Gambia national park | 196 |
| ChlosRG-12–3 | *Chlorocebus sabaeus* | 12 | 3 | River Gambia national park | 196 |
| ChlosRG-12–5 | *Chlorocebus sabaeus* | 12 | 4 | River Gambia national park | 40 |
| ChlosAN-13–1 | *Chlorocebus sabaeus* | 13 | 1 | Abuko Nature Reserve | 8526 |

Continued

**Table 1.** Continued

| Name | Source | Individual sampling no. | Colony-pick | Sampling site | ST |
|---|---|---|---|---|---|
| ChlosAN-13–2 | *Chlorocebus sabaeus* | 13 | 2 | Abuko Nature Reserve | 8550 |
| ChlosAN-13–4 | *Chlorocebus sabaeus* | 13 | 3 | Abuko Nature Reserve | 1973 |
| ChlosAN-13–5 | *Chlorocebus sabaeus* | 13 | 4 | Abuko Nature Reserve | 1973 |
| PapAN-14–1 | *Papio papio* | 14 | 1 | Abuko Nature Reserve | 2076 |
| PapAN-14–2 | *Papio papio* | 14 | 2 | Abuko Nature Reserve | 939 |
| PapAN-14–3 | *Papio papio* | 14 | 3 | Abuko Nature Reserve | 226 |
| PapAN-14–4 | *Papio papio* | 14 | 4 | Abuko Nature Reserve | 226 |
| PapAN-14–5 | *Papio papio* | 14 | 5 | Abuko Nature Reserve | 226 |
| PapAN-15–1 | *Papio papio* | 15 | 1 | Abuko Nature Reserve | 226 |
| PapAN-15–2 | *Papio papio* | 15 | 2 | Abuko Nature Reserve | 5073 |
| PapAN-15–3 | *Papio papio* | 15 | 3 | Abuko Nature Reserve | 226 |
| PapAN-15–4 | *Papio papio* | 15 | 4 | Abuko Nature Reserve | 126 |
| PapAN-15–5 | *Papio papio* | 15 | 5 | Abuko Nature Reserve | 8823 |
| ChlosAN-17–1 | *Chlorocebus sabaeus* | 17 | 1 | Abuko Nature Reserve | 681 |
| ChlosAN-17–2 | *Chlorocebus sabaeus* | 17 | 2 | Abuko Nature Reserve | 362 |
| ChlosAN-17–3 | *Chlorocebus sabaeus* | 17 | 3 | Abuko Nature Reserve | 681 |
| ChlosAN-17–4 | *Chlorocebus sabaeus* | 17 | 4 | Abuko Nature Reserve | 681 |
| ChlosAN-18–1 | *Chlorocebus sabaeus* | 18 | 1 | Abuko Nature Reserve | 681 |
| ChlosAN-18–2 | *Chlorocebus sabaeus* | 18 | 2 | Abuko Nature Reserve | 681 |
| ChlosAN-18–3 | *Chlorocebus sabaeus* | 18 | 3 | Abuko Nature Reserve | 681 |
| ChlosAN-18–4 | *Chlorocebus sabaeus* | 18 | 4 | Abuko Nature Reserve | 681 |
| ChlosAN-18–5 | *Chlorocebus sabaeus* | 18 | 5 | Abuko Nature Reserve | 349 |
| ProbAN-19–2 | *Piliocolobus badius* | 19 | 1 | Abuko Nature Reserve | 8825 |
| ChlosBP-21–1 | *Chlorocebus sabaeus* | 21 | 1 | Bijilo forest park | 677 |
| ChlosBP-21–2 | *Chlorocebus sabaeus* | 21 | 2 | Bijilo forest park | 677 |
| ChlosBP-21–3 | *Chlorocebus sabaeus* | 21 | 3 | Bijilo forest park | 677 |
| ChlosBP-21–4 | *Chlorocebus sabaeus* | 21 | 4 | Bijilo forest park | 677 |
| ChlosBP-21–5 | *Chlorocebus sabaeus* | 21 | 5 | Bijilo forest park | 677 |
| ChlosBP-23–1 | *Chlorocebus sabaeus* | 23 | 2 | Bijilo forest park | 8527 |
| ChlosBP-23–2 | *Chlorocebus sabaeus* | 23 | 3 | Bijilo forest park | 8527 |
| ChlosBP-23–3 | *Chlorocebus sabaeus* | 23 | 4 | Bijilo forest park | 3306 |
| ChlosBP-24–1 | *Chlorocebus sabaeus* | 24 | 1 | Bijilo forest park | 73 |
| ChlosBP-24–2 | *Chlorocebus sabaeus* | 24 | 2 | Bijilo forest park | 73 |
| ChlosBP-24–3 | *Chlorocebus sabaeus* | 24 | 3 | Bijilo forest park | 73 |
| ChlosBP-24–4 | *Chlorocebus sabaeus* | 24 | 4 | Bijilo forest park | 73 |
| ChlosBP-24–5 | *Chlorocebus sabaeus* | 24 | 5 | Bijilo forest park | 73 |
| ChlosBP-25–1 | *Chlorocebus sabaeus* | 25 | 1 | Bijilo forest park | 73 |

Continued

**Table 1.** Continued

| Name | Source | Individual sampling no. | Colony-pick | Sampling site | ST |
|---|---|---|---|---|---|
| ChlosBP-25–2 | *Chlorocebus sabaeus* | 25 | 2 | Bijilo forest park | 73 |
| ChlosBP-25–3 | *Chlorocebus sabaeus* | 25 | 3 | Bijilo forest park | 73 |
| ChlosBP-25–4 | *Chlorocebus sabaeus* | 25 | 4 | Bijilo forest park | 73 |
| ChlosBP-25–5 | *Chlorocebus sabaeus* | 25 | 5 | Bijilo forest park | 73 |
| ChlosM-29–1 | *Chlorocebus sabaeus* | 29 | 1 | Makasutu cultural forest | 1873 |
| ChlosM-29–2 | *Chlorocebus sabaeus* | 29 | 2 | Makasutu cultural forest | 1873 |
| PapM-31–1 | *Papio papio* | 31 | 1 | Makasutu cultural forest | 2800 |
| PapM-31–2 | *Papio papio* | 31 | 2 | Makasutu cultural forest | 135 |
| PapM-31–3 | *Papio papio* | 31 | 3 | Makasutu cultural forest | 5780 |
| PapM-31–4 | *Papio papio* | 31 | 4 | Makasutu cultural forest | 1727 |
| PapM-31–5 | *Papio papio* | 31 | 5 | Makasutu cultural forest | 5780 |
| PapM-32–1 | *Papio papio* | 32 | 2 | Makasutu cultural forest | 8532 |
| PapM-32–2 | *Papio papio* | 32 | 3 | Makasutu cultural forest | 212 |
| PapM-32–3 | *Papio papio* | 32 | 4 | Makasutu cultural forest | 212 |
| PapM-32–4 | *Papio papio* | 32 | 5 | Makasutu cultural forest | 212 |
| PapM-32–5 | *Papio papio* | 32 | 6 | Makasutu cultural forest | 212 |
| PapM-33–1 | *Papio papio* | 33 | 1 | Makasutu cultural forest | 8533 |
| PapM-33–2 | *Papio papio* | 33 | 2 | Makasutu cultural forest | 8533 |
| PapM-33–3 | *Papio papio* | 33 | 3 | Makasutu cultural forest | 8533 |
| PapM-33–4 | *Papio papio* | 33 | 4 | Makasutu cultural forest | 38 |
| PapM-33–5 | *Papio papio* | 33 | 5 | Makasutu cultural forest | 8533 |
| PapM-34–1 | *Papio papio* | 34 | 1 | Makasutu cultural forest | 676 |
| PapM-34–2 | *Papio papio* | 34 | 2 | Makasutu cultural forest | 676 |
| PapM-34–3 | *Papio papio* | 34 | 3 | Makasutu cultural forest | 676 |
| PapM-34–4 | *Papio papio* | 34 | 4 | Makasutu cultural forest | 676 |
| PapM-36–1 | *Papio papio* | 36 | 1 | Makasutu cultural forest | 8535 |
| PapM-36–2 | *Papio papio* | 36 | 2 | Makasutu cultural forest | 8535 |
| PapKW-44–1 | *Papio papio* | 44 | 1 | Kiang West national park | 442 |
| PapKW-44–2 | *Papio papio* | 44 | 2 | Kiang West national park | 442 |
| PapKW-44–3 | *Papio papio* | 44 | 3 | Kiang West national park | 442 |
| PapKW-44–4 | *Papio papio* | 44 | 4 | Kiang West national park | 442 |
| ProbK-45–1 | *Piliocolobus badius* | 45 | 1 | Kartong village | 127 |
| ProbK-45–2 | *Piliocolobus badius* | 45 | 2 | Kartong village | 127 |
| ProbK-45–3 | *Piliocolobus badius* | 45 | 3 | Kartong village | 127 |
| ProbK-45–4 | *Piliocolobus badius* | 45 | 4 | Kartong village | 127 |
| ProbK-45–5 | *Piliocolobus badius* | 45 | 5 | Kartong village | 127 |

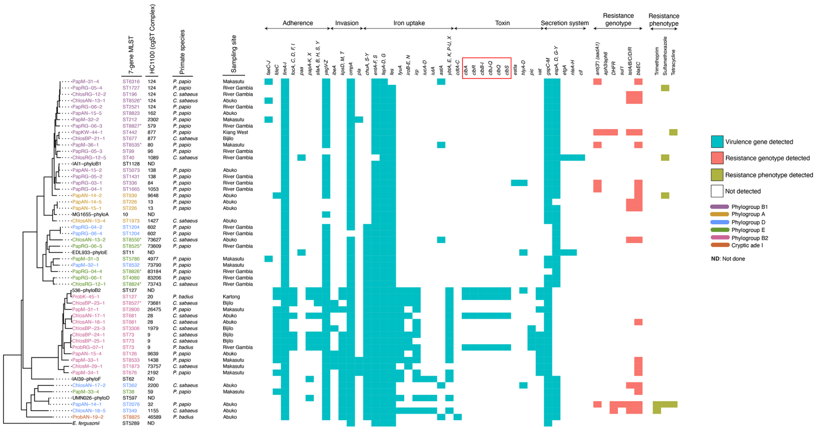


**Fig. 3.** Plot showing the maximum-likelihood phylogeny of the study isolates overlaid with the prevalence of putative virulence genes and resistance-associated genes, as well as the phenotypic antimicrobial resistance among the study isolates. The tree was reconstructed based on non-repetitive core SNPs calculated against the *E. coli* K-12 reference strain (NCBI accession: NC_000913.3), using RAxML with 1000 bootstrap replicates. *E. coli* MG1655 was used as the reference and *E. fergusonii* as the outroot species. Recombinant regions were removed using Gubbins [31]. The tip labels indicate the sample IDs, with the respective *in silico* Achtman STs and HC1100 (cgST complexes) indicated next to the tip labels. Both the sample IDs and the STs (Achtman) are colour-coded to indicate the various phylogroups as indicated. Novel STs (Achtman) are indicated by an asterisk (*). *Escherichia fergusonii* and the *E. coli* reference genomes representing the major *E. coli* phylogroups are in black. Primate species are indicated by strain names as follows: *Chlorocebus sabaeus*, 'Chlos'; *Papio papio*, 'Pap'; *Piliocolobus badius*, 'Prob'. These strain designations are also used in annotating the plot next to the tree. The sampling sites are indicated as follows: BP, Bijilo forest park; KW, Kiang-West National park; RG, River Gambia National Park; M, Makasutu Cultural forest; AN, Abuko Nature reserve; K, Kartong village. These site designations are also used in annotating the plot next to the tree. Cocolonising seven-allele (Achtman) STs in single individuals are shown by the prefix of the strain names depicting the colony as 1, 2 up to 5. We do not show multiple colonies of the same Achtman ST recovered from a single individual. In such cases, only one representative is shown. Virulence genes are grouped according to their function, with genes encoding the colibactin genotoxin highlighted with a red box. The full names of virulence factors are provided in File S7. The resistance-associated genes detected among the study isolates and the class of antibiotic to which they encode resistance are as follows: *ant*(3′′) (*aadA1*) and *aph3/aph6*, aminoglycosides; DHFR, trimethoprim; *sul1*, sulphonamides; *tetA/B/C/D/R*, tetracyclines; *bla*EC, beta-lactamase (penicillinase-type).

barcodes, using the qcat python command-line tool v 1.1.0 (https://github.com/nanoporetech/qcat). Hybrid assemblies of the Illumina and nanopore reads were created with Unicycler [41]. The quality and completion of the hybrid assemblies were assessed with QUAST v 5.0.0, de6973bb and CheckM [27, 42]. Hybrid assemblies were interrogated using ABRicate PlasmidFinder and annotated using Prokka [28]. Plasmid sequences were visualized in Artemis using coordinates from ABRicate. Prophage identification was carried out using the phage search tool, PHASTER [43].

### Antimicrobial susceptibility

We determined the MICs of amikacin, trimethoprim, sulfamethoxazole, ciprofloxacin, cefotaxime and tetracycline for the isolates from non-human primates using agar dilution [44]. Twofold serial dilutions of each antibiotic were performed in molten Mueller–Hinton agar (Oxoid, Basingstoke, UK), from 32 mg l$^{-1}$ to 0.03 mg l$^{-1}$ (512 mg l$^{-1}$ to 0.03 mg l$^{-1}$ for sulfamethoxazole), using *E. coli* NCTC 10418 as control. MICs were performed in duplicate and interpreted using breakpoint tables from the European Committee on Antimicrobial Susceptibility Testing v. 9.0, 2019 (http://www.eucast.org).

### Statistical analysis

We prepared a table to show the phylotype distribution per individual and visualized this as a heatmap in RStudio v 3.5.1. We carried out Fisher's exact tests to assess possible associations between the sampling site or non-human primate species and the phylogroups of *E. coli* that were observed using STATA version 14.2. We based our calculations on the assumption of independence across the observed phylogroups, i.e. the finding of one phylogroup does not predict or preclude the occurrence of another. Prior to the association tests, replicate phylogroups arising from copies of the same ST from a single individual were dropped from the analysis.

We calculated co-occurrence of the detected resistance genes among the study isolates and visualised this as a heatmap in RStudio v 3.5.1. In addition, we generated contingency tables to display the correlation between the phenotypic results and the detected resistance genes among the study isolates and calculated the percentage concordance between the genotypic and phenotypic resistances.

# RESULTS

## Study population

Twenty-four of 43 samples (56%) showed growth indicative of *E. coli*, yielding a total of 106 colonies. The isolates were designated by the primate species and the site from which they were sampled as follows: *Chlorocebus sabaeus*, 'Chlos'; *Papio papio*, 'Pap'; *Piliocolobus badius*, 'Prob'; Abuko Nature Reserve, 'AN'; Bijilo Forest Park, 'BP'; Kartong village, 'K'; Kiang West National Park, 'KW'; Makasutu Cultural Forest, 'M'; and River Gambia National Park, 'RG'. After genome sequencing, five isolates [PapRG-04, (*n*=1); PapRG-03 (*n*=1); ChlosRG-12 (*n*=1); ChlosAN-13 (*n*=1); ProbAN-19 (*n*=1)] were excluded due to low depth of coverage (<20×), leaving 101 genomes for subsequent analysis (Table 1 and File S2a).

## Distribution of sequence types and phylogroups

We recovered 43 seven-allele sequence types (ten of them novel), spanning five of the eight known phylogroups of *E. coli* and comprising 38 core-genome MLST complexes (Figs. 3 and 4). The majority of strains belonged to phylogroup B2 (42/101, 42%) – which encompasses strains that cause extraintestinal infections in humans (ExPEC strains) [4, 45, 46] – followed by B1 (35/101, 35%), A and D (8/101, 8% each), E (7/101, 7%) and cryptic clade I (1/101, 1%). Among the study isolates, we found several STs associated with extraintestinal infections and/or AMR in humans: ST73, ST681, ST127, ST226, ST336, ST349 [47–49]. We did not find any significant associations between the primate species and the distribution of phylogroups (*P*=0.17), nor between the sampling sites and phylogroups (*P*=0.44). The distribution of phylotype per individual is presented in Fig. S4.

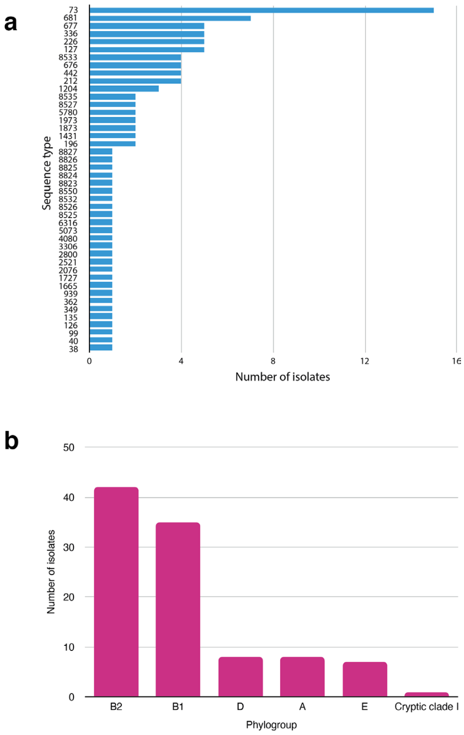


**Fig. 4.** (a) Distribution of sequence types (STs) among the study isolates. (b) Distribution of phylotypes among the study isolates.

## Prevalence of virulence factors

We detected a total of 146 virulence factors among the study isolates (Fig. 3 and File S7). The following virulence factors were largely conserved across most of the study isolates: the enterobactin-associated cluster of genes (*fepA-D, G* and *entA-F, S*), type I fimbriae (*fimA-I*) and the fimbria-associated genes (*yagV-Z*). However, iron-acquisition genes (*chuA, S-Y*) appeared to be more prevalent in strains belonging to phylogroups B2, D and E. In general, we detected a higher prevalence of virulence genes in strains belonging to phylogroup B2, compared to those from phylogroups A, B1, D and E (Fig. 3). These included additional siderophore-encoding genes (*ybt, fyu* and *irp*), capsular antigens (*kpsM1/D*), salmochelin (*iroN/C/B/D/E*), P, S and F1C fimbriae genes (*papC, D, I-K,* X, *focI/C/F* and *sfaY/B,* respectively) and the adherence factor protein gene (*fde*C) – representing colonization and fitness factors associated with extraintestinal disease in humans.

A subset of the B2 strains (13/42, 31%), belonging to STs 73, 681 and 127, carried the *pks* genomic island (*clbA-S*), which encodes the DNA alkylating genotoxin, colibactin (Fig. 3, red box). Colibactin-encoding *E. coli* frequently cause colorectal cancer, urosepsis, bacteraemia and prostatitis and commonly carry other virulence factors such as siderophores and toxins [50–52]. Also, all the ST73 (phylogroup B2) strains carried genes encoding the Serin protease autotransporter (*pic*) and 79%(33/42) of the B2 strains possessed the vacuolating autotransporter (*vat*) toxins.

Besides the B2 strains, we also detected toxins associated with intestinal and extraintestinal disease in humans among strains from other phylogroups (File S7): in particular the heat-stable enterotoxin 1 (*astA*) occurred in five isolates overall (two phylogroup B1 and one each of phylogroups E, D and the *Escherichia* cladeI); the haemolyin genes (*hlyB-D*) were detected in a single Guinea baboon (PapRG-03, phylogroup B1); the invasion of brain endothelium gene (*ibe*A) – responsible for neonatal meningitis in humans – was observed in six Guinea baboon isolates (all belonging to phylogroup B2)

**Table 2.** (a) Within-host SNP diversity between multiple genomes of the same ST recovered from the same monkey. (b) Within-host diversity in green monkey 25 (ChlosBP-25)

| (a) Sample ID | STs (colonies per ST) | Pair-wise SNP distances between multiple colonies of the same ST | Comment(s) |
|---|---|---|---|
| PapRG-03 | 336 (*n*=5) | 0–2 | |
| PapRG-04 | 1204 (*n*=2) | 4 | |
| PapRG-05 | 1431 (*n*=2) | 0 | |
| ProbRG-07 | 73 (*n*=5) | 0–1 | |
| ChlosRG-12 | 196 (*n*=2) | 25 | |
| PapAN-14 | 226 (*n*=3) | 1 | |
| PapAN-15 | 226 (*n*=2) | 1 | |
| ChlosAN-17 | 681 (*n*=3) | 0–3 | |
| ChlosAN-18 | 681 (*n*=4) | 0 | |
| ChlosBP-21 | 677 (*n*=4) | 5 | |
| ChlosBP-23 | 8527 (*n*=2) | 0 | |
| ChlosBP-24 | 73 (*n*=5) | 0–5 | |
| ChlosBP-25 | 73 (*n*=5) | 0–79 | Please see Table 2b |
| PapM-32 | 212 (*n*=4) | 0 | |
| PapM-33 | 8533 (*n*=4) | 0–4 | |
| PapM-34 | 676 (*n*=4) | 0–1 | |
| PapM-36 | 8535 (*n*=2) | 0–1 | |
| PapKW-44 | 442 (*n*=4) | 1–2 | |
| ProbK-45 | 127 (*n*=5) | 0–4 | |

In individuals where multiple colonies yielded the same genotype (*n*=19), five had entirely identicalgenotypes, while we observed a cloud of closely related genetic variants (0-5 SNPs, Table 1) in 12 individuals. However, in two monkeys (ChlosRG-12 and ChlosBP-25), pair-wise SNP comparisons suggested multiple infection events (Table 2b).

| (b) Sample ID | Clone designation |
|---|---|
| **ChlosBP-25** | |
| ChlosBP-25-1 | 1 |
| ChlosBP-25-2 | 2 |
| ChlosBP-25-3 | 2 |
| ChlosBP-25-4 | 2 |
| ChlosBP-25-5 | 3 |

**Pair-wise SNP distances between clones**

| | Clone 1 | Clone 2 | Clone 3 |
|---|---|---|---|
| Clone 1 | 0 | 12 | 79 |
| Clone 2 | 12 | 0 | 67 |
| Clone 3 | 79 | 67 | 0 |

derived from PapM33 and PapM-34 and one Green monkey (ChlosM-29).

## Within-host genomic diversity

Thirteen individuals were colonized by two or more STs and nine by two or more phylogroups (File S2a). Five colony picks from a single Guinea baboon (PapRG-06) yielded five distinct STs, two of which are novel. Two green monkeys sampled from Bijilo (ChlosBP-24 and ChlosBP-25) shared an identical ST73 genotype (zero SNP difference between the two genomes), while two Guinea baboons from Abuko shared an ST226 strain (zero SNP difference) – documenting transmission between monkeys of the same species.

In seventeen monkeys, we observed a cloud of closely related genotypes (separated by 1–5 SNPs, Table 2a) from each strain, suggesting evolution within the host after acqusition of the strain. However, in two individuals, pair-wise SNP distances between genotypes from the same ST were susbtantial enough (25 SNPs and 77 SNPs) to suggest multiple acquisitions of each strain (Table 2b). Reeves *et al.* [53] estimated a mutation rate of 1.1 SNP per genome per year from characterizing fourteen ST73 strains isolated from a single family over three years. Based on this data, with the assumption that equal rates of mutation occurred in both genomes, we can infer about 11–35 years of divergence for these strains. Thus, it is implausible that these strains represent within-host diversity and persistence in the two hosts, judging by the lifespan of a green monkey in the wild (averaged at 17 years).

## Population structure of simian *E. coli* isolates

We identified the closest neighbours of all strains from our study (Table 3). Our results suggest, in some cases, recent interactions between humans or livestock and non-human primates. However, we also found a diversity of strains specific to the non-human primate niche. Hierarchical clustering analysis revealed that simian isolates from ST442 and ST349 (Achtman) – sequence types that are associated with virulence and AMR in humans [49, 54] – were closely related to human clinical isolates, with differences of 50 alleles and seven alleles in the core-genome MLST scheme, respectively (Figs S1 and S2). Similarly, we found evidence of recent interaction between simian ST939 isolates and strains from livestock (Fig. S3) – with 40 cgMLST alleles (<40 SNPs) separating the two genomes, representing less than 18 years of divergence. Conversely, simian ST73, ST127 and ST681 isolates were genetically distinct from human isolates from these sequence types (Figs S5–S7). The only multi-drug resistant isolate (PapAN-14–1) from ST2076 was, however, closely related to an environmental isolate recovered from water (Fig. S8).

Five isolates were >1, 000 alleles away in the core-genome MLST scheme from anything in EnteroBase (Figs S9 and S10). Four of these were assigned to novel sequence types in the seven-allele scheme (Achtman) (ST8550, ST8525, ST8532, ST8826), while one belonged to ST1873, which has only two other representatives in EnteroBase: one from a species of wild bird from Australia (*Sericornis frontalis*);

**Table 2.** (a) Within-host SNP diversity between multiple genomes of the same ST recovered from the same monkey. (b) Within-host diversity in green monkey 25 (ChlosBP-25)

| (a) Sample ID | STs (colonies per ST) | Pair-wise SNP distances between multiple colonies of the same ST | Comment(s) |
|---|---|---|---|
| PapRG-03 | 336 (*n*=5) | 0–2 | |
| PapRG-04 | 1204 (*n*=2) | 4 | |
| PapRG-05 | 1431 (*n*=2) | 0 | |
| ProbRG-07 | 73 (*n*=5) | 0–1 | |
| ChlosRG-12 | 196 (*n*=2) | 25 | |
| PapAN-14 | 226 (*n*=3) | 1 | |
| PapAN-15 | 226 (*n*=2) | 1 | |
| ChlosAN-17 | 681 (*n*=3) | 0–3 | |
| ChlosAN-18 | 681 (*n*=4) | 0 | |
| ChlosBP-21 | 677 (*n*=4) | 5 | |
| ChlosBP-23 | 8527 (*n*=2) | 0 | |
| ChlosBP-24 | 73 (*n*=5) | 0–5 | |
| ChlosBP-25 | 73 (*n*=5) | 0–79 | Please see Table 2b |
| PapM-32 | 212 (*n*=4) | 0 | |
| PapM-33 | 8533 (*n*=4) | 0–4 | |
| PapM-34 | 676 (*n*=4) | 0–1 | |
| PapM-36 | 8535 (*n*=2) | 0–1 | |
| PapKW-44 | 442 (*n*=4) | 1–2 | |
| ProbK-45 | 127 (*n*=5) | 0–4 | |

In individuals where multiple colonies yielded the same genotype (*n*=19), five had entirely identicalgenotypes, while we observed a cloud of closely related genetic variants (0-5 SNPs, Table 1) in 12 individuals. However, in two monkeys (ChlosRG-12 and ChlosBP-25), pair-wise SNP comparisons suggested multiple infection events (Table 2b).

| (b) Sample ID | Clone designation | | |
|---|---|---|---|
| **ChlosBP-25** | | | |
| ChlosBP-25-1 | 1 | | |
| ChlosBP-25-2 | 2 | | |
| ChlosBP-25-3 | 2 | | |
| ChlosBP-25-4 | 2 | | |
| ChlosBP-25-5 | 3 | | |
| **Pair-wise SNP distances between clones** | | | |
| | Clone 1 | Clone 2 | Clone 3 |
| Clone 1 | 0 | 12 | 79 |
| Clone 2 | 12 | 0 | 67 |
| Clone 3 | 79 | 67 | 0 |

derived from PapM33 and PapM-34 and one Green monkey (ChlosM-29).

### Within-host genomic diversity

Thirteen individuals were colonized by two or more STs and nine by two or more phylogroups (File S2a). Five colony picks from a single Guinea baboon (PapRG-06) yielded five distinct STs, two of which are novel. Two green monkeys sampled from Bijilo (ChlosBP-24 and ChlosBP-25) shared an identical ST73 genotype (zero SNP difference between the two genomes), while two Guinea baboons from Abuko shared an ST226 strain (zero SNP difference) – documenting transmission between monkeys of the same species.

In seventeen monkeys, we observed a cloud of closely related genotypes (separated by 1–5 SNPs, Table 2a) from each strain, suggesting evolution within the host after acqusition of the strain. However, in two individuals, pair-wise SNP distances between genotypes from the same ST were susbtantial enough (25 SNPs and 77 SNPs) to suggest multiple acquisitions of each strain (Table 2b). Reeves *et al*. [53] estimated a mutation rate of 1.1 SNP per genome per year from characterizing fourteen ST73 strains isolated from a single family over three years. Based on this data, with the assumption that equal rates of mutation occurred in both genomes, we can infer about 11–35 years of divergence for these strains. Thus, it is implausible that these strains represent within-host diversity and persistence in the two hosts, judging by the lifespan of a green monkey in the wild (averaged at 17 years).

### Population structure of simian *E. coli* isolates

We identified the closest neighbours of all strains from our study (Table 3). Our results suggest, in some cases, recent interactions between humans or livestock and non-human primates. However, we also found a diversity of strains specific to the non-human primate niche. Hierarchical clustering analysis revealed that simian isolates from ST442 and ST349 (Achtman) – sequence types that are associated with virulence and AMR in humans [49, 54] – were closely related to human clinical isolates, with differences of 50 alleles and seven alleles in the core-genome MLST scheme, respectively (Figs S1 and S2). Similarly, we found evidence of recent interaction between simian ST939 isolates and strains from livestock (Fig. S3) – with 40 cgMLST alleles (<40 SNPs) separating the two genomes, representing less than 18 years of divergence. Conversely, simian ST73, ST127 and ST681 isolates were genetically distinct from human isolates from these sequence types (Figs S5–S7). The only multi-drug resistant isolate (PapAN-14–1) from ST2076 was, however, closely related to an environmental isolate recovered from water (Fig. S8).

Five isolates were >1, 000 alleles away in the core-genome MLST scheme from anything in EnteroBase (Figs S9 and S10). Four of these were assigned to novel sequence types in the seven-allele scheme (Achtman) (ST8550, ST8525, ST8532, ST8826), while one belonged to ST1873, which has only two other representatives in EnteroBase: one from a species of wild bird from Australia (*Sericornis frontalis*);

**Table 3.** Genomic relationship between study isolates and publicly available *E. coli* genomes

| Seven-allele ST | HC100 subgroups | Non-human primate host | Closest neighbours' source | Neighbours' country of isolation | Allelic distance |
|---|---|---|---|---|---|
| **349** | – | *Chlorocebus sabaeus* 18 | Human (bloodstream infection) | Canada | 7 |
| **2076** | – | *Papio papio* 14 | Environment (water) | Unknown | 25 |
| **939** | – | *Papio papio* 14 | Livestock | US | 40 |
| **442** | – | *Papio papio* 44 | Human | China | 50 |
| **2800** | – | *Papio papio* 31 | Unknown | Vietnam | 59 |
| **1973** | – | *Chlorocebus sabaeus* 13 | Unknown | Unknown | 64 |
| **8533** | – | *Papio papio* 33 | Environment (water) | Unknown | 69 |
| **6316** | – | *Papio papio* 05 | Human | Kenya | 97 |
| **1727** | – | *Papio papio* 34 | Human | Kenya | 98 |
| **676** | – | *Papio papio* 34 | Human (bloodstream infection) | UK | 98 |
| **8823** | – | *Papio papio* 15 | Rodent (guinea pig) | Kenya | 101 |
| **1431** | – | *Papio papio* 05 | Human | US | 109 |
| **5073** | – | *Papio papio* 15 | Human | US | 112 |
| **226** | 73 641 | *Papio papio* 14 | Human | Tanzania | 112 |
| **8827** | – | *Papio papio* 06 | Human | Unknown | 122 |
| **1204** | 83 197 | *Papio papio* 04 | Livestock | Japan | 127 |
| **1204** | 83 197 | *Papio papio* 04 | Livestock | Japan | 130 |
| **677** | – | *Chlorocebus sabaeus* 21 | Human | US | 132 |
| **40** | – | *Chlorocebus sabaeus* 12 | Human | UK | 137 |
| **1204** | 83 164 | *Papio papio* 06 | Livestock | Japan | 173 |
| **99** | – | *Papio papio* 05 | Human | UK | 180 |
| **362** | – | *Chlorocebus sabaeus* 17 | Food | Kenya | 180 |
| **8825** | – | *Piliocolobus badius* 19 | Human | France | 189 |
| **336** | – | *Papio papio* 03 | Poultry | Kenya | 189 |
| **73** | – | *Chlorocebus sabaeus* 24 | Human | Sweden | 189 |
| **196** | – | *Chlorocebus sabaeus* 12 | Human | Sweden | 197 |
| **2521** | – | *Papio papio* 06 | Livestock | US | 201 |
| **127** | | *Pioliocolobus badius* 45 | Companion animal | US | 229 |
| **681** | | *ChlosAN* 17 | Human | Norway | 251 |
| **38** | – | *Papio papio* 33 | human | UK | 265 |
| **135** | – | *Papio papio* 31 | Poultry | US | 281 |
| **8824** | – | *Chlorocebus sabaeus* 12 | Environmental* | US | 296 |
| **226** | 100 039 | *Papio papio* 14 | Human | Sri Lanka | 318 |
| **8527** | – | *Chlorocebus sabaeus* 23 | Human | Kenya | 323 |
| **8535** | – | *Papio papio* 36 | Environmental (soil) | US | 368 |
| **1665** | – | *Papio papio* 04 | Livestock | UK | 371 |
| **4080** | – | *Papio papio* 06 | Human | Denmark | 507 |

Continued

Table 3. Continued

| Seven-allele ST | HC100 subgroups | Non-human primate host | Closest neighbours' source | Neighbours' country of isolation | Allelic distance |
|---|---|---|---|---|---|
| **8526** | – | *Chlorocebus sabaeus* 13 | Livestock | US | 708 |
| **8532** | – | *Papio papio* 32 | Non-human primate | Gambia (PapM-31–3) | 1102 |
| **8826** | – | *Papio papio* 04 | Livestock | Mozambique | 1255 |
| **8525** | – | *Papio papio* 06 | Livestock/companion animal | Switzerland | 1659 |
| **1873** | – | *Chorocebus sabaeus* 29 | Environment | US | 1685 |
| **8550** | – | *Chlorocebus sabaeus* 13 | Unknown | Unknown | 2006 |

*Source details unknown.
Isolates from humans were recovered from stools, except where indicated otherwise.

the other from water. Besides, ST8550, ST8525, ST8532 and ST8826 belonged to novel HierCC 1100 groups (cgST complexes), indicating that they were distinct from any other publicly available *E. coli* genomes.

Besides our study isolates, there were 94 *E. coli* genomes sourced from non-human primates from the rest of the world within EnteroBase: the USA (83), Uganda (6), Kenya (4), Mexico (1). A total of 52 STs were found among these primates from other parts of the world (Fig. S11a), four of which were also found among our study isolates (ST 73, ST127, ST681 and ST939). As observed in our monkey isolates, the most common ST among primates from the rest of the world was ST73 (11%). Also, most of the non-Gambian primate isolates belonged to phylogroup B2 (41%) and B1 (21%), consistent with what we found in our study population (Fig. S11b). Hierarchical clustering based on cgMLST types revealed clustering patterns that were largely consistent with the phylotype designations to which the primate isolates belonged. No discernible segregation of primate *E. coli* phylotypes based on geography was observed.

### Prevalence of AMR-associated genes

We observed a modest prevalence of genotypic antimicrobial resistance in our study population. The AMR-associated genotypes we found among the monkey isolates included *bla*EC (beta-lactamase, penicillinase-type), *ant(3') (aadA1)* (streptomycin and spectinomycin), *aph3/aph6* (neomycin and kanamycin), DHFR (trimethoprim), *sul1* (sulphonamides) and *tetA/B/C/D/R* (tetracyclines) (Fig. 3). A total of 22 isolates encoded resistance genes to a single antibiotic agent; 22 to two antibiotic classes and three isolates to three or more antibiotic classes. Pair-wise co-occurrence of AMR-associated genes in the same genome was sparse. The most common gene network was *bla*EC-*tetA/B/C/D/R* (12%), followed by *bla*EC- *ant(3') (aadA1)* (5%), DHFR-*tetA/B/C/D/R* (3%), then *ant(3') (aadA1)*-DHFR (2%) (Fig. S12). Although phenotypic susceptibility tests were performed for all the isolates, phenotypic resistance to single agents was confirmed in ten isolates only: to trimethoprim in a single isolate, to sulfamethoxazole in four unrelated isolates and to tetracycline in four closely related isolates from a single animal. A single ST2076 (Achtman) isolate (PapAN-14-1) belonging to the ST349 lineage was phenotypically resistant to trimethoprim, sulfamethoxazole and tetracycline (multi-drug resistance). The associated resistance genes were harboured on an IncFIB plasmid. The genotypic resistance predictions were largely concordant with the results of phenotypic testing (range, 90–99%, File S3). Due to logistic constraints, we could not carry out phenotypic confirmation of the predicted penicillinase-type beta-lactamase resistance.

A higher prevalence of genotypic antimicrobial resistance was found in *E. coli* isolates from humans in the Gambia, compared to what prevails in the monkey isolates (Fig. 5). Notably, a range of beta-lactamase resistance genes were found among *E. coli* from humans in the Gambia (*bla*OXA-*1, bla*TEM-*1B, bla*TEM-*1B, bla*TEM-*1C, bla*SHV-*1*), while only the *bla*EC gene occurred in our study isolates.

### Prevalence of plasmid replicons

Eighty percent (81/101) of the study isolates harboured one or more plasmids. We detected the following plasmid replicon types: IncF (various subtypes), IncB/K/O/Z, I1, IncX4, IncY, Col plasmids (various subtypes) and plasmids related to p0111 (rep B) (File S4a). Long-read sequencing of six representative samples showed that the IncFIB plasmids encoded acquired antibiotic resistance, fimbrial adhesins and colicins (File S4b). Also, the IncFIC/FII, ColRNAI, Col156 and IncB/O/K/Z plasmids encoded fimbrial proteins and colicins. Besides, the IncX and Inc-I-Aplha encoded bundle forming pili *bfp*B and the heat-stable enterotoxin protein *StbB,* respectively.

### Polished assemblies of novel strains

We generated complete genome sequences of five novel sequence types of *E. coli* (ST8525, ST8527, ST8532, ST8826, ST8827) within the seven-allele scheme (Achtman) (File S4a). Although none of these new genomes encoded AMR genes, one of them (PapRG-04–4) contained an IncFIB plasmid

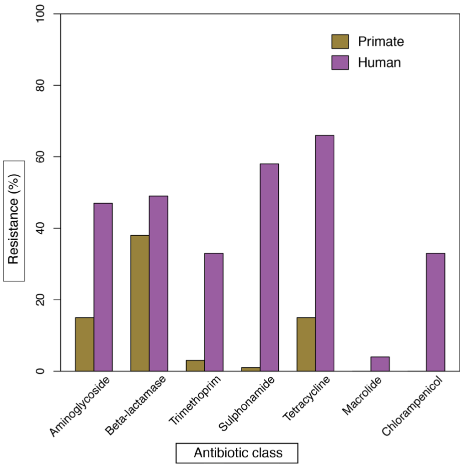


**Fig. 5.** Bar graph comparing the prevalence of antimicrobial resistance genotypes in *E. coli* isolated from humans in the Gambia (*n*=128) as found in EnteroBase [30] to that found among the study isolates (*n*=101). The antimicrobial resistance genes detected were as follows: Aminoglycoside: *aph*(6)-Id, ant *aac*(3)-IIa, *ant*(3'')-Ia, *aph*(3'')-Ib, *aad*A1, *aad*A2; Beta-lactamase: *bla*EC, *bla*OXA-1, *bla*TEM-1B, *bla*TEM-1B, *bla*TEM-1C, *bla*SHV1; Trimethoprim: *dfrA*/DHFR; Sulphonamide: *sul1*, *sul2*; Tetracycline: *tet*(A), *tet*(B), *tet*(34), *tet*(D); *tet*(R) Macrolide, *mph*(A); Chloramphenicol, catA1. Screening of resistance genes was carried out using ARIBA ResFinder [38] and confirmed by ABRicate (*https://github.com/tseemann/abricate*). A percentage identity of ≥ 90% and coverage of ≥70% of the respective gene length were taken as a positive result.

encoding fimbrial proteins and a cryptic ColRNA plasmid. PHASTER identified thirteen intact prophages and four incomplete phage remnants (File S4B). Two pairs of genomes from Guinea baboons from different parks shared common prophages: one pair carrying PHAGE_Entero_933W, the other PHAGE-Entero_lambda.

## DISCUSSION

We have described the population structure of *E. coli* in diurnal non-human primates living in rural and urban habitats from the Gambia. Although our sample size was relatively small, we have recovered isolates that span the diversity previously described in humans and have also identified ten new sequence types (five of them now with complete genome sequences). This finding is significant, considering the vast number of *E. coli* genomes that have been sequenced to date (9, 597 with MLST via sanger sequencing and 127, 482 via WGS) [30].

Increasing contact between animal species facilitates the potential exchange of pathogens [55]. Accumulating data shows that ExPEC strains are frequently isolated from diseased companion animals and livestock – highlighting the potential for zoonotic as well as anthroponotic transmission [54, 56]. In a previous study, green monkeys from Bijilo Park were found to carry lineages of *Staphylococcus aureus* thought to be acquired from humans [16]. Our analyses similarly suggest exchange of *E. coli* strains between humans and wild non-human primates – with only seven cgMLST alleles separating a simian ST349 isolate and a human bloodstream isolate from Canada. This simian ST349 isolate was recovered from a green monkey in Abuko Nature Reserve, where tourists sometimes handfeed monkeys, despite prohibitions. A limitation of our study is that we could not sample *E. coli* from humans living in close proximity to the study primates. Comparisons between simian isolates and those from sympatric humans may shed light on possible transmission routes between humans and primates in this setting. However, beside human–monkey or monkey–human transmission, it is possible for the spread of pathogenic strains to have originated from an environmental reservoir to both humans and monkeys. Our results also show that non-human primates harbour *E. coli* genotypes that are clinically important in humans, such as ST73, ST127 and ST681, yet are distinct from those circulating in humans – probably reflecting lineages that have existed in this niche for long periods.

We found that several monkeys were colonized with multiple STs, often encompassing two or more phylotypes. Colonization with multiple serotypes of *E. coli* is common in humans [20–23]. Our results indicate that a single monkey can carry as many as five STs. Sampling multiple colonies from single individuals also revealed within-host diversity arising from microevolution. However, we also found evidence of acquisition in the same animal of multiple lineages of the same sequence type, although it is unclear whether this reflects a single transmission event involving more than one strain or serial transfers.

We found a relatively lower prevalence of genotypic antimicrobial resistance among our study isolates, compared to the genotypic resistance observed among isolates sourced from humans in the Gambia – probably reflecting differing selective pressures from antibiotic use. The Gambia does not have national AMR surveillance data and background data on the use of antimicrobials is limited. However, a recent study on the aetiology of diarrhoea among children less than 5 years old reported the frequent use of trimethoprim/sulphamethoxazole in the treatment of diarrhoea in the Gambia [57]. This probably accounts, at least in part, for the observed high rates of genotypic resistance to trimethoprim and sulphonamides among human *E. coli* isolates from the Gambia. The excretion of resistant bacteria and active antimicrobials from humans and domesticated animals and their persistence in the environment is known to facilitate the proliferation of AMR in the environment [19].

Antimicrobial resistance in wildlife is known to spread on plasmids through horizontal gene transfer [58]. Given the challenge of resolving large plasmids using short-read sequences [59], we exploited long-read sequencing to document the contribution of plasmids to the genomic diversity that we observed in our study population. Consistent with previous reports [60], we found IncF plasmids, which encoded antimicrobial resistance genes. Virulence-encoding plasmids, particularly colicin-encoding and the F incompatibility group ones, have long been associated with several pathotypes of *E. coli* [61]. Consistent with this, we found plasmids that contributed to the dissemination of virulence factors such as the heat-stable enterotoxin protein *StbB*, colicins and fimbrial proteins.

This study could have been enhanced by sampling human populations living near those of our non-human primates. We compensated for this limitation by leveraging the wealth of genomes in publicly available databases. Furthermore, we did not sample nocturnal monkeys due to logistic challenges; however, these have more limited contact with humans than the diurnal species. Despite these limitations, our study provides insight into the diversity and population structure of *E. coli* among non-human primates in the Gambia, highlighting the impact of human continued encroachment on natural habitats and revealing important phylogenomic relationships between strains from humans and non-human primates.

Funding information
M.J.P., E.F.N., N.M.T., A.R., G.T., J.O. and G.K. were supported by the BBSRC Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent projects 44414000A and 4408000A. N.F.A. and D.B. were supported by the Quadram Institute Bioscience BBSRC funded Core Capability Grant (project number BB/ CCG1860/1). The sample collection was funded by a faculty research grant to J.D.C. by the American Public University System. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements
We want to thank Dr Andrew Page and Dr Thanh Le Viet for their thoughtful advice on the long-read analysis. We also thank Dr Mark Webber for proofreading the manuscript and giving constructive feedback.

Author contributions
Conceptualization, M.A, M.J.P.; data curation, M.J.P., N.F.A.; formal analysis, E.F.N., analytical support, G.T.; funding, M.J.P. and M.A.; sample collection, J.D.C., laboratory experiments, E.F.N., D.B.; supervision, A.R., N.F.A., G.K., J.O., M.P., M.A.; manuscript preparation – original draft, E.F.N.; review and editing, N.M.T., A.R., J.O., N.F.A., M.J.P.; review of final manuscript, all authors.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Ethical statement
This study was approved by the Gambian Department of Parks and Wildlife Management and the joint MRC-Gambia Scientific Coordinating Committee and Ethics Committee.

References

1. Sousa CP. The versatile strategies of *Escherichia coli* pathotypes: a mini review. *J Venom Anim Toxins Incl Trop Dis* 2006;12:363–373.
2. Kaper JB, Nataro JP, Mobley HLT. Pathogenic *Escherichia coli*. *Nat Rev Microbiol* 2004;2:123–140.
3. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol* 2010;8:207–217.
4. Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O. ClermonTyping: an easy-to-use and accurate in silico method for *Escherichia* genus strain phylotyping. *Microbial Genomics* 2018;4:e000192.
5. Clayton JB, Danzeisen JL, Trent AM, Murphy T, Johnson TJ. Longitudinal characterization of *Escherichia coli* in healthy captive non-human primates. *Front Vet Sci* 2014;1:24.
6. Feng Y, Mannion A, Madden CM, Swennes AG, Townes C *et al*. Cytotoxic *Escherichia coli* strains encoding colibactin and cytotoxic necrotizing factor (CNF) colonize laboratory macaques. *Gut Pathog* 2017;9:71.
7. Thomson JA, Scheffler JJ. Hemorrhagic typhlocolitis associated with attaching and effacing *Escherichia coli* in common marmosets. *Laboratory Animal Science* 1996;46:275–279.
8. Mansfield KG, Lin K-C, Newman J, Schauer D, MacKey J *et al*. Identification of enteropathogenic *Escherichia coli* in simian immunodeficiency virus-infected infant and adult rhesus macaques. *J Clin Microbiol* 2001;39:971–976.
9. Mansfield KG, Lin Kuei-Chin, Xia D, Newman JV, Schauer DB *et al*. Enteropathogenic *Escherichia coli* and ulcerative colitis in cotton-top tamarins (*Saguinus oedipus*). *J Infect Dis* 2001;184:803–807.
10. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol* 2008;6:776–788.
11. Moeller AH, Caro-Quintero A, Mjungu D, Georgiev AV, Lonsdorf EV *et al*. Cospeciation of gut microbiota with hominids. *Science* 2016;353:380–382.
12. Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D *et al*. Meta-analyses of studies of the human microbiota. *Genome Res* 2013;23:1704–1714.
13. RdO I, Dejuste C, Miranda F, Filoni C, Bueno MG *et al*. Isolation of *Escherichia coli* and *Salmonella* spp. from free-ranging wild animals. *Brazillian Journal of Microbiology* 2015;46:1257–1263.
14. Bublitz DC, Wright PC, Rasambainarivo FT, Arrigo-Nelson SJ, Bodager JR *et al*. Pathogenic enterobacteria in lemurs associated with anthropogenic disturbance. *Am J Primatol* 2015;77:330–337.
15. Weiss D, Wallace RM, Rwego IB, Gillespie TR, Chapman CA *et al*. Antibiotic-resistant *Escherichia coli* and Class 1 integrons in humans, domestic animals, and wild primates in Rural Uganda. *Appl Environ Microbiol* 2018;84:e01632–18.
16. Senghore M, Bayliss SC, Kwambana-Adams BA, Foster-Nyarko E, Manneh J *et al*. Transmission of *Staphylococcus aureus* from humans to green monkeys in the Gambia as revealed by whole-genome sequencing. *Appl Environ Microbiol* 2016;82:5910–5917.
17. Goldberg TL, Gillespie TR, Rwego IB, Estoff EL, Chapman CA. Forest fragmentation as cause of bacterial transmission among nonhuman primates, humans, and livestock, Uganda. *Emerg Infect Dis* 2008;14:1375–1382.
18. Rwego IB, Isabirye-Basuta G, Gillespie TR, Goldberg TL. Gastrointestinal bacterial transmission among humans, mountain gorillas, and livestock in Bwindi impenetrable National Park, Uganda. *Conserv Biol* 2008;22:1600–1607.
19. Arnold KE, Williams NJ, Bennett M. 'Disperse abroad in the land': the role of wildlife in the dissemination of antimicrobial resistance. *Biol Lett* 2016;12:20160137.
20. Chen SL, Wu M, Henderson JP, Hooton TM, Hibbing ME *et al*. Genomic diversity and fitness of *E. coli* strains recovered from the intestinal and urinary tracts of women with recurrent urinary tract infection. *Sci Transl Med* 2013;5:184ra60.
21. Richter TKS, Hazen TH, Lam D, Coles CL, Seidman JC *et al*. Temporal variability of *Escherichia coli* diversity in the gastrointestinal tracts of Tanzanian children with and without exposure to antibiotics. *mSphere* 2018;3:e00558–18.
22. Dixit OVA, O'Brien CL, Pavli P, Gordon DM. Within-host evolution *versus* immigration as a determinant of *Escherichia coli* diversity in the human gastrointestinal tract. *Environ Microbiol* 2018;20:993–1001.
23. Schlager TA, Hendley JO, Bell AL, Whittam TS. Clonal diversity of *Escherichia coli* colonizing stools and urinary tracts of young girls. *Infect Immun* 2002;70:1225–1229.

24. Lidin-Janson G, Kaijser B, Lincoln K, Olling S, Wedel H. The homogeneity of the faecal coliform flora of normal school-girls, characterized by serological and biochemical properties. *Med Microbiol Immunol* 1978;164:247–253.

25. Connor TR, Loman NJ, Thompson S, Smith A, Southgate J *et al*. CLIMB (the cloud infrastructure for microbial bioinformatics): an online resource for the medical microbiology community. *Microbial Genomics* 2016;2:e00008.

26. Wingett SW, Andrews S. FastQ screen: a tool for multi-genome mapping and quality control. *F1000Res* 2018;7:1338.

27. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–1075.

28. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.

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

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

31. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43:e15.

32. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.

33. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 2013;30:2725–2729.

34. Frentrup M, Zhou Z, Steglich M, Meier-Kolthoff JP, Göker M *et al*. Global genomic population structure of *Clostridioides difficile*. *BioRxiv* 2019;727230.

35. Wheeler TJ. Large-Scale neighbor-joining with NINJA. in: proceedings of the 9th workshop on algorithms in bioinformatics 2009;5724:375–389.

36. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J *et al*. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microbial Genomics* 2017;3:e000131.

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

38. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S *et al*. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* 2012;67:2640–2644.

39. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O *et al*. *In Silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 2014;58:3895–3903.

40. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* 2012;28:464–469.

41. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;13:e1005595.

42. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;25:1043–1055.

43. Arndt D, Grant JR, Marcu A, Sajed T, Pon A *et al*. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;44:W16–W21.

44. Wiegand I, Hilpert K, Hancock REW. Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat Protoc* 2008;3:163–175.

45. Escobar-Páramo P, Clermont O, Blanc-Potard A-B, Bui H, Le Bouguénec C *et al*. A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. *Mol Biol Evol* 2004;21:1085–1094.

46. Picard B, Garcia JS, Gouriou S, Duriez P, Brahimi N *et al*. The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. *Infect Immun* 1999;67:546–553.

47. Manges AR, Johnson JR. Reservoirs of extraintestinal pathogenic *Escherichia coli*. *Microbiology Spectrum* 2015;3:UTI-0006-2012..

48. Kamjumphol W, Wongboot W, Suebwongsa N, Kluabwang P, Chantaroj S *et al*. Draft genome sequence of a colistin-resistant *Escherichia coli* ST226: A clinical strain harbouring an mcr-1 variant. *J Glob Antimicrob Resist* 2019;16:168–169.

49. Manges AR, Geum HM, Guo A, Edens TJ, Fibke CD *et al*. Global Extraintestinal Pathogenic *Escherichia coli* (ExPEC) Lineages. *Clin Microbiol Rev* 2019;32:e00135–18.

50. Krieger JN, Dobrindt U, Riley DE, Oswald E. Acute *Escherichia coli* prostatitis in previously health young men: bacterial virulence factors, antimicrobial resistance, and clinical outcomes. *Urology* 2011;77:1420–1425.

51. Faïs T, Delmas J, Barnich N, Bonnet R, Dalmasso G. Colibactin: more than a new bacterial toxin. *Toxins* 2018;10:151.

52. Johnson JR, Johnston B, Kuskowski MA, Nougayrede J-P, Oswald E. Molecular epidemiology and phylogenetic distribution of the *Escherichia coli* pks genomic island. *J Clin Microbiol* 2008;46:3906–3911.

53. Reeves PR, Liu B, Zhou Z, Li D, Guo D *et al*. Rates of mutation and host transmission for an *Escherichia coli* clone over 3 years. *PLoS One* 2011;6:e26907.

54. Zogg AL, Zurfluh K, Schmitt S, Nüesch-Inderbinen M, Stephan R. Antimicrobial resistance, multilocus sequence types and virulence profiles of ESBL producing and non-ESBL producing uropathogenic *Escherichia coli* isolated from cats and dogs in Switzerland. *Vet Microbiol* 2018;216:79–84.

55. Keusch GT, Pappaioanou M, Gonzalez MC, National Research Council (US) Committee on Achieving Sustainable Global Capacity for Surveillance and Response to Emerging Diseases of Zoonotic Origin. *Sustaining Global Surveillance and Response to Emerging Zoonotic Diseases*, 3. Washington, DC: National Academies Press (US); 2009.

56. Ewers C, Grobbel M, Stamm I, Kopp PA, Diehl I *et al*. Emergence of human pandemic O25:H4-ST131 CTX-M-15 extended-spectrum-beta-lactamase-producing *Escherichia coli* among companion animals. *J Antimicrob Chemother* 2010;65:651–660.

57. Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH *et al*. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the global enteric multicenter study, GEMs): a prospective, case-control study. *The Lancet* 2013;382:209–222.

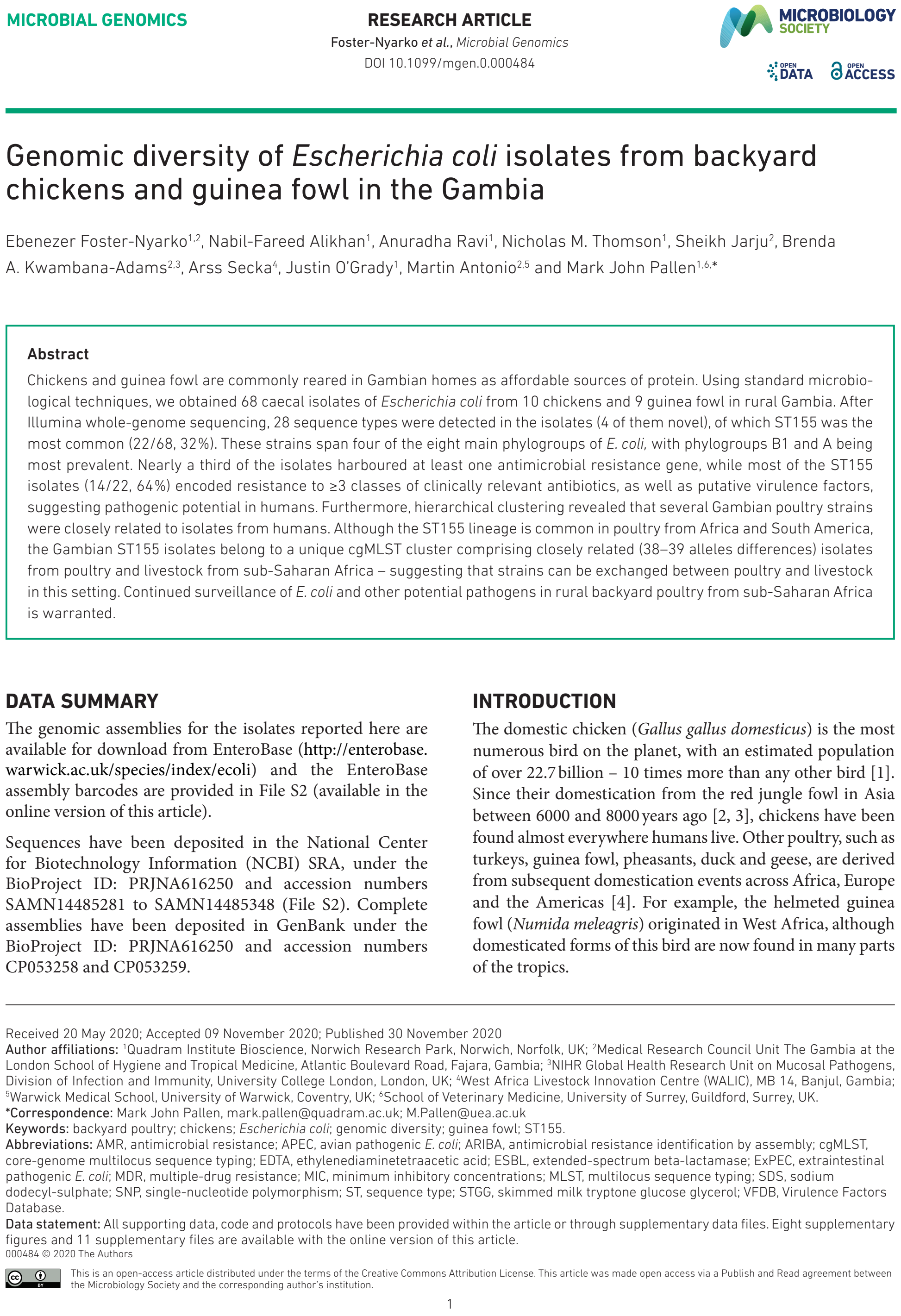
58. Vittecoq M, Godreuil S, Prugnolle F, Durand P, Brazier L *et al*. Antimicrobial resistance in wildlife. *J Appl Ecol* 2016;53:519–529.

59. Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microbial Genomics* 2017;3:e000128.

60. Carattoli A. Resistance plasmid families in *Enterobacteriaceae*. *Antimicrob Agents Chemother* 2009;53:2227–2238.

61. Johnson TJNolan LK, . Pathogenomics of the virulence plasmids of *Escherichia coli*. *Microbiol Mol Biol Rev* 2009;73:750–774.

# APPENDIX 2: Genomic diversity of Escherichia coli isolates from backyard chickens and guinea fowl in the Gambia



MICROBIAL GENOMICS

RESEARCH ARTICLE
Foster-Nyarko *et al.*, *Microbial Genomics*
DOI 10.1099/mgen.0.000484

## Genomic diversity of *Escherichia coli* isolates from backyard chickens and guinea fowl in the Gambia

Ebenezer Foster-Nyarko[1,2], Nabil-Fareed Alikhan[1], Anuradha Ravi[1], Nicholas M. Thomson[1], Sheikh Jarju[2], Brenda A. Kwambana-Adams[2,3], Arss Secka[4], Justin O'Grady[1], Martin Antonio[2,5] and Mark John Pallen[1,6,*]

### Abstract

Chickens and guinea fowl are commonly reared in Gambian homes as affordable sources of protein. Using standard microbiological techniques, we obtained 68 caecal isolates of *Escherichia coli* from 10 chickens and 9 guinea fowl in rural Gambia. After Illumina whole-genome sequencing, 28 sequence types were detected in the isolates (4 of them novel), of which ST155 was the most common (22/68, 32%). These strains span four of the eight main phylogroups of *E. coli,* with phylogroups B1 and A being most prevalent. Nearly a third of the isolates harboured at least one antimicrobial resistance gene, while most of the ST155 isolates (14/22, 64%) encoded resistance to ≥3 classes of clinically relevant antibiotics, as well as putative virulence factors, suggesting pathogenic potential in humans. Furthermore, hierarchical clustering revealed that several Gambian poultry strains were closely related to isolates from humans. Although the ST155 lineage is common in poultry from Africa and South America, the Gambian ST155 isolates belong to a unique cgMLST cluster comprising closely related (38–39 alleles differences) isolates from poultry and livestock from sub-Saharan Africa – suggesting that strains can be exchanged between poultry and livestock in this setting. Continued surveillance of *E. coli* and other potential pathogens in rural backyard poultry from sub-Saharan Africa is warranted.

## DATA SUMMARY

The genomic assemblies for the isolates reported here are available for download from EnteroBase (http://enterobase.warwick.ac.uk/species/index/ecoli) and the EnteroBase assembly barcodes are provided in File S2 (available in the online version of this article).

Sequences have been deposited in the National Center for Biotechnology Information (NCBI) SRA, under the BioProject ID: PRJNA616250 and accession numbers SAMN14485281 to SAMN14485348 (File S2). Complete assemblies have been deposited in GenBank under the BioProject ID: PRJNA616250 and accession numbers CP053258 and CP053259.

## INTRODUCTION

The domestic chicken (*Gallus gallus domesticus*) is the most numerous bird on the planet, with an estimated population of over 22.7 billion – 10 times more than any other bird [1]. Since their domestication from the red jungle fowl in Asia between 6000 and 8000 years ago [2, 3], chickens have been found almost everywhere humans live. Other poultry, such as turkeys, guinea fowl, pheasants, duck and geese, are derived from subsequent domestication events across Africa, Europe and the Americas [4]. For example, the helmeted guinea fowl (*Numida meleagris*) originated in West Africa, although domesticated forms of this bird are now found in many parts of the tropics.

Received 20 May 2020; Accepted 09 November 2020; Published 30 November 2020

Author affiliations: [1]Quadram Institute Bioscience, Norwich Research Park, Norwich, Norfolk, UK; [2]Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine, Atlantic Boulevard Road, Fajara, Gambia; [3]NIHR Global Health Research Unit on Mucosal Pathogens, Division of Infection and Immunity, University College London, London, UK; [4]West Africa Livestock Innovation Centre (WALIC), MB 14, Banjul, Gambia; [5]Warwick Medical School, University of Warwick, Coventry, UK; [6]School of Veterinary Medicine, University of Surrey, Guildford, Surrey, UK.
*Correspondence: Mark John Pallen, mark.pallen@quadram.ac.uk; M.Pallen@uea.ac.uk

Keywords: backyard poultry; chickens; *Escherichia coli*; genomic diversity; guinea fowl; ST155.

Abbreviations: AMR, antimicrobial resistance; APEC, avian pathogenic *E. coli*; ARIBA, antimicrobial resistance identification by assembly; cgMLST, core-genome multilocus sequence typing; EDTA, ethylenediaminetetraacetic acid; ESBL, extended-spectrum beta-lactamase; ExPEC, extraintestinal pathogenic *E. coli*; MDR, multiple-drug resistance; MIC, minimum inhibitory concentrations; MLST, multilocus sequence typing; SDS, sodium dodecyl-sulphate; SNP, single-nucleotide polymorphism; ST, sequence type; STGG, skimmed milk tryptone glucose glycerol; VFDB, Virulence Factors Database.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Eight supplementary figures and 11 supplementary files are available with the online version of this article.

Poultry are reared for meat, eggs and feathers [5]. Poultry production is classified into four sectors, based on the marketing of poultry products and the level of biosecurity [6]. Intensive poultry farming falls under sectors 1 to 3, characterized by moderate to high levels of biosecurity, while sector 4 pertains to the 'backyard', 'village' or 'family' poultry system, with few or no biosecurity measures.

Backyard poultry fulfil important social, economic and cultural roles in many societies. Seventy per cent of poultry production in low-income countries comes from backyard poultry [7]. The sale of birds and eggs generates income, while occasional consumption of poultry meat provides a source of protein in the diet. It is estimated that meat and eggs from backyard poultry contribute about 30% of the total animal protein supply of households in low-income countries [8]. In rural Gambia, backyard poultry can be offered as gifts for newlyweds or sold to solve family needs such as paying school fees, buying new clothes or other household needs [9]. The proximity between backyard poultry and humans may facilitate transmission of pathogens such as *Escherichia coli* between the two host species.

*E. coli* is a generalist bacterium that commonly colonizes the gastrointestinal tract of mammals and avian species [10]. Based on their pathogenic potential, *E. coli* can be divided into three categories: commensals, diarrhoeagenic *E. coli* and extraintestinal pathogenic *E. coli* (ExPEC). ExPEC frequently colonize the gut asymptomatically; however, they possess a wide range of unique virulence factors that enable them to colonize extraintestinal tissues in humans, pets and poultry [11, 12]. A sub-pathotype of ExPEC strains, known as avian pathogenic *E. coli* (APEC), causes colibacillosis – an extraintestinal disease in birds, with manifestations such as septicaemia, air sacculitis and cellulitis [13]. As a result of the high mortality and condemnation of birds associated with avian colibacillosis [14], antimicrobials are often used in intensive farming systems to prevent bacterial infections and treat sick birds – a practice that has been linked to the development of antimicrobial resistance (AMR) in poultry.

Although previous studies have focused on the detection of AMR and documented the emergence of multiple-drug resistance (MDR) in this niche [15–18], little is known about the population structure of *E. coli* in rural backyard poultry. The Gambia does not have genomic data on *E. coli* from poultry prior to this study, and data on the circulating MLST types among poultry *E. coli* strains from sub-Saharan Africa is limited. However, reports from Ghana, Senegal and Nigeria have indicated the prevalence of ST624, ST69, ST540, ST7473, ST155, ST297, ST226, ST10, ST3625 and ST58 among *E. coli* isolates from commercial poultry [19–22]. Given the increased exposure to humans, the natural environment and other animals, the population of *E. coli* in birds raised under the backyard system may differ considerably from those reared in intensive systems. It is also possible that the lineages of *E. coli* within local genotypes of rural poultry might differ between geographical regions. Previous studies have suggested that several *E. coli* clones are shared between poultry and humans, including isolates recovered from clinical cases. These include ST10, ST69, ST95, ST117, ST131, ST155, ST371, ST100, ST88 and ST23, ST38, ST3541, ST3018, ST58, ST6359, ST1011, ST746 and ST2676 [21, 23–29]. The absence of biosecurity measures in backyard poultry farming increases the potential for zoonotic transmission of pathogenic and/or antimicrobial-resistant strains to humans.

**Impact Statement**

Domestic birds play a crucial role in human society, in particular contributing to food security in low-income countries. Many households in sub-Saharan Africa rear free-range chickens and guinea fowl, which are often left to scavenge for feed in and around the family compound, where they are frequently exposed to humans, other animals and the environment. Such proximity between backyard poultry and humans is likely to facilitate transmission of pathogens such as *Escherichia coli* or antimicrobial resistance between the two host species. Little is known about the population structure of *E. coli* in rural chickens and guinea fowl, although this information is needed to contextualize the potential risks of transmission of bacterial strains between humans and rural backyard poultry. Thus, we sought to investigate the genomic diversity of *E. coli* in backyard poultry from rural Gambia.

In a recent study of commercial broiler chickens, multiple colony sampling revealed that a single broiler chicken could harbour up to nine sequence types of *E. coli* [30]. However, within-host diversity of *E. coli* in backyard poultry, particularly in guinea fowl, has not been well studied and so we do not know how many lineages of *E. coli* can co-colonize a single backyard bird. To address these gaps in our knowledge, we exploited whole-genome sequencing to investigate the genomic diversity and burden of AMR among *E. coli* isolates from backyard chickens and guinea fowl in rural Gambia, West Africa.

## METHODS

### Study population

The study population comprised 10 local-breed chickens and 9 guinea fowl from a village in Sibanor in the Western Region of the Gambia (Table 1). Sibanor covers an area of approximately 90 km$^2$ and is representative of rural areas in the Gambia [31]. It has a population of about 10, 000. Most of the villagers are subsistence farmers growing peanuts, maize and millet. Households within this community comprise extended family units of up to 15 people, which make up the 'compound'. All guinea fowl were of the pearl variety, characterized by purplish-grey feathers dotted with white.

**Table 1.** Characteristics of the study population

| Sample ID | Poultry species | Gender | Household | Colony picks | Recovered sequence types (No. of colonies per ST) | Phylogroup distribution (STs per phylogroup) |
|---|---|---|---|---|---|---|
| C1 | Chicken | Rooster | 1 | No *E. coli* isolated | | |
| C2 | Chicken | Hen | 3 | 1 | 155 (1) | B1 (155) |
| C3 | Chicken | Rooster | 2 | 5 | 155 (1), 48 (1), 746 (1) 2461 (1), 542 (1) | A (48, 746, 2461, 542), B1 (155) |
| C4 | Chicken | Rooster | 2 | 5 | 1423 (1), 337 (1), 9285* (1), 540 (1), 58 (1) | A (540), B1 (1423, 337, 9285*, 58) |
| C5 | Chicken | Hen | 2 | 2 | 155 (2) | B1 (155) |
| C6 | Chicken | Rooster | 2 | 5 | 155 (3), 9284* (2) | B1 (155), E (9284*) |
| C7 | Chicken | Rooster | 3 | 5 | 155 (4), 602 (1) | B1 (155, 602) |
| C8 | Chicken | Rooster | 4 | 5 | 5286 (1), 2772 (2), 6186 (1), 2165 (1) | A (5286), B1 (2772, 6186, 2165) |
| C9 | Chicken | Hen | 5 | No *E. coli* isolated | | |
| C10 | Chicken | Rooster | 5 | No *E. coli* isolated | | |
| GF1 | Guinea fowl | Rooster | 1 | 5 | 540 (5) | A (540) |
| GF2 | Guinea fowl | Rooster | 1 | 5 | 155 (4), 540 (1) | A (540), B1 (155) |
| GF3 | Guinea fowl | Rooster | 3 | 5 | 540 (2), 443 (1), 6025 (1), 10654* (1) | A (540), B1 (443), D (6025), E (10654) |
| GF4 | Guinea fowl | Rooster | 6 | 5 | 155(4), 9286* (1) | B1 (155, 9286) |
| GF5 | Guinea fowl | Hen | 6 | 5 | 155 (2), 4392 (1), 86 (1), 942 (1) | B1 (155, 4392, 86, 942) |
| GF6 | Guinea fowl | Hen | 1 | 5 | 540 (1), 2067 (4) | A (540), B1 (2067) |
| GF7 | Guinea fowl | Rooster | 2 | 5 | 212 (4), 155 (1) | B1 (155, 212) |
| GF8 | Guinea fowl | Rooster | 7 | No *E. coli* isolated | | |
| GF9 | Guinea fowl | Rooster | 8 | 5 | 2614 (2), 295 (1) 196 (1), 2067 (1) | B1 (2614, 295, 196) |
| Total | | | | 68 | | |

*Novel sequence types.

## Sample collection

The sampling was done in November 2016. Poultry birds were first observed in motion for the presence of any abnormalities. Healthy-looking birds were procured from eight contiguous households within 0.3–0.4 km of each other and transported to the Abuko Veterinary Station, the Gambia in an air-conditioned vehicle. A qualified veterinarian then euthanized the birds and removed their caeca under aseptic conditions. These were placed into sterile Falcon tubes and flash-frozen on dry ice in a cooler box. The samples were transported to the Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine labs in Fajara, where the caecal contents were aseptically emptied into new Falcon tubes for storage at −80 °C within 3 h. A peanut-sized aliquot was taken from each sample into a 1.8 ml Nunc tube containing 1 ml of skim-milk-tryptone-glucose-glycerol (STGG) transport and storage medium (Oxoid, Basingstoke, UK), vortexed at 4200 r.p.m. for 2 min and frozen at −80 °C. Fig. 1 summarizes the sample processing flow.

## Microbiological processing

The caecal–STGG suspension was removed from −80˚C storage and allowed to thaw briefly on wet ice. A 100 µl aliquot was then taken into 900 µl of physiological saline (0.85%) and taken through four 10-fold serial dilutions. A 100 µl aliquot each was then taken from the dilutions and uniformly streaked onto tryptone–bile–X-glucoronide agar plates using the spread plate technique. The inoculated plates

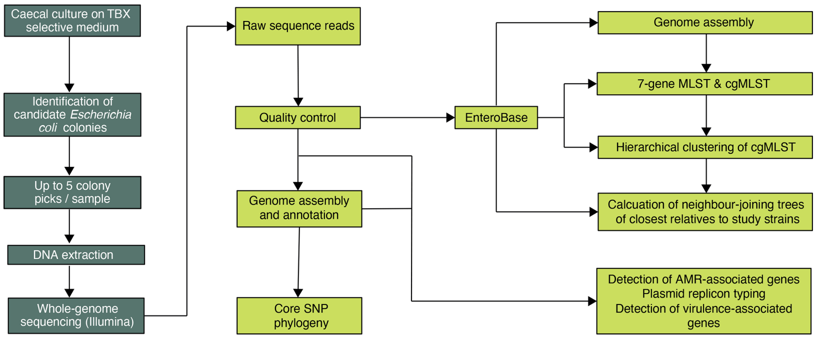

Caecal culture on TBX selective medium
Identification of candidate *Escherichia coli* colonies
Up to 5 colony picks / sample
DNA extraction
Whole-genome sequencing (Illumina)
Raw sequence reads
Quality control
Genome assembly and annotation
Core SNP phylogeny
EnteroBase
Genome assembly
7-gene MLST & cgMLST
Hierarchical clustering of cgMLST
Calcuation of neighbour-joining trees of closest relatives to study strains
Detection of AMR-associated genes
Plasmid replicon typing
Detection of virulence-associated genes

**Fig. 1.** Study sample-processing flow diagram. TBX, tryptone–bile–X-glucoronide agar; MLST, multilocus sequence typing; cgMLST, core genome multilocus sequence typing.

were incubated at 37 °C for 18–24 h under aerobic conditions. Following overnight incubation, colony counts were determined for raised, translucent and entire colonies that exhibited bluish-green pigmentation typical of *E. coli*. Up to five candidate colonies were selected per sample and sub-cultured on MacConkey agar. These were incubated at 37 °C in air for 18–24 h and stored in 20% glycerol broth at −80 °C. The isolates from chickens were designated C1–C10, while those from guinea fowl were prefixed by GF1–GF9, followed by the respective colony number (1 up to 5).

### Genomic DNA extraction

Genomic DNA was extracted from overnight broth cultures prepared from each single colony sub-culture using an in-house 96-well plate lysate method as described previously [32]. The DNA was eluted in Tris/Cl (pH, 8.0) and quantified using the Qubit high-sensitivity DNA assay kit (Invitrogen, MA, USA). DNA samples were kept at −20 °C until the Illumina sequencing library preparation. Broth cultures were spun at 3500 r.p.m. for 2 min and lysed using lysozyme, proteinase K, 10% SDS and RNase A in Tris EDTA buffer (pH 8.0).

### Illumina sequencing

Whole-genome shotgun sequencing of the DNA extracts was performed for all the study isolates on the Illumina NextSeq 500 instrument (Illumina, San Diego, CA, USA) using a modified Illumina Nextera library preparation protocol as described previously [32]. We ran the final pooled library at a concentration of 1.8 pM on a mid-output flow cell (NSQ 500 Mid Output KT v2 300 cycles; Illumina catalogue no. FC-404–2003) according to the manufacturer's instructions. Following sequencing, FASTQ files were downloaded from BaseSpace to a local server hosted at the Quadram Institute Bioscience.

### Genome assembly and phylogenetic analysis

The raw sequences were initially analysed on the Cloud Infrastructure for Microbial Bioinformatics [33]. This included concatenating paired-end short reads, quality checks with FastQC v0.11.7 [34], trimming of low-quality reads (median quality below a Phred score of~30 and read lengths below 36 bp) and Illumina adapters with Trimmomatic v0.39 [35] and assembly by Spades v3.13.2 [36]. The quality of the assemblies was checked using QUAST v5.0.0, de6973bb [37] and annotation of the draft genomes was carried out using Prokka v1.13.3 [38]. We used the mlst software (https://github.com/tseemann/mlst) to call multilocus sequence types (MLSTs) using the Achtman scheme [39], based on the seven house-keeping genes, *adk*, *fum*C, *gyr*B, *icd*, *mdh*, *pur*A and *rec*A. We used Snippy v4.3.2 (https://github.com/tseemann/snippy) for variant calling and to generate a core-genome alignment, from which a maximum-likelihood phylogenetic tree was reconstructed using RAxML v8.2.4 [40], based on a general time-reversible nucleotide substitution model with 1, 000 bootstrap replicates. We included representative reference genome sequences for the major phylogroups of *E. coli* and *Escherichia fergusonii* as an outgroup (File S1). Given that recombination is widespread in *E. coli* and tends to blur phylogenetic signals [39], we used Gubbins (Genealogies Unbiased By recomBinations In Nucleotide Sequences) [41] to detect and mask recombinant regions of the core-genome alignment prior to the phylogenetic reconstruction. We used the GrapeTree [42] to visualize and annotate phylogenetic trees. We calculated pair-wise single-nucleotide polymorphism (SNP) distances between genomes from the core-genome alignment using snp-dists v0.6 (https://github.com/tseemann/snp-dists).

Subsequently, the short-read sequences were uploaded to EnteroBase [43], an online genome database and integrated software environment that currently hosts more than 138 164 *E. coli* genomes, sourced from all publicly available sequence databases and user uploads. EnteroBase routinely retrieves short-read *E. coli* sequences from the public domain, performs quality control and *de novo* assemblies of Illumina short-read sequences, annotates these and assigns seven-allele MLST (ST) and phylogroups from genome assemblies using standardized pipelines. In addition, EnteroBase assigns unique core-genome MLST (cgMLST) numbers to each genome, based on the typing of 2, 512 genes in *E. coli*.

### Population structure analysis

We utilized the hierarchical clustering (HierCC) algorithm in EnteroBase to assign our poultry genomes to eleven stable clusters designated as HC0 up to HC1100, based on pair-wise differences between genomes at cgMLST alleles. In *Salmonella,* the HC100 or HC200 clusters seem to correspond to long-term strain endemicity, while in *E. coli*, HC1100 corresponds to the seven-allele MLST clonal complexes [43]. The HierCC algorithm therefore lends itself as a very useful tool for the analysis of bacterial population structures at multiple levels of resolution. In a recent study of the population structure of *Clostridioides difficile*, Frentrup *et al.* [44] showed that HierCC allows closely related neighbours to be detected at 89% consistency between cgMLST pair-wise allelic differences and SNPs. We determined the closest relatives to our study *E. coli* isolates using the HC1100 cluster and reconstructed neighbour-joining trees using NINJA [45]. In order to compare the strain distribution that we observed among our study isolates with what pertains in poultry *E. coli* isolates from elsewhere, we further retrieved genomic assemblies from all publicly available poultry *E. coli* isolates, stratified by their source continent and reconstructed NINJA neighbour-joining trees depicting the prevalence of STs per continent.

### Analysis of accessory gene content

We used ARIBA v2.12.1 [46] to detect virulence factors, antimicrobial resistance genes and plasmid replicons among our study isolates. Briefly, this tool scans the short-read sequences against the core Virulence Factors Database (VFDB) [47] (virulence factors), ResFinder (AMR) [48] and PlasmidFinder (plasmid-associated genes) [49] databases and generates customized outputs, based on a percentage identity of ≥90% and coverage of ≥70%. The VFDB-core, ResFinder and PlasmidFinder databases were downloaded on 29 October 2018. As a quality check, the results were confirmed by running ABRicate v0.9.8 (https://github.com/tseemann/abricate) (databases updated 12 October 2020) using the assembled contigs. Virulence factors were visualized by overlaying them onto the phylogenetic tree using the ggtree, ggplot2 and phangorn packages in RStudio v3.5.1.

We determined the prevalence of AMR genes among poultry *E. coli* isolates from the rest of the world, for comparison with what we found in isolates from this study. To do this, we interrogated the downloaded continent-stratified genomes as above using ABRicate v0.9.8 (https://github.com/tseemann/abricate) to predict AMR-associated genes by scanning against the ResFinder database (accessed 28 July 2019), based on a percentage identity threshold of ≥90% and a coverage of ≥70%.

### Antimicrobial susceptibility

Due to logistic constraints, a third of the study isolates (20/68, 29%) were randomly selected for phenotypic susceptibility testing by minimum inhibitory concentrations (MICs). MICs were performed by the agar dilution method [50], according to the European Committee on Antimicrobial Susceptibility Testing v9.0 (EUCAST, 2019) guidelines. Stock solutions of 1000 mg $l^{-1}$ were initially prepared, from which the working solutions were made. For each antibiotic, duplicate twofold serial dilutions (from 32 mg $l^{-1}$ to 0.03 mg $l^{-1}$) were done in molten Müller–Hinton agar (Oxoid, Basingstoke, UK). The results were interpreted according to EUCAST breakpoint tables (http://www.eucast.org). Where EUCAST cut-off values were not available, the recommended cut-off values from the Clinical Laboratory Standards Institute (https://www.clsi.org) were used.

### Oxford Nanopore sequencing

Two novel strains recovered from guinea fowl were long-read sequenced on the Oxford Nanopore platform as follows. Prior to sequencing, DNA fragments were assessed using the Agilent 2200 TapeStation (Agilent catalogue no. 5067–5579) to determine the fragment lengths. Long-read sequencing was carried out using the rapid barcoding kit (Oxford Nanopore catalogue no. SQK-RBK004). Libraries were prepared following the manufacturer's instructions. An input DNA concentration of 400 ng was used for the library preparation and a final concentration of 75 µl of the prepared library was loaded onto an R9.4 MinION flow cell. The final concentration of the library pool was assessed using the Qubit high-sensitivity DNA assay (Invitrogen, MA, USA).

### Hybrid assembly and analysis of plasmids and phages

The long reads were base-called with Guppy, the Oxford Nanopore Technologies' post-sequencing processing software (https://nanoporetech.com/). The base-called FASTQ files were then concatenated into a single file each and demultiplexed based on their respective barcodes, using the qcat Python command-line tool v1.1.0 (https://github.com/nanoporetech/qcat). We performed hybrid assemblies of the Illumina and Nanopore reads with Unicycler v0.4.8.0 [51]. The quality of the hybrid assemblies was assessed with QUAST v5.0.0, de6973bb [37]. The hybrid assemblies were then analysed for the presence of plasmids and prophages using ABRicate PlasmidFinder and PHASTER [52] respectively. Annotations of the assemblies were carried out using Prokka v1.13.3 [38].

**Table 2.** Prevalence of the study sequence types in EnteroBase

| ST | Source | Phylotype | Prevalence in EnteroBase |
|---|---|---|---|
| ST48 | Chicken | A | Human, livestock, *Celebes* ape |
| ST58 | Chicken | B1 | Human, livestock, poultry |
| ST86 | Guinea fowl | B1 | Human, livestock, companion animal, poultry |
| ST155 | Chicken, guinea fowl | B1 | Human, poultry, mink, livestock |
| ST196 | Guinea fowl | B1 | Human, livestock, companion animal, environment |
| ST212 | Guinea fowl | B1 | Human, poultry, deer, companion animal |
| ST295 | Guinea fowl | B1 | Human, poultry, livestock, companion animal, environment, food, |
| ST337 | Chicken | B1 | Human, rhinoceros, poultry, environment (soil and water) |
| ST443 | Guinea fowl | B1 | Human, environment, livestock |
| ST540 | Chicken, guinea fowl | A | Human, environment (water and sewage), livestock, poultry, gull, rabbit, plant, oyster, fish |
| ST542 | Chicken | A | Human, livestock, poultry |
| ST602 | Chicken | B1 | Human, poultry, livestock, bird, fish, reptile |
| ST746 | Chicken | A | Human, poultry, fish, livestock, environment (water) |
| ST942 | Guinea fowl | B1 | Environment, food, companion animal, livestock |
| ST1423 | Chicken | B1 | Human, reptile, livestock |
| ST2067 | Guinea fowl | B1 | Human, environment |
| ST2165 | Chicken | B1 | Livestock, companion animal, reptile, bird |
| ST2461 | Chicken | A | Sheep, poultry |
| ST2614 | Guinea fowl | B1 | Human |
| ST2772 | Chicken | B1 | Human, livestock, environment |
| ST4392 | Guinea fowl | B1 | Livestock, wild animal, companion animal |
| ST5826 | Chicken | A | Poultry |
| ST6025 | Guinea fowl | D | Unknown source |
| ST6186 | Chicken | B1 | Livestock, environment |
| ST9284 | Chicken | E | Novel |
| ST9285 | Chicken | B1 | Novel |
| ST9286 | Guinea fowl | B1 | Novel |
| ST10654 | Guinea fowl | D | Novel |

*ST6025 occurred in only one other isolate in EnteroBase, beside the study strain. However, the source of isolation of this other isolate was not available.

## RESULTS

### Study population

We analysed 19 caecal samples obtained from 10 chickens and 9 guinea fowl. Fifteen out of the 19 (79%) samples yielded growth of *E. coli* on culture, from which 68 colonies were recovered (5 colonies from each of 13 birds, 2 from a single bird, and 1 colony from another bird).

### Sequence type and phylogroup distribution

We recovered 28 7-allele sequence types (STs), of which ST155 was the most common (22/68, 32%). Four of the STs were novel – two from chickens and two from guinea fowl. The allelic profiles of the novel strains are provided in File S2. Seventeen of the 28 STs have previously been isolated from humans or other vertebrates, 6 (ST942, ST2165, ST2461, ST4392, ST5826 and 6186) have not been seen in humans

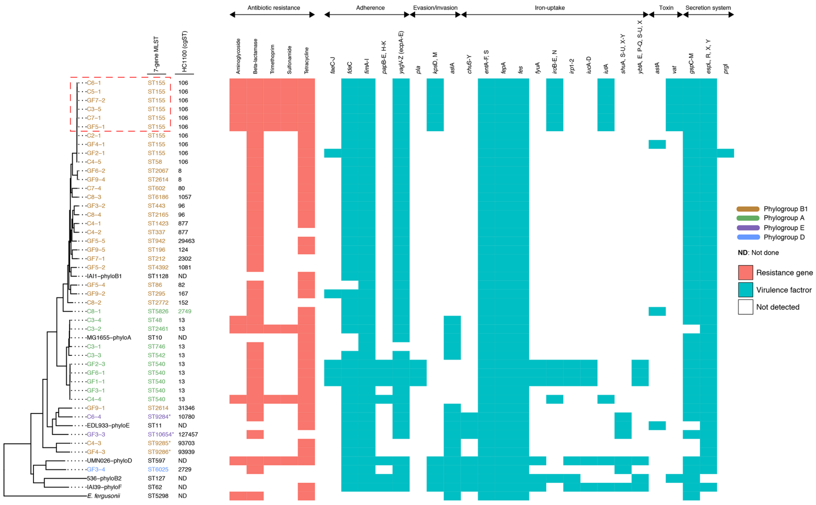


**Fig. 2.** A maximum-likelihood phylogeny of the study isolates reconstructed with RAxML, based on non-repetitive, non-recombinant core SNPs, using a general time-reversible nucleotide substitution model with 1000 bootstrap replicates. The tip labels indicate the sample names, with the respective Achtman sequence types (STs) and HC1100 (cgST complexes) indicated next to the sample names. The colour codes indicate the respective phylogroups to which the isolates belong. The outgroup and the other *E. coli* reference genomes denoting the major *E. coli* phylogroups are in black. Asterisks (*) are used to indicate novel STs. Overlaid on the tree are the predicted antimicrobial resistance genes and virulence factors for each isolate. The virulence genes are grouped according to their function. Chicken isolates are denoted 'C' and guinea fowl samples 'GF', with the suffix indicating the colony pick. We have not shown multiple colonies of the same Achtman ST recovered from a single individual – in such instances, only one representative isolate is shown. Nor have we shown virulence factors that were detected only in the reference genomes. The red box highlights multi-drug-resistant isolates that concurrently harbour putative fitness and colonization factors that are important for invasion of host tissues and evasion of host immune defences. The full names of virulence factors and their known functions are provided in File S6.

before and 1 (ST6025) only occurred in 1 other isolate in EnteroBase, beside the study strain. However, the source of isolation of this other isolate was not available (Table 2). The isolates were spread over phylogroups B1, A, B2 and D, but most belonged to phylogroups B1 and A, which are home to strains associated with human intestinal infections and avian colibacillosis [53, 54] (Fig. 2). Hierarchical clustering resolved the study strains into 22 cgMLST complexes, indicating a high level of genomic diversity (File S2).

We generated complete, circular genome assemblies of the two novel sequence types isolated from guinea fowl: ST10654 (GF3-3) and ST9286 (GF4-3). Although neither strain encoded AMR genes or plasmids, GF3-3 contained three prophages (two intact, one incomplete), while GF4-3 harboured four prophages (three intact, one incomplete) (File S3).

## Within-host genomic diversity and transmission of strains

Several birds (12/19, 63%) were colonized by two or more STs; in most cases, the STs spanned more than two phylotypes (Table 1). In two chickens, all five colony picks belonged to distinct STs. We observed some genetic diversity among multiple colonies of the same ST recovered from the same host (Table 3a). Most of these involved variants that differed by 0–4 SNPs, i.e. variation likely to have arisen due to within-host evolution. However, in one instance, pair-wise SNP differences (ranging from 4 to 255) suggested independent acquisition of distinct clones. Pair-wise SNP analysis also suggested transmission of strains (including MDR isolates) between chickens and between chickens and guinea fowl (Table 3b, c) from the same household (File S4).

## Prevalence of AMR, virulence factors and plasmid replicons among the study isolates

Twenty isolates (20/68, 29%) harboured at least one AMR gene and 16 (16/68, 24%) were MDR, i.e. positive for genes predicted to convey resistance to three or more classes of antibiotics (Fig. 3; File S5). Fourteen of the 16 MDR isolates belonged to ST155 – representing 64% (14/22) of the ST155 isolates recovered in this study. Notable among the resistance genes detected was the class A broad-spectrum

**Table 3a.** Within-host single-nucleotide polymorphism diversity between multiple genomes of the same ST recovered from the same bird

| Sample ID | Sequence type (ST) | Colonies per ST | Pair-wise SNP distances between multiple colonies of the same ST |
|---|---|---|---|
| C5 | ST155 | 2 | 0 |
| C6 | ST155 | 3 | 0 |
| C6 | ST9284 | 2 | 4 |
| C7 | ST155 | 4 | 0 |
| C8 | ST2772 | 2 | 4 |
| GF1 | ST540 | 5 | 0–3 |
| GF2 | ST155 | 4 | 0 |
| GF3 | ST540 | 2 | 2 |
| GF4 | ST155 | 4 | 0–4 |
| GF5 | ST155 | 2 | 0 |
| GF6 | ST2067 | 4 | 0 |
| GF7 | ST212 | 4 | 4–255 |
| GF9 | ST2614 | 2 | 0 |

'C' denotes chickens and 'GF' denotes guinea fowl.

**Table 3b.** Single-nucleotide polymorphism differences between isolates recovered from chicken 3, chicken 5, chicken 6 and guinea fowl 7. All the isolates in this transmission network encoded resistance to ≥3 classes of antimicrobials

| | C3-5 | C5-1 | C5-2 | GF7-2 | C6-1 | C6-2 | C6-3 |
|---|---|---|---|---|---|---|---|
| C3-5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C5-1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C5-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| GF7-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C6-1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C6-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C6-3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

'C'' denotes chickens and 'GF' denotes guinea fowl.

**Table 3c.** Single-nucleotide diversity differences between isolates recovered from guinea fowls 1, 2 and 6

| | GF1-1 | GF1-2 | GF1-3 | GF1-4 | GF1-5 | GF2-3 | GF6-1 |
|---|---|---|---|---|---|---|---|
| GF1-1 | 0 | 2 | 3 | 1 | 1 | 2 | 3 |
| GF1-2 | 2 | 0 | 3 | 1 | 1 | 2 | 3 |
| GF1-3 | 3 | 3 | 0 | 2 | 2 | 3 | 2 |
| GF1-4 | 1 | 1 | 2 | 0 | 0 | 1 | 2 |
| GF1-5 | 1 | 1 | 2 | 0 | 0 | 1 | 2 |
| GF2-3 | 2 | 2 | 3 | 1 | 1 | 0 | 3 |
| GF6-1 | 3 | 3 | 2 | 2 | 2 | 3 | 0 |

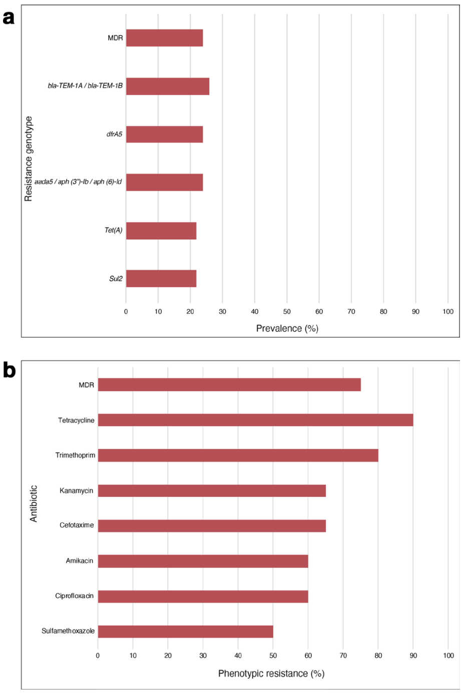


**Fig. 3.** (a) A bar graph showing the prevalence of resistance genes found among the study isolates, using the core Virulence Factors Database (reference 47) (virulence factors), ResFinder (AMR) (reference 48) and PlasmidFinder (plasmid-associated genes) (reference 49) databases, with a cut-off percentage identity of ≥90% and coverage of ≥70%. The full list of the resistance genes that were detected is presented in File S5. (b) A bar graph depicting the prevalence of phenotypic antimicrobial resistance in 20 isolates. The results were interpreted using the recommended breakpoint tables from EUCAST (*http://www.eucast.org*) or the Clinical Laboratory Standards Institute (*https://www.clsi.org*) where EUCAST cut-off values were not available.

beta-lactamase resistance (*bla*$_{\text{TEM-1A/B}}$) (18/68, 26%). Phenotypic resistance was confirmed in >50% of the isolates tested, with an MDR rate of 75% (15/20).

Interestingly, the MDR isolates also harboured more genes encoding putative virulence factors than did less-resistant isolates (Fig. 2). Overall, 125 unique virulence-associated genes were detected from the study isolates (File S6). Notably, the virulence and AMR profiles of co-colonizing STs tended to differ from each other.

One or more plasmid replicons were detected in 69% (47/68) of the study isolates, with 17 plasmid types detected overall (File S7). IncF plasmids were the most common. A single isolate carried the col156 virulence plasmid. The MDR isolates often co-carried large IncF plasmids [IncFIA_1, ~27 kb; IncFIB(AP001918)_1, ~60 kb; IncFIC(FII)_1, ~56 kb]. Scrutiny of annotated assemblies revealed that the resistance genes were often co-located on the same contig as one of the IncF plasmids. In three birds (guinea fowl 2, guinea fowl 5 and guinea fowl 7), co-colonizing strains (belonging to different STs) shared the same plasmid profile. The results of ARIBA ResFinder, PlasmidFinder and VFDB were 100% concordant with those produced by ABRicate for our study isolates.

### Population dynamics of study strains

Hierarchical clustering analyses provided evidence of genomic relationships between strains from poultry and those from humans (Table 4); however, this warrants further investigation using samples collected from poultry and humans living in close proximity from the same setting. Significant among these were ST2772 and ST4392, which were separated from human isolates belonging to these STs by just 41 and 68 alleles in the core-genome MLST scheme, respectively (Figs 4 and 5). Similarly, ST86, ST6186 and ST602 were closest to isolates from livestock (Figs S9–S11), suggesting possible exchange of strains between livestock species.

By contrast, three of the novel STs from this study (ST10654, ST9285, ST9286) were genetically distinct from anything else in the public domain. These belonged to unique HC1100 clusters in the cgMLST scheme and did not have any relatives in the seven-allele MLST scheme, even after allowing for two mismatches. Two of these (ST10654 from Guinea fowl 3 and ST9286 from Guinea fowl 4) now have complete genomic assemblies.

### The global prevalence of strains and AMR among avian *E. coli* isolates

Phylogenomic analyses of 4, 846 poultry *E. coli* isolates from all over the world revealed that ST155 is common among poultry isolates from Africa and South America (Figs S1 and S2). In contrast, ST117 is prevalent among poultry isolates from Europe and North America (Figs S3 and S4), with ST156 and ST254 being the most common *E. coli* STs found in poultry from Asia and Oceania, respectively (Figs S5 and S6).

Our phylogenetic analyses revealed that ST155 strains from Africa were dispersed among other ST155 isolates from the rest of the world; however, the majority of ST155 strains from this study belonged to a tight genomic cluster, comprising isolates from poultry and livestock from sub-Saharan Africa (separated by 38–39 alleles), except for a single isolate sourced from poultry in the USA. In the cgMLST scheme, all the study ST155 isolates fell into four HC100 sub-clusters (100 alleles difference) (Fig. S7). The largest sub-cluster (sub-cluster 1, HC100_43137) comprised ST155 isolates from this study and isolates from Uganda and Kenya; while sub-clusters 2 (HC100_73903), 3 (HC100_73905) and 4 (HC100_93719) occurred in the Gambia only, although distantly related to isolates from humans and a companion animal (Fig. S8).

**Table 4.** Closest relatives to the Gambian poultry strains

| Seven-gene ST | cgST HC100 sub-cluster designation | Study poultry host | Neighbour host | Neighbour's country of isolation | Allelic distance |
|---|---|---|---|---|---|
| ST9286 | NA | Guinea fowl | Chicken | Gambia (this study) | 945 |
| ST9285 | NA | Chicken | Guinea fowl | Gambia (this study) | 945 |
| ST10654 | NA | Guinea fowl | Unknown avian source | Kenya | 1324 |
| ST155 | 43137 | Chicken and guinea fowl | Poultry | USA | 32–34 |
| ST2772 | NA | Chicken | Human | Kenya | 41 |
| ST6186 | NA | Chicken | Livestock | USA | 58 |
| ST540 | 10207 | Guinea fowl | Human | UK | 59 |
| ST58 | 25133 | Chicken | Unknown | Unknown | 59 |
| ST2461 | 93699 | Chicken | Human | Kenya | 64 |
| ST2165 | 12281 | Chicken | Food | Kenya | 66 |
| ST4392 | NA | Guinea fowl | Human | UK | 68 |
| ST602 | NA | Chicken | Livestock | USA | 70 |
| ST540 | 70056 | Chicken | Food | UK | 72 |
| ST540 | 1320 | Guinea fowl | Poultry | USA | 73 |
| ST942 | NA | Guinea fowl | Environment (tap water) | Australia | 76 |
| ST212 | NA | Guinea fowl | Seagull | Australia | 81 |
| ST5826 | NA | Chicken | Water | UK | 91 |
| ST1423 | 27957 | Chicken | Reptile | USA | 96 |
| ST337 | 73054 | Chicken | Reptile | USA | 96 |
| ST196 | NA | Guinea fowl | Human | Kenya | 102 |
| ST155 | 93719 | Chicken | Tanzania | Human | 106 |
| ST86 | NA | Guinea fowl | US | Livestock | 131 |
| ST155 | 73905 | Guinea fowl | Companion animal | USA | 137 |
| ST542 | 93732 | Chicken | Poultry | USA | 148 |
| ST746 | NA | Chicken | Poultry | USA | 148 |
| ST295 | NA | Guinea fowl | Human | Mexico | 162 |
| ST48 | 93724 | Chicken | Unknown | UK | 163 |
| ST542 | 93697 | Chicken | Environment (soil/dust) | USA | 194 |
| ST155 | 73903 | Guinea fowl | Nepal | Human | 195 |
| ST443 | 93721 | Guinea fowl | Unknown | Unknown | 224 |
| ST6025 | NA | Guinea fowl | Unknown | USA | 245 |
| ST2614 | NA | Guinea fowl | Human | PR China | 284 |
| ST9284 | NA | Chicken | Environment (soil/dust) | North America | 293 |
| ST2067 | NA | Guinea fowl | Human | Gambia | 458 |

NA, Not applicable.

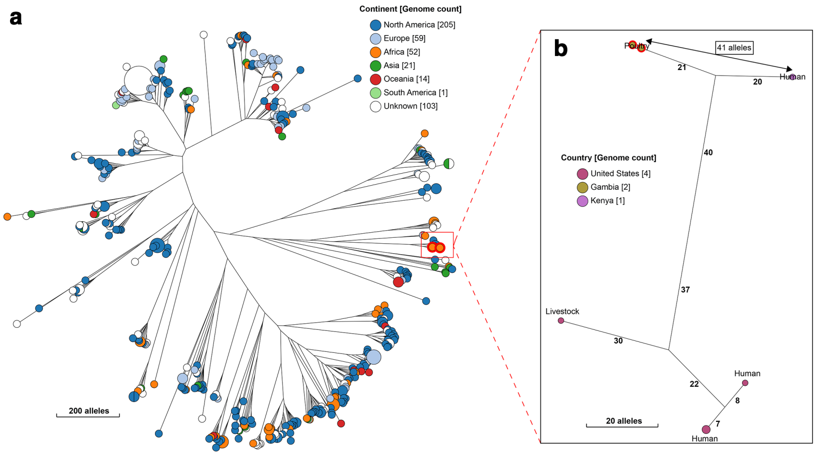


**Fig. 4.** A NINJA neighbour-joining tree showing the phylogenetic relationship between our study ST2772 (Achtman) strain and all other publicly available genomes that fell within the same HC1100 cluster (cgST complex, corresponding to clonal complex in the seven-allele MLST scheme). The locations of the isolates are displayed in the legends, with the genome counts displayed in square brackets. The branch lengths are annotated with the allelic distances separating the genomes. Strains from this study are highlighted in red. The subtree (b) shows the closest relatives to the study strains, with the allelic distance separating them displayed with the arrow (41 alleles).

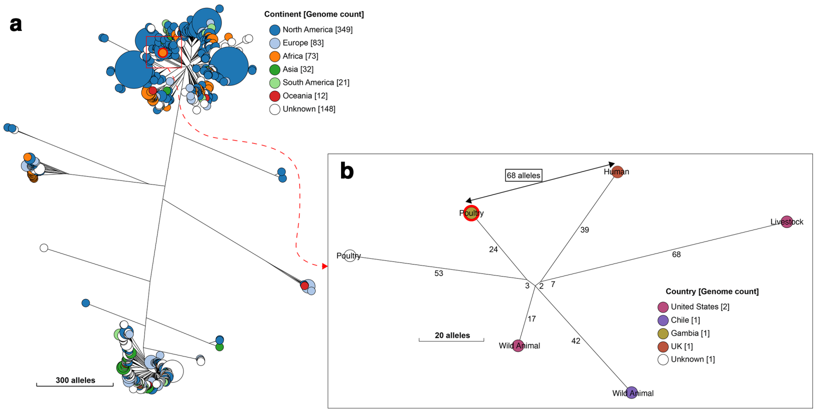


**Fig. 5.** A NINJA neighbour-joining tree showing the phylogenetic relationship between the avian ST4392 (Achtman) strain from this study and all other publicly available genomes that cluster together at HC1100 level (cgST complex, corresponding to clonal complex in the seven-allele MLST scheme). The legend shows the continent of isolation of the isolates, with genome counts displayed in square brackets. Gambian poultry strains are highlighted in red. The study ST strain is separated from a human ST4392 isolate by 68 alleles, as shown in the subtree (b).

**Table 5.** Global prevalence of AMR genes

| | Europe | Africa | South America | North America | Asia | Oceania |
|---|---|---|---|---|---|---|
| Tetracycline | 564/752, 75% | 559/591, 95% | 108/131, 83% | 2480/2975, 83% | 228/249, 92% | 132/148, 90% |
| Aminoglycoside | 303/752, 40% | 378/591, 64% | 94/131, 72% | 1497/2975, 50% | 172/249, 69% | 56/148, 38% |
| Beta-lactamase | 303/752, 40% | 246/591, 42% | 127/131, 98% | 933/2975, 31% | 157/249, 63% | 61/148, 41% |
| Sulphonamide | 338/752, 45% | 377/591, 64% | 84/131, 65% | 1174/2975, 39% | 167/249, 67% | 52/148, 35% |
| Trimethoprim | 192/752, 25% | 353/591, 52% | 58/131, 45% | 176/2975, 6% | 143/249, 57% | 66/148, 45% |
| Chloramphenicol | 303/752, 40% | 69/591, 13% | 36/131, 28% | 69/2975, 2% | 131/249, 53% | 0/148, 0% |
| Quinolone | 51/752, 7% | 144/591, 24% | 24/131, 18% | 17/2975, 1% | 74/249, 30% | 0/148, 0% |
| Lincosamide | 57/752, 8% | 0/591, 0% | 12/131, 9% | 0/2975, 0% | 14/249, 6% | 1/148, 1% |
| Macrolide | 20/752, 3% | 79/591, 13% | 3/131, 2% | 30/2975, 1% | 92/249, 37% | 0/148, 0% |
| Fosfomycin | 8/752, 1% | 4/591, 1% | 31/131, 24% | 19/2975, 1% | 71/249, 29% | 0/148, 0% |
| Streptogrammin | 0/752, 0% | 0/591, 0% | 23/131, 18% | 0/2975, 0% | 0/249, 0% | 0/148, 0% |
| Colistin | 29/752, 4% | 0/591, 0% | 9/131, 7% | 0/2975, 0% | 119/249, 48% | 0/148, 0% |
| MDR | 406/752, 54% | 392/591, 66% | 100/131, 77% | 1236/2975, 42% | 175/249, 70% | 56/148, 44% |

The full list of resistance genes that were detected is presented in File S8.

Antimicrobial resistance was high across the continents, with the highest prevalence of MDR in South America (100/131, 77%), followed by Asia (175/249, 70%) and then Africa (392/591, 66%) (Table 5; File S8). Of note, the highest percentages of resistance globally were those for broad-spectrum beta-lactamases, while the lowest percentages of resistance were to colistin (File S8). Interestingly, the prevalence of colistin resistance was highest in Europe but did not occur in Oceania and North America.

## DISCUSSION

Here, we have described the genomic diversity of *E. coli* from backyard chickens and guinea fowl reared in households in rural Gambia, West Africa. Backyard poultry from this rural setting harbour a remarkably diverse population of *E. coli* strains that encode antimicrobial resistance genes and virulence factors that are important for infections in humans. Furthermore, we provide evidence of sharing of strains (including MDR strains) from poultry to poultry and between poultry, livestock and humans, with potential implications for public health.

Our results reflect the rich diversity that exists within the *E. coli* population from backyard poultry. Although our sample size was small (19 birds), we recovered as many as 28 STs of *E. coli*, 4 of which have not been seen before – even though more than quarter of a million *E. coli* strains had been sequence typed to date (March 2020). Three of our novel STs differed by >945 alleles from their nearest relative. Two of these now have complete assemblies. Also, some of the strains from this study were found in unique cgMLST HierCC clusters containing strains only from this study.

Our results confirm previous reports that phylogroups B1 and A are the dominant phylogroups among *E. coli* isolates from both intensive and backyard poultry [55–58]. Hierarchical clustering analysis suggested that ST155 is common in African poultry. However, most of our ST155 strains belong to a unique cgMLST cluster containing closely related (38–39 alleles differences, and so presumably recently diverged) isolates from poultry and livestock from sub-Saharan Africa, suggesting that strains can be exchanged between livestock and poultry in this setting.

Rural backyard poultry can act as a source of transmission of infections to humans, due to the absence of biosecurity and daily contact with humans [59]. Indirect contact might occur through food or through contact with faeces; for example, by children who are often left to play on the ground [60].

We observed a high prevalence of AMR genes among *E. coli* isolates sourced from African poultry. Similarly, high rates of genotypic MDR were detected among poultry *E. coli* isolates from the rest of the world, with ESBL (various types) being the most significant resistance gene detected. Poultry-associated ESBL genes have also been found among human clinical isolates [61]. Strikingly, most of our ST155 isolates encoded resistance to ≥3 classes of clinically relevant antibiotics, with the highest percentages seen for *bla*$_{\text{TEM-1}}$ beta-lactamase and tetracycline (*tetA*). This is worrying, as beta-lactamase-positive isolates are often resistant to several other classes of antibiotics [62, 63].

Our results are consistent with previous studies that reported ST155 isolates to be commonly associated with MDR [64, 65], but differ from other studies that have reported a low prevalence of AMR in backyard poultry. For example, in a study that compared the prevalence of ESBL genes in backyard poultry and commercial flocks from West Bengal, India, none of the 272 *E. coli* isolates from backyard birds harboured any ESBL gene [66], while 30% of commercial birds carried ESBL genes. The absence of resistance in that study was attributed to a lack of exposure to antimicrobials. Similarly, *E. coli* from organic poultry in Finland were reported to be highly susceptible to most of the antimicrobials studied and no ESBL resistance was detected [67].

Although tetracycline is commonly used in poultry farming for therapeutic purposes [68], resistance to this antibiotic is known to be prevalent in poultry, even in the absence of the administration of this antibiotic [69]. Our results also suggest that IncF plasmids may play a role in the dissemination of AMR in our study population. Conjugation assays are needed to confirm the association of these plasmids with the observed resistance genes and the mobilisability of the plasmids and thus, the potential for exchange among co-colonizing strains in a single host; however, these could not be performed due to coronavirus disease 2019 (Covid-19) restrictions.

Many sub-Saharan countries lack clear guidelines on the administration of antibiotics in agriculture, although an increasing trend in the veterinary use of antimicrobials has been documented [70]. The use of antimicrobials in developing countries is likely to increase because of increasingly intensive farming practices [71]. Europe has banned the use of antimicrobials as growth promoters since 2006 [72] and the use of all essential antimicrobials for prophylaxis in animal production since 2011 [73]. However, AMR may be less well controlled in other parts of the world.

Although APEC strains span several phylogroups (A, B1, B2 and D) and serogroups [54], the majority of APEC strains encode virulence genes associated with intestinal or extra-intestinal disease in humans. These include adhesion factors, toxins, iron acquisition genes and genes associated with serum resistance, such as *fyuA*, *iucD*, *iroN*, *iss*, *irp2*, *hlyF*, *vat*, *kpsM* and *ompT*. Although APEC isolates present different combinations of virulence factors, each retains the capability to cause colibacillosis [13, 74]. We did not detect haemolysin or serum survival genes in our study isolates; however, we recovered some of the known markers of intestinal and extraintestinal virulence in some study isolates, such as the enteroaggregative *E. coli* heat-stable enterotoxin and the vacuolating autotransporter toxin (*vat*, *astA*), invasion and evasion factors (*kpsM*, *kpsD*, *pla*) and adherence factors (*fim* and *pap* genes) that are associated with intestinal and extraintestinal infections in humans. Thus, these strains could cause disease in humans, should they gain access to the appropriate tissues.

Several birds were colonized with two or more STs and at least two phylotypes of *E. coli*. This level of diversity is probably a consequence of the frequent exposure of backyard poultry to the environment, livestock and humans. Co-colonization of single hosts with multiple strains may facilitate the spread of AMR- and virulence-associated genes from resistant strains to other bacteria via both horizontal and vertical gene transfer [75]. A high co-colonization rate of *E. coli* has been described in humans [76, 77] and in non-human primates [32], involving pathogenic strains of *E. coli*. Recently, Li *et al*. reported three to nine sequence types of colistin-resistant *E. coli* to co-exist within a single broiler chicken [30]. Here, we report co-colonization with different lineages of *E. coli* in backyard chickens and guinea fowl. Unsurprisingly, co-colonizing strains often had different AMR and virulence patterns.

An obvious limitation of our study is the small sample size. This study could have also been enhanced by sampling *E. coli* from humans within close proximity to our backyard birds, but we could not perform an analysis of *E. coli* from sympatric humans from our study setting due to logistic reasons and funding limitations of our study. Nonetheless, the inclusion of publicly available sequences strengthens our analysis and inference of the population of *E. coli* in this setting. We also could not perform phenotypic susceptibility testing on all isolates. We acknowledge that a minor percentage of genotypic resistance predictions fail to correspond with phenotypic resistance [78].

Taken together, our results indicate a rich diversity of *E. coli* within backyard poultry from the Gambia, characterized by strains with a high prevalence of AMR and the potential to contribute to infections in humans. This, coupled with the potential for the exchange of strains between poultry and livestock within this setting, might have important implications for human health and warrants continued surveillance.

Funding information
M. J. P., E. F. N., N. M. T., A. R. and J. O. were supported by the BBSRC Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent projects 44 414 000A and 4 408 000A. N. F. A. was supported by the Quadram Institute Bioscience BBSRC funded Core Capability Grant (project number BB/ CCG1860/1). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements
We thank Dr Duto Fofana and Dr Ousman Ceesay of the Department of Livestock Services, Veterinary Services, Ministry of Agriculture, the Gambia for kindly providing access to their veterinary station at Abuko. We gratefully acknowledge the Research Support Office at the Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine, in particular Ms Isatou Cham, for project management support. We also thank Mr Steven Rudder and Mr David Baker for their assistance with the Illumina and Oxford Nanopore sequencing.

Author contributions
Conceptualization, M. A., M. J. P.; data curation, M. J. P., N. F. A.; formal analysis, E. F. N.; funding, M. J. P., M. A.; sample collection, E. F. N., S. J., A. S.; laboratory experiments, E. F. N.; supervision, B. A. K. A., A. R., N. F. A., J. O., M. J. P., M. A.; manuscript preparation – original draft, E. F. N.; review and editing, N. M. T., N. F. A., M. J. P.; review of final manuscript, all authors.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Ethical statement
The study was approved by the Medical Research Council Unit The Gambia at London School of Hygiene and Tropical Medicine's Scientific Co-ordinating Committee (SCC 1615).

## References

1. Bennett CE, Thomas R, Williams M, Zalasiewicz J, Edgeworth M *et al*. The broiler chicken as a signal of a human reconfigured biosphere. *R Soc Open Sci* 2018;5:180325.

2. Storey AA, Athens JS, Bryant D, Carson M, Emery K *et al*. Investigating the global dispersal of chickens in prehistory using ancient mitochondrial DNA signatures. *PLoS One* 2012;7:e39171.
3. Miao YW, Peng MS, Wu GS, Ouyang YN, Yang ZY *et al*. Chicken domestication: an updated perspective based on mitochondrial genomes. *Heredity* 2013;110:277–282.
4. Alders RG, Pym RAE. Village poultry: still important to millions, eight thousand years after domestication. *Worlds Poult Sci J* 2009;65:181–190.
5. Alders RG, Dumas SE, Rukambile E, Magoke G, Maulaga W *et al*. Family poultry: multiple roles, systems, challenges, and options for sustainable contributions to household nutrition security through a planetary health lens. *Matern Child Nutr* 2018;14:e12668.
6. Food and Agriculture Organization of the United Nations (FAO). *Recommendations on the Prevention, Control and Eradication of Highly Pathogenic Avian Influenza in Asia*. Rome, Italy: FAO Position Paper September; 2004.
7. Alders R, Costa R, Gallardo RA, Sparks N, Zhou H. Smallholder poultry: contributions to food and nutrition security, in encyclopedia of food security and sustainability. *Elsevier: Oxford* 2019:292–298.
8. Branckaert RDS, Guèye EF. FAO's programme for support to family poultry production. In: Dolberg F, Petersen PH (editors). *Poultry as a Tool in Poverty Eradication and Promotion of Gender Equality*. Tune Landboskole, Denmark: Proceedings workshop; 1999. pp. 244–256.
9. Olaniyan OF, Camara S. Rural household chicken management and challenges in the upper river region of the Gambia. *Trop Anim Health Prod* 2018;50:1921–1928.
10. Nataro JP, Kaper JB. Diarrheagenic *Escherichia coli*. *Clin Microbiol Rev* 1998;11:142–201.
11. Picard B, Garcia JS, Gouriou S, Duriez P, Brahimi N *et al*. The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. *Infect Immun* 1999;67:546–553.
12. Escobar-Páramo P, Le Menac'h A, Le Gall T, Amorin C, Gouriou S *et al*. Identification of forces shaping the commensal *Escherichia coli* genetic structure by comparing animal and human isolates. *Environ Microbiol* 2006;8:1975–1984.
13. Rodriguez-Siek KE, Giddings CW, Doetkott C, Johnson TJ, Nolan LK. Characterizing the APEC pathotype. *Vet Res* 2005;36:241–256.
14. Barnes HJ NL, Vaillancourt J. Colibacillosis. In: Saif YM, Fadly AM, Glisson JR, Mcdougald LR, Nolan LK *et al*. (editors). *Diseases of Poultry*, 12th. Iowa: Iowa State University Press; 2008. pp. 691–738.
15. Hedman HD, Eisenberg JNS, Trueba G, Rivera DLV, Herrera RAZ *et al*. Impacts of small-scale chicken farming activity on antimicrobial-resistant *Escherichia coli* carriage in backyard chickens and children in rural Ecuador. *One Health* 2019;8:100112.
16. Sarba EJ, Kelbesa KA, Bayu MD, Gebremedhin EZ, Borena BM *et al*. Identification and antimicrobial susceptibility profile of *Escherichia coli* isolated from backyard chicken in and around ambo, central Ethiopia. *BMC Vet Res* 2019;15:85.
17. Langata LM, Maingi JM, Musonye HA, Kiiru J, Nyamache AK. Antimicrobial resistance genes in *Salmonella* and *Escherichia coli* isolates from chicken droppings in Nairobi, Kenya. *BMC Res Notes* 2019;12:22.
18. Borzi MM, Cardozo MV, Oliveira ES, Pollo AS, Guastalli EAL *et al*. Characterization of avian pathogenic *Escherichia coli* isolated from free-range helmeted guineafowl. *Braz J Microbiol* 2018;49:107–112.
19. Vounba P, Kane Y, Ndiaye C, Arsenault J, Fairbrother JM *et al*. Molecular characterization of *Escherichia coli* isolated from chickens with colibacillosis in Senegal. *Foodborne Pathog Dis* 2018;15:517–525.
20. Adzitey F, Assoah-Peprah P, Teye GA. Whole-genome sequencing of *Escherichia coli* isolated from contaminated meat samples collected from the Northern Region of Ghana reveals the presence of multiple antimicrobial resistance genes. *J Glob Antimicrob Resist* 2019;18:179–182.
21. Falgenhauer L, Imirzalioglu C, Oppong K, Akenten CW, Hogan B *et al*. Detection and characterization of ESBL-producing *Escherichia coli* from humans and poultry in Ghana. *Front Microbiol* 2018;9:3358.
22. Chah KF, Ugwu IC, Okpala A, Adamu KY, Alonso CA *et al*. Detection and molecular characterisation of extended-spectrum *β*-lactamase-producing enteric bacteria from pigs and chickens in Nsukka, Nigeria. *J Glob Antimicrob Resist* 2018;15:36–40.
23. Bergeron CR, Prussing C, Boerlin P, Daignault D, Dutil L *et al*. Chicken as reservoir for extraintestinal Pathogenic *Escherichia coli* in humans, Canada. *Emerg Infect Dis* 2012;18:415–421.
24. Manges AR, Harel J, Masson L, Edens TJ, Portt A *et al*. Multilocus sequence typing and virulence gene profiles associated with *Escherichia coli* from human and animal sources. *Foodborne Pathog Dis* 2015;12:302–310.
25. Tijet N, Faccone D, Rapoport M, Seah C, Pasterán F *et al*. Molecular characteristics of mcr-1-carrying plasmids and new mcr-1 variant recovered from polyclonal clinical *Escherichia coli* from Argentina and Canada. *PLoS One* 2017;12:e0180347.
26. El Garch F, Sauget M, Hocquet D, LeChaudee D, Woehrle F *et al*. mcr-1 is borne by highly diverse *Escherichia coli* isolates since 2004 in food-producing animals in Europe. *Clin Microbiol Infect* 2017;23:51.e1–5151.
27. Lazarus B, Paterson DL, Mollinger JL, Rogers BA. Do human extraintestinal *Escherichia coli* infections resistant to expanded-spectrum cephalosporins originate from food-producing animals? A systematic review. *Clin Infect Dis* 2015;60:439–452.
28. Huijbers PMC, Graat EAM, Haenen APJ, van Santen MG, van Essen-Zandbergen A *et al*. Extended-spectrum and AmpC *β*-lactamase-producing *Escherichia coli* in broilers and people living and/or working on broiler farms: prevalence, risk factors and molecular characteristics. *J Antimicrob Chemother* 2014;69:2669–2675.
29. Manges AR. *Escherichia coli* and urinary tract infections: the role of poultry-meat. *Clin Microbiol Infect* 2016;22:122–129.
30. X-P L, Sun R-Y, Song J-Q, Fang L-X, Zhang R-M *et al*. Within-Host heterogeneity and flexibility of mcr-1 transmission in chicken gut. *Int J Antimicrob Agents* 2020;55:105806.
31. Roca A, Hill PC, Townend J, Egere U, Antonio M *et al*. Effects of community-wide vaccination with PCV-7 on pneumococcal nasopharyngeal carriage in the Gambia: a cluster-randomized trial. *PLoS Med* 2011;8:e1001107.
32. Foster-Nyarko E, Alikhan NF, Ravi A, Thilliez G, Thomson NM *et al*. Genomic diversity of *Escherichia coli* isolates from non-human primates in the Gambia. *Microbial Genomics* 2020;6.
33. Connor TR, Loman NJ, Thompson S, Smith A, Southgate J *et al*. CLIMB (the cloud infrastructure for microbial bioinformatics): an online resource for the medical microbiology community. *Microb Genom* 2016;2:e000086.
34. Wingett SW, Andrews S. FastQ screen: a tool for multi-genome mapping and quality control. *F1000Res* 2018;7:1338
35. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
36. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
37. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–1075.
38. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
39. Wirth T, Falush D, Lan R, Colles F, Mensa P *et al*. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* 2006;60:1136–1151.
40. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.

41. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43:e15.

42. Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C *et al*. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res* 2018;28:1395–1404.

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

44. Frentrup M, Zhou Z, Steglich M, Meier-Kolthoff JP, Göker M *et al*. Global genomic population structure of *Clostridioides difficile*. *bioRxiv* 2019;727230.

45. Wheeler TJ. Large-scale neighbor-joining with NINJA. *Algorithms in Bioinformatics*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009.

46. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J *et al*. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genom* 2017;3:e000131.

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

48. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S *et al*. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* 2012;67:2640–2644.

49. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O *et al*. *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 2014;58:3895–3903.

50. Wiegand I, Hilpert K, Hancock REW. Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat Protoc* 2008;3:163–175.

51. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;13:e1005595.

52. Arndt D, Grant JR, Marcu A, Sajed T, Pon A *et al*. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;44:W16–W21.

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

54. Mellata M. Human and avian extraintestinal pathogenic *Escherichia coli*: infections, zoonotic risks, and antibiotic resistance trends. *Foodborne Pathog Dis* 2013;10:916–932.

55. Coura FM, Diniz Sde A, Silva MX, Mussi JM, Barbosa SM *et al*. Phylogenetic group determination of *Escherichia coli* isolated from animals samples. *The Scientific World Journal* 2015;2015:258424.

56. Asadi A, Zahraei Salehi T, Jamshidian M, Ghanbarpour R. ECOR phylotyping and determination of virulence genes in *Escherichia coli* isolates from pathological conditions of broiler chickens in poultry slaughter-houses of southeast of Iran. *Veterinary Research Forum* 2018;9:211–216.

57. Messaili C, Messai Y, Bakour R. Virulence gene profiles, antimicrobial resistance and phylogenetic groups of fecal *Escherichia coli* strains isolated from broiler chickens in Algeria. *Veterinaria Italiana* 2019;55:35–46.

58. Vounba P, Kane Y, Ndiaye C, Arsenault J, Fairbrother JM *et al*. Molecular characterization of *Escherichia coli* Isolated from chickens with colibacillosis in Senegal. *Foodborne Pathog Dis* 2018;15:517–525.

59. Patyk KA, Helm J, Martin MK, Forde-Folle KN, Olea-Popelka FJ *et al*. An epidemiologic simulation model of the spread and control of highly pathogenic avian influenza (H5N1) among commercial and backyard poultry flocks in South Carolina, United States. *Prev Vet Med* 2013;110:510–524.

60. Dione MM, Ikumapayi UN, Saha D, Mohammed NI, Geerts S *et al*. Clonal differences between Non-Typhoidal *Salmonella* (NTS) recovered from children and animals living in close contact in the Gambia. *PLoS Negl Trop Dis* 2011;5:e1148.

61. Castellanos LR, Donado-Godoy P, León M, Clavijo V, Arevalo A *et al*. High heterogeneity of *Escherichia coli* sequence types harbouring ESBL/AmpC genes on IncI1 plasmids in the Colombian poultry chain. *PLoS One* 2017;12:e0170777

62. Mathers AJ, Peirano G, Pitout JD. The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant *Enterobacteriaceae*. *Clin Microbiol Rev* 2015;28:565–591.

63. Reich F, Atanassova V, Klein G. Extended-Spectrum $\beta$-lactamase- and AmpC-Producing enterobacteria in healthy broiler chickens, Germany. *Emerg Infect Dis* 2013;19:1253–1259.

64. Dominguez JE, Faccone D, Tijet N, Gomez S, Corso A *et al*. Characterization of *Escherichia coli* carrying mcr-1-plasmids recovered from food animals arom Argentina. *Frontiers in Cellular and Infection Microbiology*, 9. Brief Research Report; 2019.

65. van Hoek AHAM, Veenman C, Florijn A, Huijbers PMC, Graat EAM *et al*. Longitudinal study of ESBL *Escherichia coli* carriage on an organic broiler farm. *J Antimicrob Chemother* 2018;68:3298–3304.

66. Samanta I, Joardar SN, Das PK, Sar TK. Comparative possession of Shiga toxin, intimin, enterohaemolysin and major extended spectrum beta lactamase (ESBL) genes in *Escherichia coli* isolated from backyard and farmed poultry. *Iran J Vet Res* 2015;16:90–93.

67. Pohjola L, Nykäsenoja S, Kivistö R, Soveri T, Huovilainen A *et al*. Zoonotic public health hazards in backyard chickens. *Zoonoses Public Health* 2016;63:420–430.

68. Fairchild AS, Smith JL, Idris U, Lu J, Sanchez S *et al*. Effects of orally administered tetracycline on the intestinal community structure of chickens and on tet determinant carriage by commensal bacteria and *Campylobacter jejuni*. *Appl Environ Microbiol* 2005;71:5865–5872.

69. van den Bogaard A, Stobberingh EE. Epidemiology of resistance to antibiotics links between animals and humans. *Int J Antimicrob Agents* 2000;14:327–335.

70. Alonso CA, Zarazaga M, Ben Sallem R, Jouini A, Ben Slama K *et al*. Antibiotic resistance in *Escherichia coli* in husbandry animals: the African perspective. *Lett Appl Microbiol* 2017;64:318–334.

71. Van Boeckel TP, Brower C, Gilbert M, Grenfell BT, Levin SA *et al*. Global trends in antimicrobial use in food animals. *Proc Natl Acad Sci U S A* 2015;112:5649–5654.

72. Castanon JIR. History of the use of antibiotic as growth promoters in European poultry feeds. *Poult Sci* 2007;86:2466–2471.

73. Maron DF, Smith TJS, Nachman KE. Restrictions on antimicrobial use in food animal production: an international regulatory and economic survey. *Global Health* 2013;9:48.

74. Schouler C, Schaeffer B, Bree A, Mora A, Dahbi G *et al*. Diagnostic strategy for identifying avian pathogenic *Escherichia coli* based on four patterns of virulence genes. *J Clin Microbiol* 2012;50:1673–1678.

75. Laxminarayan R, Duse A, Wattal C, Zaidi AKM, Wertheim HFL *et al*. Antibiotic resistance—the need for global solutions. *Lancet Infect Dis* 2013;13:1057–1098.

76. Lidin-Janson G, Kaijser B, Lincoln K, Olling S, Wedel H. The homogeneity of the faecal coliform flora of normal school-girls, characterized by serological and biochemical properties. *Med Microbiol Immunol* 1978;164:247–253.

77. Schlager TA, Hendley JO, Bell AL, Whittam TS. Clonal diversity of *Escherichia coli* colonizing stools and urinary tracts of young girls. *Infect Immun* 2002;70:1225–1229.

78. Doyle RM, O'Sullivan DM, Aller SD, Bruchmann S, Clark T *et al*. Discordant bioinformatic predictions of antimicrobial resistance from whole-genome sequencing data of bacterial isolates: an inter-laboratory study. *bioRxiv* 2019;793885.

# APPENDIX 3: Genomic diversity of Escherichia coli from healthy children in rural Gambia

# Genomic diversity of *Escherichia coli* from healthy children in rural Gambia

Ebenezer Foster-Nyarko[1,2], Nabil-Fareed Alikhan[1], Usman N. Ikumapayi[2], Golam Sarwar[2], Catherine Okoi[2], Peggy-Estelle Maguiagueu Tientcheu[2], Marianne Defernez[1], Justin O'Grady[1], Martin Antonio[2,3] and Mark J. Pallen[1,4]

[1] Quadram Institute Bioscience, Norwich Research Park, Norfolk, United Kingdom
[2] Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine, Fajara, The Gambia
[3] Microbiology and Infection Unit, Warwick Medical School, University of Warwick, Coventry, United Kingdom
[4] School of Veterinary Medicine, University of Surrey, Surrey, United Kingdom

## ABSTRACT

Little is known about the genomic diversity of *Escherichia coli* in healthy children from sub-Saharan Africa, even though this is pertinent to understanding bacterial evolution and ecology and their role in infection. We isolated and whole-genome sequenced up to five colonies of faecal *E. coli* from 66 asymptomatic children aged three-to-five years in rural Gambia (n = 88 isolates from 21 positive stools). We identified 56 genotypes, with an average of 2.7 genotypes per host. These were spread over 37 seven-allele sequence types and the *E. coli* phylogroups A, B1, B2, C, D, E, F and *Escherichia* cryptic clade I. Immigration events accounted for three-quarters of the diversity within our study population, while one-quarter of variants appeared to have arisen from within-host evolution. Several isolates encode putative virulence factors commonly found in Enteropathogenic and Enteroaggregative *E. coli,* and 53% of the isolates encode resistance to three or more classes of antimicrobials. Thus, resident *E. coli* in these children may constitute reservoirs of virulence- and resistance-associated genes. Moreover, several study strains were closely related to isolates that caused disease in humans or originated from livestock. Our results suggest that within-host evolution plays a minor role in the generation of diversity compared to independent immigration and the establishment of strains among our study population. Also, this study adds significantly to the number of commensal *E. coli* genomes, a group that has been traditionally underrepresented in the sequencing of this species.

**Subjects** Ecology, Genomics, Microbiology, Molecular Biology
**Keywords** *Escherichia coli*, Genomic diversity, Within-host evolution

Submitted 8 September 2020
Accepted 23 November 2020
Published 6 January 2021

Corresponding author
Mark J. Pallen,
m.pallen@warwick.ac.uk,
mark.pallen@quadram.ac.uk

Academic editor
Valeria Souza

Additional Information and Declarations can be found on page 16

DOI **10.7717/peerj.10572**

## INTRODUCTION

Ease of culture and genetic tractability account for the unparalleled status of *Escherichia coli* as "the biological rock star", driving advances in biotechnology (Blount, 2015), while also providing critical insights into biology and evolution (Good et al., 2017). However, *E. coli* is also a widespread commensal, as well as a versatile pathogen, linked to diarrhoea (particularly in the under-fives), urinary tract infection, neonatal sepsis, bacteraemia and multi-drug resistant infection in hospitals (Camins et al., 2011; Rodríguez-Baño et

**How to cite this article** Foster-Nyarko E, Alikhan N-F, Ikumapayi UN, Sarwar G, Okoi C, Tientcheu P-EM, Defernez M, O'Grady J, Antonio M, Pallen MJ. 2021. Genomic diversity of *Escherichia coli* from healthy children in rural Gambia. *PeerJ* **9**:e10572 http://doi.org/10.7717/peerj.10572

al., 2010; Russo & Johnson, 2003). Yet, most of what we know about *E. coli* stems from the investigation of laboratory strains, which fail to capture the ecology and evolution of this key organism "in the wild" (Hobman, Penn & Pallen, 2007). What is more, most studies of non-lab strains have focused on pathogenic strains or have been hampered by low-resolution PCR methods, so we have relatively few genomic sequences from commensal isolates, particularly from low- to middle-income countries (Ahmed et al., 2014; Ferjani et al., 2017; Moremi et al., 2017; Oshima et al., 2008; Rasko et al., 2008; Stoesser et al., 2015; Touchon et al., 2009).

We have a broad understanding of the population structure of *E. coli,* with eight significant phylogroups loosely linked to ecological niche and pathogenic potential (B2, D and F linked to extraintestinal infection; A and B1 linked to severe intestinal infections such as haemolytic-uraemic syndrome) (Alm, Walk & Gordon, 2011; Escobar-Paramo et al., 2004a; Escobar-Páramo et al., 2004b; Mellata, 2013; Walk et al., 2009). All phylogroups can colonise the human gut, but it remains unclear how far commensals and pathogenic strains compete or collaborate—or engage in horizontal gene transfer—within this important niche (Laxminarayan et al., 2013; Stoppe et al., 2017).

Although clinical microbiology typically relies on single-colony picks (which has the potential to underestimate species diversity and transmission events), within-host diversity of *E. coli* in the gut is crucial to our understanding of inter-strain competition and co-operation and also for accurate diagnosis and epidemiological analyses. Pioneering efforts using serotyping, molecular typing and whole-genome sequencing have shown that normal individuals typically harbour more than one strain of *E. coli*, with one individual carrying 24 distinct clones (Chen et al., 2013; Schlager et al., 2002; Shooter et al., 1977; Dixit et al., 2018; Richter et al., 2018; Bettelheim, Faiers & Shooter, 1972; Sears, Brownlee & Uchiyama, 1950; Sears & Brownlee, 1952). More recently, whole-genome sequencing has illuminated molecular epidemiological investigations (Stoesser et al., 2015), for example, studies of the transmission of extended-spectrum beta-lactamase-encoding *E. coli*, multidrug-resistant *Acinetobacter baumannii*, and the genomic surveillance of multidrug-resistant *E. coli* carriage. Whole-genome data has also been applied to studies of *E. coli* adaptation during and after infection (McNally et al., 2013; Nielsen et al., 2016), as well as the intra-clonal diversity in healthy hosts (Stegger et al., 2020).

There are two plausible sources of within-host genomic diversity. Although a predominant strain usually colonises the host for extended periods (Hartl & Dykhuizen, 1984), successful immigration events mean that incoming strains can replace the dominant strain or co-exist alongside it as minority populations (Bettelheim, Faiers & Shooter, 1972). Strains originating from serial immigration events are likely to differ by hundreds or thousands of single-nucleotide polymorphisms (SNPs). Alternatively, within-host evolution can generate clouds of intra-clonal diversity, where genotypes differ by just a handful of SNPs (Dixit et al., 2018).

Most relevant studies have been limited to Western countries, except for a recent report from Tanzania (Richter et al., 2018), so, little is known about the genomic diversity of *E. coli* in sub-Saharan Africa. The Global Enteric Multicenter Study (GEMS) (Kotloff et al., 2013; Liu et al., 2016) has documented a high burden of diarrhoea attributable to *E. coli*

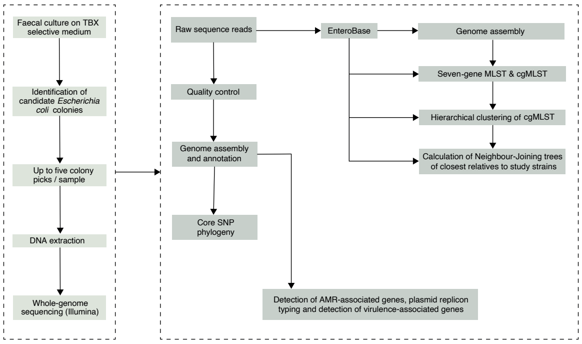

Faecal culture on TBX selective medium
Identification of candidate *Escherichia coli* colonies
Up to five colony picks / sample
DNA extraction
Whole-genome sequencing (Illumina)
Raw sequence reads
Quality control
Genome assembly and annotation
Core SNP phylogeny
EnteroBase
Genome assembly
Seven-gene MLST & cgMLST
Hierarchical clustering of cgMLST
Calculation of Neighbour-Joining trees of closest relatives to study strains
Detection of AMR-associated genes, plasmid replicon typing and detection of virulence-associated genes

**Figure 1** **The study sample-processing flow diagram.**

Full-size DOI: 10.7717/peerj.10572/fig-1

(including *Shigell*a) among children from the Gambia, probably as a result of increased exposure to this organism through poor hygiene and frequent contact with animals and the environment. GEMS was a prospective case-control study which investigated the aetiology of moderate-to-severe diarrhoea in children aged less than five years residing in sub-Saharan Africa and South Asia. In the Gambia, children with moderate-to-severe diarrhoea seeking care at the Basse Health centre in the Upper River Division of the country were recruited, with one to three matched control children randomly selected from the community along with each case. In also facilitating access to stool samples from healthy Gambian children, the GEMS study has given us a unique opportunity to study within-host genomic diversity of commensal *E. coli* in this setting.

# METHODS

## Study population

We initially selected 76 faecal samples from three- to five-four-old (36–59 months) asymptomatic Gambian children, who had been recruited into the GEMS study (*Kotloff et al., 2013*) as healthy controls from December 1, 2007, to March 3, 2011. Samples had been collected according to a previously described sampling protocol (*Kotloff et al., 2012*) and the results of the original study are publicly available at ClinEpiDB.org. Ten of the original 76 samples were depleted and were therefore unavailable for processing in this study. Of the remaining 66 stools, 62 had previously tested positive for *E. coli*. GEMS isolated three *E. coli* colonies per stool sample but pooled these into a single tube for frozen storage. Thus, we needed to re-culture the stools with multiple colony picks, as the original isolate collection was unsuitable for the investigation of within-host diversity. Archived stool samples were retrieved from −80 °C storage and allowed to thaw on ice. A 100–200 mg aliquot from each sample was transferred aseptically into 1.8 ml Nunc tubes for microbiological processing below (Fig. 1).

## Bacterial growth and isolation

1 ml of physiological saline (0.85%) was added to each sample tube and vigorously vortexed at 4,200 rpm for at least 2 min. Next, the homogenised sample suspensions were taken through four ten-fold dilution series. A100 µl aliquot from each dilution was then spread evenly on a plate of tryptone-bile-X-glucuronide differential and selective agar. The inoculated plates were incubated overnight at 37 °C under aerobic conditions. Colony counts were performed on the overnight cultures for each serial dilution for translucent colonies with entire margins and blue–green pigmentation indicative of *E. coli*. Up to five representative colonies were selected from each sample and sub-cultured on MacConkey agar overnight at 37 °C before storing in 20% glycerol broth at −80 °C. Individual isolates were assigned a designation comprised of the subject ID followed by the colony number ("1–5").

## Genomic DNA extraction and genome sequencing

Broth cultures were prepared from pure, fresh cultures of each colony-pick in 1 ml Luria-Bertani broth and incubated overnight to attain between $10^9$–$10^{10}$ cfu per ml. Genomic DNA was then extracted from the overnight broth cultures using the lysate method described in *Foster-Nyarko et al. (2020)*. The eluted DNA was quantified by the Qubit high sensitivity DNA assay kit (Invitrogen, MA, USA) and sequenced on the Illumina NextSeq 500 instrument (Illumina, San Diego, CA), using a modified Nextera XT DNA protocol for the library preparation as described previously (*Foster-Nyarko et al., 2020*). The pooled library was loaded on a mid-output flow cell (NSQ 500 Mid Output KT v2 300 cycles; Illumina Catalogue No. FC-404–2003) at a final concentration of 1.8 pM, following the Illumina recommended denaturation and loading parameters—including a 1% PhiX spike (PhiX Control v3; Illumina Catalogue FC-110–3001).

Following *Dixit et al. (2018)*, we sequenced a random selection of ten isolates twice, using DNA obtained from independent cultures, to help in the determination of clones and the analysis of within-host variants (File S1). Bioinformatic analyses of the genome sequences were carried out on the Cloud Infrastructure for Microbial Bioinformatics (CLIMB) platform (*Connor et al., 2016*).

## Genome assembly and phylogenetic analysis

The paired 150 bp reads were concatenated, then quality checked using the FastQC tool v0.11.7 (*Wingett & Andrews, 2018*) and assembled using SPAdes genome assembler v3.12.0 (*Bankevich et al., 2012*), under default parameters. The quality of the assemblies was assessed using QUAST v5.0.0, de6973bb (*Gurevich et al., 2013*). We used Snippy v4.3.2 (https://github.com/tseemann/snippy)—a rapid command line tool that finds SNPs (substitutions and insertions/deletions) between a haploid reference genome and input sequence reads and generates a core SNP alignment which can be used to reconstruct a high-resolution phylogeny—to generate a core-genome alignment based on core SNPs under default parameters. The complete genome sequence of commensal *E. coli* str. K12 substr. MG1655 was used as a reference strain (NCBI accession: NC_000913.3). From the core-genome alignment, we then reconstructed a maximum-likelihood phylogeny with 1,000 bootstrap replicates using RAxML v8.2.4 (*Stamatakis, 2006*), based on a general time-reversible nucleotide substitution model. The phylogenetic tree was rooted using the genomic sequence of *E. fergusonii* as an outgroup (NCBI accession: GCA_000026225.1). The phylogenetic tree was visualised in FigTree v1.4.3 (https://github.com/rambaut/figtree/) and annotated in RStudio v3.5.1 and Adobe Illustrator v 23.0.3 (Adobe Inc., San Jose, California). As recombination is known to be widespread in *E. coli* and can blur phylogenetic signals (*Wirth et al., 2006*), we detected and masked any recombinant regions of the core-genome alignment using Gubbins (Genealogies Unbiased By recomBinations In Nucleotide Sequences) (*Croucher et al., 2015*) before the phylogenetic reconstruction. For visualisation, a single colony was chosen to represent replicate colonies of the same strain (ST) with identical virulence, plasmid and antimicrobial resistance profiles and a de-replicated phylogenetic tree reconstructed using the representative isolates. We computed pairwise SNP distances between genomes from the core-genome alignment using snp-dists v0.6 (https://github.com/tseemann/snp-dists).

### Multi-locus sequence typing and Clermont typing

The merged reads were uploaded to EnteroBase (*Zhou et al., 2020*), where de novo assembly and genome annotation were carried out, and in-silico multi-locus sequence types (MLST) assigned based on the Achtman scheme, allocating new sequence types (ST) if necessary. EnteroBase assigns phylogroups using ClermontTyper and EzClermont (*Clermont et al., 2013*; *Clermont, Gordon & Denamur, 2015*) and unique core-genome MLST types (cgMLST) based on 2, 513 core loci in *E. coli*. Publicly available *E. coli* sequences in EnteroBase (http://enterobase.warwick.ac.uk/species/index/ecoli) (*Zhou et al., 2020*) were included for comparative analysis, including 23 previously sequenced isolates obtained from diarrhoeal cases recruited in the GEMS study in the Gambia (File S2). The isolates can be searched in EnteroBase using the 'Search Strains' parameter and under 'Strain Metadata', selecting the 'Name' option and entering the study sample name (column 1 of File S2) in the 'Value' box.

### Determination of immigration events and within-host variants

For the whole genome sequences of the strains sequenced twice, we used SPAdes v3.13.2 (*Bankevich et al., 2012*) to assemble each set of reads and map the raw sequences from one sequencing run to the assembly of the other run and vice versa, as described previously (*Dixit et al., 2018*). Briefly, mapping was done using the BWA-MEM algorithm v0.7.17-r1188 under default parameters to generate a SAM alignment. This was then converted to BAM files using Samtools view v1.9 (*Li et al., 2009*), sorted and indexed. Next, variants were called and written to a VCF file using Samtools mpileup and the "view" module of BCFtools (which is part of the Samtools v1.9 package) and visualised in Tablet v1.19.09.13 (*Milne et al., 2013*). The number of SNPs, and their positions were determined and compared between the two steps, counting only those SNPs that were detected in both sets of reads as accurate.

In line with *Dixit et al. (2018)*, isolates belonging to different STs recovered from the same host were considered to be separate strains derived from independent exposures and immigration events. As described in Dixit et al. (2018), we determined the number of SNP differences that existed between assemblies of the same isolate that were sequenced on two separate occasions, to determine if multiple isolates of the same ST from a single host were distinct variants (clones). If the SNP difference between two isolates belonging to the same ST recovered from the same host was less than the SNP difference between the sequences of the same isolate sequenced on two separate occasions, then the two isolates were taken to represent replicate copies of the same clone. Otherwise, they were considered as within-host variants (separate, distinct clones of the same strain)—provided the SNP differences between such distinct clones were no more than eleven SNPs. This cut-off was chosen based on an estimated mutation rate of 1.1 SNP per genome per year (Reeves et al., 2011), assuming equal rates of mutation in both genomes being compared. Based on these data, we inferred replicate clones with SNP differences of greater than 11 SNPs to represent a divergence of more than five years. Thus, it seems implausible that such replicate clones would have emerged from within-host evolution, considering the age of the study participants (<5 years old).

We produced a contingency table to summarise the distribution of variants derived from migration events and within-host evolution and visualised this using a clustered bar graph. We then performed Fisher's exact test to investigate the association between phylogroup and the distribution of variants (migration versus within-host evolution). Our calculations were based on the assumption of independence among the observed phylogroups—that is, the finding of one phylogroup does not preclude or predict the co-occurrence of another.

## Accessory gene content

We used ABRicate v0.9.8 (https://github.com/tseemann/abricate) to predict virulence factors, acquired antimicrobial resistance (AMR) genes and plasmid replicons by scanning the contigs against the VFDB, ResFinder and PlasmidFinder databases respectively, using an identity threshold of $\geq$ 90% and a coverage of $\geq$ 70%. Virulence factors and AMR genes were plotted next to the phylogenetic tree using the ggtree, ggplot2 and phangorn packages in RStudio v3.5.1. We calculated co-occurrence of AMR genes among study isolates by transforming the binary AMR gene content matrix and visualising this as a heat map using the pheatmap package v 1.0.12 (https://CRAN.R-project.org/package=pheatmap) in RStudio v3.5.1. We computed Fisher's exact tests between the detected virulence factors and the observed phylogroups in RStudio v3.5.1.

## Population structure and comparison of commensal and pathogenic strains

We assessed the population structure using the hierarchical clustering algorithm in EnteroBase. Briefly, the isolates were assigned stable population clusters at eleven levels (from HC0 to HC 2350) based on pairwise cgMLST allelic differences. Hierarchical clustering at 1,100 alleles differences (HC1100) resolves populations into cgST (core-genome MLST type) complexes, the equivalent of clonal complexes achieved with the legacy MLST clustering approaches (Zhou et al., 2020). We reconstructed neighbour-joining phylogenetic trees using NINJA (Wheeler, 2009), based on clustering at HC1100 to display the population sub-clusters at this level as an indicator of the genomic diversity within our study population and to infer the evolutionary relationship among our strains and others in the public domain.

Next, we interrogated the HC1100 clusters that encompassed our study isolates and Gambian pathogenic isolates recovered from diarrhoeal cases and commensal *E. coli* strains recovered from the GEMS study. For the clusters that encompassed commensal and pathogenic strains belonging to the same ST (HC1100_200 cluster, comprising pathogenic isolates from GEMS cases 100415, 102106 and 102098 and the resident ST38 strain recovered from our study subject 18), we reconstructed both neighbour-joining and SNP phylogenetic trees to display the genetic relationships among these strains. We visualised the accessory genomes for the overlapping STs mentioned above to determine genes associated with phages, virulence factors and AMR. The resulting phylogenetic trees were annotated in Adobe Illustrator v 23.0.3 (Adobe Inc., San Jose, California).

### Ethical statement

The parent study was approved by the joint Medical Research Council Unit The Gambia-Gambian Government ethical review board (SCC 1331). Written informed consents were obtained from all the study participants as previously reported in *Kotloff et al. (2013)*. The Medical Research Council Unit The Gambia at London School of Hygiene and Tropical Medicine's Scientific Coordinating Committee gave approval for the use of the stool samples analysed in this study.

## RESULTS

### Population structure

The study population included 27 females and 39 males (File S3). All but one reported the presence of a domestic animal within the household. Twenty-one samples proved positive for the growth of *E. coli*, yielding 88 isolates (File S4). We detected 37 seven-allele sequence types (STs) among the isolates, with a fairly even distribution (Fig. 2). Five STs were completely novel (ST9274, ST9277, ST9278, ST9279 and ST9281). These study strains were scattered over all the eight main phylogroups of *E. coli*: A (27%), B1 (32%), B2 (9%), D (15%), C and F (5% each), E (1%), and the cryptic Clade I (7%), although the majority belonged to phylogroups A and B1 (Table 1). Hierarchical clustering of core genomic STs revealed twenty-seven cgST clonal complexes (File S4). The raw genomic sequences of the study isolates have been deposited in the NCBI SRA under the BioProject ID PRJNA658685, (accession numbers SAMN15880274 to SAMN15880361).

### Within-host diversity

Just a single ST colonised nine individuals, six carried two STs, four carried four STs, and two carried six STs. We found 56 distinct genotypes, which equates to an average of 2.7 genotypes per host. Two individuals (H-18 and H-2) shared an identical strain belonging to ST9274 (zero SNP difference) (File S5, yellow highlight), suggesting recent transfer from one child to another or recent acquisition from a common source.

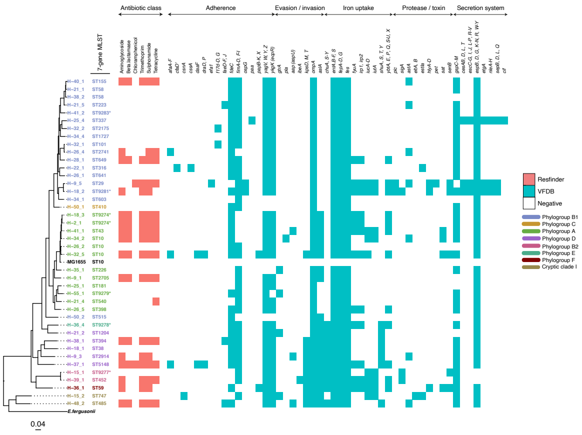


**Figure 2 A maximum-likelihood tree depicting the phylogenetic relationships among the study isolates.** The tree was reconstructed with RAxML, using a general time-reversible nucleotide substitution model and 1,000 bootstrap replicates. The genome assembly of *E. coli* str. K12 substr. MG1655 was used as the reference, and the tree rooted using the genomic assembly of *E. fergusonii* as an outgroup. The sample names are indicated at the tip, with the respective Achtman sequence types (ST) indicated beside the sample names. The respective phylogroups the isolates belong to are indicated with colour codes as displayed in the legend. The *E. coli* reference genome and *E. fergusonii* are denoted in black. Asterisks (*) are used to indicate novel STs. The predicted antimicrobial resistance genes and putative virulence factors for each isolate are displayed next to the tree, with the virulence genes clustered according to their function. Multiple copies of the same strain (ST) isolated from a single host are not shown. Instead, we have shown only one representative isolate from each strain. Virulence and resistance factors were not assessed in the reference strains either. A summary of the identified virulence factors and their known functions are provided in File S3.

Full-size DOI: 10.7717/peerj.10572/fig-2

We observed thirteen within-host variants in ten hosts (intra-clonal diversity) (subjects H-15, H-18, H-22, H-25, H-28, H-34, H-36, H-37, H-38 and H-39), compared to forty-one immigration events (Tables 1 and 2). Overall, immigration events accounted for the majority (76%) of variants (Fig. S1). The proportion of migration versus within-host evolution events did not appear to be affected by phylogroup ($p = 0.42$). Twenty-two percent of within-host mutations represented synonymous changes, 43% were non-synonymous mutations, while 31% occurred in non-coding regions, and 4% represented stop-gained mutations (File S6). On an average, Ka/Ks ratios were greater than 1, which seems to suggest that these mutations were under positive Darwinian selection—indicating that most of the mutations were likely to have little effect on fitness. However, these remain to be investigated further. Also, the observed non-synonymous mutations were spread across genes with a variety of functions, including metabolism, transmembrane transport, pathogenesis and iron import into the cell. However, the bulk (42%) occurred in genes involved in metabolism. The average number of SNPs among within-host variants was

**Table 1** **Phylogroup and sequence types of the distinct clones isolated in each study subject.**

| | Colony or isolate number | | | | | Number of distinct genotypes (clones) | Migration events | Within-host evolution events |
|---|---|---|---|---|---|---|---|---|
| Host | 1 | 2 | 3 | 4 | 5 | | Phylotype (number of events) | Phylotype (number of events) |
| H-2 | A (9274) | A (9274) | A (9274) | A (9274) | A (9274) | 1 | A (1) | 0 |
| H-9 | A (2705) | A (2705) | A (2705) | D (2914) | B1 (29) | 3 | A (1), D (1), B1 (1) | 0 |
| H-15 | B2 (9277) | B2 (9277) | B2 (9277) | Clade I (747) | Clade I (747) | 3 | B2 (1), Clade I (1) | Clade I (1) |
| H-18 | D (38) | D (38) | B1 (9281) | A (9274) | | 4 | D (1), B1 (1), A (1) | D (1) |
| H-21 | B1 (58) | B1 (58) | B1 (223) | A (540) | D (1204) | 4 | B1(2) A (1), D (1) | 0 |
| H-22 | B1 (316) | B1 (316) | B1 (316) | B1 (316) | | 2 | B (1) | B1(1) |
| H-25 | A (181) | A (181) | A (181) | A (181) | B1 (337) | 4 | A (1), B1 (1) | A (2) |
| H-26 | B1 (641) | B1 (2741) | A (10) | A (398) | | 4 | B1(2), A (1), D (1) | 0 |
| H-28 | B1 (469) | B1 (469) | B1 (469) | B1 (469) | | 2 | B1(1) | B1(1) |
| H-32 | B1 (101) | B1 (101) | B1 (101) | B1 (2175) | A (10) | 3 | B1(2), A (1) | 0 |
| H-34 | B1 (603) | B1 (603) | B1 (603) | B1 (1727) | A (10) | 4 | B1(2), A (1) | B1(1) |
| H-35 | A (226) | | | | | 1 | A (1) | 0 |
| H-36 | F (59) | F (59) | F (59) | F (59) | E (9278) | 4 | F (1), E (1) | F (1) |
| H-37 | D (5148) | D (5148) | D (5148) | D (5148) | D (5148) | 3 | D (1) | D (2) |
| H-38 | D (394) | D (394) | D (394) | D (394) | B1 (58) | 4 | D (1), B1(1) | D (2) |
| H-39 | B2 (452) | B2 (452) | B2 (452) | B2 (452) | B2 (452) | 2 | B2(1) | B2 (1) |
| H-40 | B1 (155) | | | | | 1 | B1(1) | 0 |
| H-41 | A (43) | A (43) | A (43) | A (43) | B1 (9283) | 2 | A (1), B1(1) | 0 |
| H-48 | Clade I (485) | Clade I (485) | Clade I (485) | Clade I (485) | | 3 | Clade I (1) | 0 |
| H-50 | C (410) | C (410) | C (410) | C (410) | B1 (515) | 2 | C (1), B1(1) | 0 |
| H-55 | A (9279) | | | | | 1 | A(1) | 0 |

5 (range 0–18) (Table 2). However, in two subjects (H36 and H37), pairwise distances between genomes from the same ST (ST59 and ST5148) were as large as 14 and 18 SNPs respectively (File S5, grey highlight).

## Accessory gene content and relationships with other strains

A quarter of our isolates were most closely related to commensal strains from humans, with smaller numbers most closely related to human pathogenic strains or strains from livestock, poultry or the environment (File S7). One isolate was most closely related to a canine isolate from the UK. Three STs (ST38, ST10 and ST58) were shared by our study isolates and diarrhoeal isolate from the GEMS study (Fig. S2), with just eight alleles separating our commensal ST38 strain from a diarrhoeal isolate from the GEMS study (Fig. 3). For ST10 and ST58, hierarchical clustering placed the commensal strains from this study into separate clusters from the pathogenic isolates from diarrhoeal cases, indicating that they were genetically distinct to each other. Yet, the closest relative of our study ST58 strain was an extraintestinal strain isolated from the blood of a 69-year-old male (87 alleles differences, Fig. 4). Also, the resident ST10 isolates recovered from this study (H-26_2,

**Table 2** Pairwise SNP distances between variants arising from within-host evolution.

| Host | Sequence type (ST) | Colonies per ST | Pairwise SNP distances between multiple colonies of the same ST |
|---|---|---|---|
| H2 | 9274 | 5 | 0–9 |
| H9 | 2705 | 3 | 0–1 |
| H15 | 9277 | 3 | 0–1 |
| H15 | 747 | 2 | 3 |
| H18 | 38 | 2 | 3 |
| H21 | 58 | 2 | 0 |
| H22 | 316 | 4 | 0–3 |
| H25 | 181 | 4 | 1–5 |
| H28 | 469 | 4 | 0–3 |
| H32 | 101 | 3 | 1–9 |
| H34 | 603 | 3 | 2–8 |
| H36 | 59 | 4 | 0–14 |
| H37 | 5148 | 5 | 2–18 |
| H38 | 394 | 4 | 1–3 |
| H39 | 452 | 5 | 0–2 |
| H41 | 43 | 4 | 0–1 |
| H48 | 485 | 4 | 1–9 |
| H50 | 410 | 4 | 0 |

H-34_2, and H-32_5) had their closest neighbours in isolates from livestock (83 and 111 alleles each), and an isolate of an unspecified source (18 alleles differences) respectively (File S7).

We detected 130 genes encoding putative virulence factors across the 88 study isolates (Fig. 2; File S8). Notable among these were genes associated with pathogenesis in Enteroaggregative *E. coli* and *Salmonella* referred to as the Serine Protease Autotransporters of *Enterobacteriaceae* (SPATEs) (*Pokharel et al., 2019*), such as *sat* (13%), *sigA* (11%) and *pic* (1%). Besides, eight isolates harboured known markers of Enteropathogenic *E. coli* (*eltAB* or *estA*). Several strains (across all phylogroups) also harboured virulence genes associated with intestinal or extraintestinal disease in humans, including adhesins, invasins, toxins and iron-acquisition genes such as *fyuA*, several *fim* and *pap genes*, *iroN*, *irp1,* 2, *ibeA* and *aslA*. We did not detect any of the well-known markers of EPEC (*eae, bfpA, stx1*, or *stx2*) (Fig. 2, File S8).

The prevalence of some virulence factors involved in invasion/evasion, iron uptake, adherence, and secretion systems appeared to be more or less likely to occur in one or a few phylotypes ($p \leq 0.05$) as follows (File S9). The iron acquisition genes *chuA, S-Y* and *shuA, S, T, Y* were found to be present in all cases for phylogroup D ($n = 5$), and absent in virtually all cases for phylogroups A ($n = 13$) and B1 ($n = 16$). On the other hand, *iutA* and *iucA-D* were observed in the two cases from phylogroup B2, and absent from all samples from phylogroup D ($n = 5$). The invasion/evasion genes *kpsD, M, T* and *aslA* were found

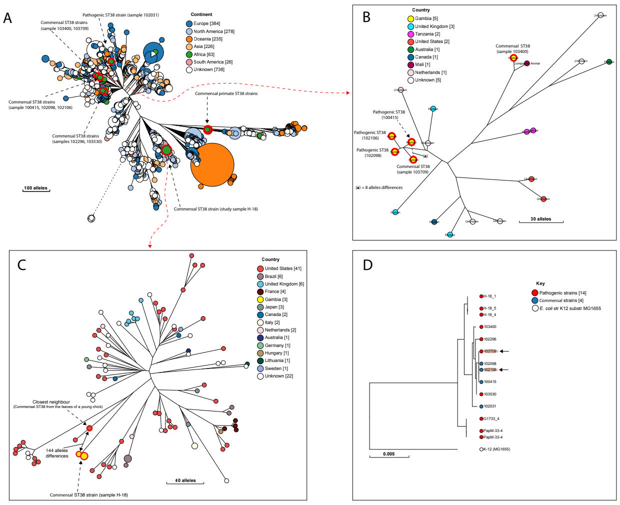


**Figure 3** **The population structure of ST38.** (A) A NINJA neighbour-joining tree showing the population structure of *E. coli* ST38, drawn using the genomes found in the core-genome MLST hierarchical cluster at HC1100, which corresponds to ST38 clonal complex. The size of the nodes represents the number of isolates per clade. The geographical locations where isolates were recovered are displayed in the legend; with the genome counts shown in square brackets. The study resident ST38 strains and the pathogenic ST38 strains recovered from GEMS cases are highlighted with red circles around the nodes. (B) The closest neighbour to a pathogenic strain reported in GEMS *Kotloff et al. (2013)* is shown to be a commensal isolate recovered from a healthy individual. The size of the nodes represents the number of isolates per clade. The geographical locations where isolates were recovered are displayed in the legend; with the genome counts shown in square brackets. Red circles around the nodes are used to highlight the study resident ST38 strains and the pathogenic ST38 strains recovered from GEMS cases within this cluster. (C) The closest relatives to the commensal ST38 strain recovered from this study is shown (red highlights), with the number of core-genome MLST alleles separating the two genomes displayed. The geographical locations where isolates were recovered are displayed in the legend; with the genome counts shown in square brackets, and the size of the nodes depicting the number of isolates per clade. (D) A maximum-likelihood phylogenetic tree reconstructed using the genomes found in the cluster in C above, comprising both pathogenic and commensal ST38 strains is presented, depicting the genetic relationship between strain 100415 (pathogenic) and 103709 (commensal) (red highlights). The nodes are coloured to depict the status of the strains, as pathogenic (red) or commensal (blue). The size of the nodes represents the number of isolates per clade. The geographical locations where isolates were recovered are displayed in the legend; with the genome counts shown in square brackets.

Full-size DOI: 10.7717/peerj.10572/fig-3

to be present in almost all cases for phylogroups D ($n=5$), B2 ($n=2$), and Clade I ($n=2$), and absent in B1 ($n=16$). The secretion system gene cluster *espB, D, G, K-N, R, W-Y* was observed in all cases except the two belonging to phylogenetic group B2. The protease gene *sigA* was absent from most samples, except two samples from phylotype B2. The adherence gene *fdeC* was observed in all cases for phylotype D ($n=5$) and most for B1 ($n=16$).

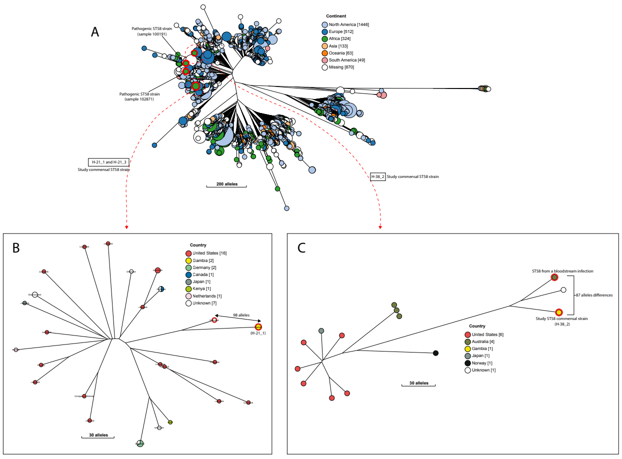


**Figure 4** **The population structure of ST58.** (A) A NINJA neighbour-joining tree depicting the population structure of *E. coli* ST58, drawn using the genomes found that clustered together in the same HC1100 hierarchical cluster in the core-genome MLST scheme in EnteroBase (*Zhou et al., 2020*). Commensal ST58 strains from this study and Gambian pathogenic ST58 isolates from GEMS are highlighted in red. The geographical locations where isolates were recovered are displayed in the legend; with the genome counts shown in square brackets. The size of the nodes represents the number of isolates per clade. (B and C) The closest relatives to the study ST58 strains are shown. Geographical locations where isolates were recovered are displayed in the legend, with the genome counts displayed in square brackets. The red highlights around the nodes depict the study commensal ST58 strains and their closest neighbours. The size of the nodes represents the number of isolates per clade, and the geographical locations where isolates were recovered are displayed in the legend; with the genome counts shown in square brackets.

Full-size 🖼 DOI: 10.7717/peerj.10572/fig-4

More than half of the isolates encoded resistance to three or more clinically relevant classes of antibiotics such as aminoglycosides, penicillins, trimethoprim, sulphonamides and tetracyclines (Fig. 5; Fig. S3). The most common resistance gene network was *-aph(6)-Id_1-sul2* (41% of the isolates), followed by *aph(3′′)-Ib_5-sul2* (27%) and *bla-TEM-aph(3′′)-Ib_5* (24%). Most isolates (67%) harboured two or more plasmid types (Fig. 6). Of the 24 plasmid types detected, IncFIB was the most common (41%), followed by col156 (19%) and IncI_1-Alpha (15%). Nearly three-quarters of the multi-drug resistant isolates carried IncFIB (AP001918) plasmids (∼50 kb), suggesting that these large plasmids may be linked to the dissemination of resistance genes within our study population.

## DISCUSSION

This study provides an overview of the within-host genomic diversity of *E. coli* in healthy children from a rural setting in the Gambia, West Africa. Surprisingly, we were able to recover *E. coli* from only 34% of stools which had previously tested positive for *E. coli* in the original study. This low rate of recovery may reflect some hard-to-identify effect

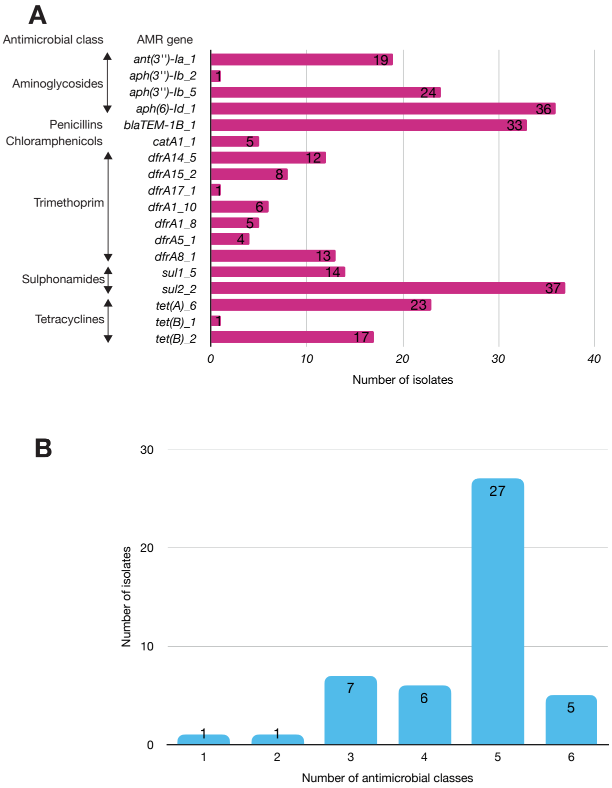


**Figure 5** **The prevalence of antimicrobial-associated genes detected in the study isolates.** (A) The *y*-axis shows the prevalence of the detected AMR-associated genes in the study isolates, grouped by antimicrobial class. (B) A histogram depicting the number of antimicrobial classes to which resistance genes were detected in the corresponding strains.

Full-size 🖼 DOI: 10.7717/peerj.10572/fig-5

of long-term storage (nine to thirteen years) or the way the samples were handled, even though they were kept frozen and thawed only just before culture.

Several studies have shown that sampling a single colony is insufficient to capture *E. coli* strain diversity in stools (*Dixit et al., 2018*; *Richter et al., 2018*; *Shooter et al., 1977*). *Lidin-Janson et al. (1978)* claim that sampling five colonies provides a >99% chance of recovering dominant genotypes from single stool specimens, while *Schlager et al. (2002)* calculate that sampling twenty-eight colonies provides a >90% chance of recovering minor genotypes. Our results confirm the importance of multiple-colony picks in faecal

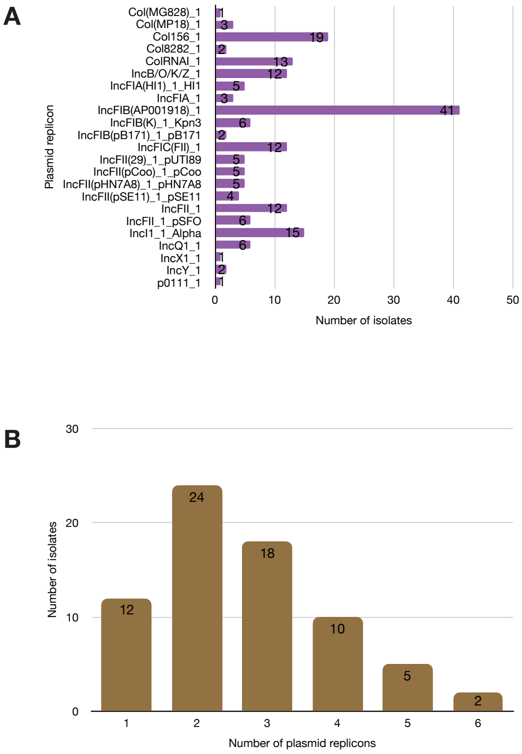


**Figure 6** **Prevalence of plasmid replicons among the study isolates.** (A) Plasmid replicons detected in the study isolates. (B) A histogram depicting the number of plasmids co-harboured in a single strain.

Full-size DOI: 10.7717/peerj.10572/fig-6

surveillance studies, as over half (57%) of our strains would have been missed by picking a single colony.

Our strains encompassed all eight major phylotypes of *E. coli,* however, the majority fell into the A and B1 phylogenetic groups, in line with previous reports that these phylogroups dominate in stools from people in low- and middle-income countries (*Duriez et al., 2001*; *Escobar-Paramo et al., 2004a*; *Escobar-Páramo et al., 2004b*). Although not fully understood, there appear to be host-related factors that influence the composition of *E. coli* phylogroups in human hosts. For example, the establishment of strains belonging to phylogroups E or F seems to favour subsequent colonisation by other phylotypes, compared to the establishment of phylogroup B2 strains, which tend to limit the heterogeneity within individual hosts (Gordon, O'Brien & Pavli, 2015). Geographical differences have also been reported, with phylogroups A and B1 frequently dominating the stools of people living in developing countries (Duriez et al., 2001; Escobar-Paramo et al., 2004a; Escobar-Páramo et al., 2004b). Conversely, phylogroups B2 and D strains appear to be pervasive among people living in developed countries (Massot et al., 2016; Skurnik et al., 2008). These locale-specific patterns in the distribution of *E. coli* phylotypes have been attributed to differences in diet and climate (Duriez et al., 2001; Escobar-Paramo et al., 2004a; Escobar-Páramo et al., 2004b).

The prevalence of putative virulence genes in most of our isolates highlights the pathogenic potential of commensal intestinal strains—regardless of their phylogroup—should they gain access to the appropriate tissues, for example, the urinary tract. Our results complement previous studies reporting genomic similarities between faecal *E. coli* isolates and those recovered from urinary tract infection (McNally et al., 2013; Wold et al., 1992).

We found that within-host evolution plays a minor role in the generation of diversity in our study population. This might be due to the low prevalence of B2 strains, which are thought to inhibit the establishment of strains from other phylogroups, as discussed above (Gordon, O'Brien & Pavli, 2015); or it may indicate that members of phylogroups A and B1 might favour a more heterogeneous composition of *E. coli* phylotypes in stools of healthy individuals. However, this remains to be properly investigated, as we did not find statistical evidence that the distribution of variants (independent migration versus within-host evolution) was influenced by phylogroup. Our findings are similar to that reported by Dixit et al. (2018), who reported that 83% of diversity originates from immigration events, and with epidemiological data suggesting that the recurrent immigration events account for the high faecal diversity of *E. coli* in the tropics (Tenaillon et al., 2010).

The estimated mutation rate for *E. coli* lineages is around one SNP per genome per year (Reeves et al., 2011), so that two genomes with a most recent common ancestor in the last five years would be expected to be around ten SNPs apart. However, in two subjects, pairwise distances between genomes from the same ST (ST59 and ST5148) were large enough (14 and 18 respectively) to suggest that they might have arisen from independent immigration events, as insufficient time had elapsed in the child's life for such divergence to occur within the host. However, it remains possible that the mutation rate was higher than expected in these lineages, although we found no evidence of damage to DNA repair genes. Co-colonising variants belonging to the same ST tended to share an identical virulence, AMR and plasmid profile, signalling similarities in their accessory gene content.

The sources of novel variation that account for within-host diversity include point mutation and small insertions or deletions (indels), indels and the loss or acquisition of mobile genetic elements. Among the variants inferred to have been derived from within-host evolution, we observed a dominance of mutations that were predicted to result in changes in protein function, in the form of missense mutations and non-sense mutations (leading to a premature stop codon). Although the mutations appeared to be heterogeneously distributed, a higher number was observed in genes associated with metabolism. These appeared to be under positive selection, although it remains to be seen if these changes confer any effects on fitness. It will be desirable to investigate this in future studies. Due to the cross-sectional nature of our sampling, we were unable to analyse the dynamics of strain gain or loss and variation in gene content over time. Homologous recombination has also been noted to contribute to the generation of diversity (Golubchik et al., 2013; González-González et al., 2013), however, we detected and removed recombinant regions prior to phylogenetic reconstruction and thus focused our analysis on SNPs.

More than half of our isolates encode resistance to three or more classes of antimicrobials echoing the high rate of MDR (65%; confirmed by phenotypic testing) in the GEMS study. IncFIB (AP001918) was the most common plasmid Inc type from our study, in line with the observation that IncF plasmids are frequently associated with the dissemination of resistance (Carattoli, 2009). However, a limitation of our study is that we did not perform phenotypic antimicrobial resistance testing, although Doyle et al. (2020) reported that only a small proportion of genotypic AMR predictions are discordant with phenotypic results.

Comparative analyses confirm the heterogeneous origins of the strains reported here, documenting links to other human commensal strains or isolates sourced from livestock or the environment. This is not surprising, as almost all the study participants reported that animals are kept in their homes and children in rural Gambia are often left to play on the ground, close to domestic animals such as pets and poultry (Dione et al., 2011).

## CONCLUSIONS

Our results show that the commensal *E. coli* population in the gut of healthy children in rural Gambia is richly diverse, with the independent immigration and establishment of strains contributing to the bulk of the observed diversity. An obvious limitation to our study is the low recovery of *E. coli* from frozen stools—which potentially implies we may have underestimated the extent of genetic diversity present within our study population. Although solely observational, our study paves the way for future studies aimed at a mechanistic understanding of the factors driving the diversification of *E. coli* in the human gut and what it takes to make a strain of *E. coli* successful in this habitat. Besides, this work has added significantly to the number of commensal *E. coli* genomes, which are underrepresented in public repositories.

## ACKNOWLEDGEMENTS

We gratefully acknowledge the study participants in GEMS and all clinicians, field workers and the laboratory staff of the Medical Research Council Unit The Gambia at London School of Hygiene and Tropical Medicine involved in the collection and storage of stools in the GEMS study in Basse Field Station and Fajara.

## ADDITIONAL INFORMATION AND DECLARATIONS

### Funding

Martin Antonio, Usman N. Ikumapayi, Golam Sarwar, Catherine Okoi, Peggy-Estelle Maguiagueu Tientcheu and Mark J. Pallen were supported by the Medical Research

Council Unit, The Gambia at London School of Hygiene and Tropical Medicine. The BBSRC Institute Strategic Programme, Microbes in the Food Chain (BB/R012504/1 and its constituent projects 44414000A and 4408000A) supported Ebenezer Foster-Nyarko and Mark J. Pallen. Marianne Defernez and Nabil-Fareed Alikhan were supported by the Quadram Institute Bioscience BBSRC funded Core Capability Grant (project number BB/CCG1860/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Grant Disclosures

The following grant information was disclosed by the authors:
Medical Research Council Unit, The Medical Research Council Unit The Gambia at London School of Hygiene and Tropical Medicine" is one name.
The Gambia at London School of Hygiene and Tropical Medicine.
BBSRC Institute Strategic Programme.
Microbes in the Food Chain (BB/R012504/1 and its constituent projects 44414000A and 4408000A).
Core Capability Grant: BB/ CCG1860/1.

## Competing Interests

The authors declare there are no competing interests.

## Author Contributions

- Ebenezer Foster-Nyarko conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Nabil-Fareed Alikhan, Usman N. Ikumapayi, Golam Sarwar, Catherine Okoi, Peggy-Estelle Maguiagueu Tientcheu, Marianne Defernez and Justin O'Grady performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Martin Antonio conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Mark J. Pallen conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

## Human Ethics

The following information was supplied relating to ethical approvals (i.e., approving body and any reference numbers):

The study was approved by the Medical Research Council Unit The Gambia at London School of Hygiene and Tropical Medicine's Scientific Coordinating Committee.

## DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

All genomic assemblies for the strains included in this study are freely available from EnteroBase (http://enterobase.warwick.ac.uk/species/index/ecoli). The EnteroBase genome assembly barcodes are available in the Supplemental Files. The isolates can be found in

EnteroBase using the 'Search Strains' parameter and under 'Strain Metadata', selecting the 'Name' option and entering the study sample name in the 'Value' box.

The raw genomic sequences are available at NCBI SRA, BioProject ID: PRJNA658685, (SAMN15880274 to SAMN15880361).

## Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.10572#supplemental-information.

## REFERENCES

**Ahmed SF, Ali MM, Mohamed ZK, Moussa TA, Klena JD. 2014.** Fecal carriage of extended-spectrum β-lactamases and AmpC-producing *Escherichia coli* in a Libyan community. *Annals of Clinical Microbiology and Antimicrobials* **13**:22–30 DOI 10.1186/1476-0711-13-22.

**Alm EW, Walk ST, Gordon DM. 2011.** The niche of *Escherichia coli*. In: Walk ST, Feng PCH, eds. *Population Genetics of Bacteria: American Society of Microbiology, Chapter 6*. Washington, D.C.: ASP Press, 67–89 DOI 10.1128/9781555817114.ch6.

**Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012.** SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19**:455–477 DOI 10.1089/cmb.2012.0021.

**Bettelheim KA, Faiers M, Shooter RA. 1972.** Serotypes of *Escherichia coli* in normal stools. *The Lancet* **2**:1223–1224.

**Blount ZD. 2015.** The unexhausted potential of *E. coli*. *Elife* **4**:e05826 DOI 10.7554/eLife.05826.

**Camins BC, Marschall J, DeVader SR, Maker DE, Hoffman MW, Fraser VJ. 2011.** The clinical impact of fluoroquinolone resistance in patients with *E. coli* bacteremia. *Journal of Hospital Medicine* **6**:344–349 DOI 10.1002/jhm.877.

**Carattoli A. 2009.** Resistance plasmid families in *Enterobacteriaceae*. *Antimicrobial Agents and Chemotherapy* **53**:2227–2238 DOI 10.1128/AAC.01707-08.

**Chen SL, Wu M, Henderson JP, Hooton TM, Hibbing ME, Hultgren SJ, Gordon JI. 2013.** Genomic diversity and fitness of *E. coli* strains recovered from the intestinal and urinary tracts of women with recurrent urinary tract infection. *Science Translational Medicine* **5**:184ra160–184ra174 DOI 10.1126/scitranslmed.3005497.

**Clermont O, Christenson JK, Denamur E, Gordon DM. 2013.** The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environmental Microbiology Reports* **5**:58–65 DOI 10.1111/1758-2229.12019.

**Clermont O, Gordon D, Denamur E. 2015.** Guide to the various phylogenetic classification schemes for *Escherichia coli* and the correspondence among schemes. *Microbiology* **161**:980–988 DOI 10.1099/mic.0.000063.

**Connor TR, Loman NJ, Thompson S, Smith A, Southgate J, Poplawski R, Bull MJ, Richardson E, Ismail M, Thompson SE, Kitchen C, Guest M, Bakke M, Sheppard SK, Pallen MJ. 2016.** CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *Microbial Genomics* **2**:e000086 DOI 10.1099/mgen.0.000086.

**Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015.** Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* **43**:e15–e15 DOI 10.1093/nar/gku1196.

**Dione MM, Ikumapayi UN, Saha D, Mohammed NI, Geerts S, Ieven M, Adegbola RA, Antonio M. 2011.** Clonal differences between Non-Typhoidal *Salmonella* (NTS) recovered from children and animals living in close contact in the Gambia. *PLOS Neglected Tropical Diseases* **5**:e1148 DOI 10.1371/journal.pntd.0001148.

**Dixit OVA, O'Brien CL, Pavli P, Gordon DM. 2018.** Within-host evolution versus immigration as a determinant of *Escherichia coli* diversity in the human gastrointestinal tract. *Environmental Microbiology* **20**:993–1001 DOI 10.1111/1462-2920.14028.

**Doyle RM, O'Sullivan DM, Aller SD, Bruchmann S, Clark T, Coello Pelegrin A, Cormican M, Diez Benavente E, Ellington MJ, McGrath E, Motro Y, Phuong Thuy Nguyen T, Phelan J, Shaw LP, Stabler RA, Van Belkum A, Van Dorp L, Woodford N, Moran-Gilad J, Huggett JF, Harris KA. 2020.** Discordant bioinformatic predictions of antimicrobial resistance from whole-genome sequencing data of bacterial isolates: an inter-laboratory study. *Microbial Genomics* **6**:e000335 DOI 10.1099/mgen.0.000335.

**Duriez P, Clermont O, Bonacorsi S, Bingen E, Chaventré A, Elion J, Picard B, Denamur E. 2001.** Commensal *Escherichia coli* isolates are phylogenetically distributed among geographically distinct human populations. *Microbiology* **147**:1671–1676 DOI 10.1099/00221287-147-6-1671.

**Escobar-Paramo P, Clermont O, Blanc-Potard AB, Bui H, Le Bouguenec C, Denamur E. 2004a.** A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. *Molecular Biology and Evolution* **21**:1085–1094 DOI 10.1093/molbev/msh118.

**Escobar-Páramo P, Grenet K, Le Menac'h A, Rode L, Salgado E, Amorin C, Gouriou S, Picard B, Rahimy MC, Andremont A, Denamur E, Ruimy R. 2004b.** Large-scale population structure of human commensal *Escherichia coli* isolates. *Applied and Environmental Microbiology* **70**:5698–5700 DOI 10.1128/AEM.70.9.5698-5700.2004.

**Ferjani S, Saidani M, Hamzaoui Z, Alonso CA, Torres C, Maamar E, Slim AF, Boutiba BB. 2017.** Community fecal carriage of broad-spectrum cephalosporin-resistant *Escherichia coli* in Tunisian children. *Diagnostic Microbiology and Infectious Disease* **87**:188–192 DOI 10.1016/j.diagmicrobio.2016.03.008.

**Foster-Nyarko E, Alikhan NF, Ravi A, Thilliez G, Thomson NM, Baker D, Kay G, Cramer JD, O'Grady J, Antonio M, Pallen MJ. 2020.** Genomic diversity of *Escherichia coli* isolates from non-human primates in the Gambia. *Microbial Genomics* **6**:e000428 DOI 10.1099/mgen.0.000428.

**Golubchik T, Batty EM, Miller RR, Farr H, Young BC, Larner-Svensson H, Fung R, Godwin H, Knox K, Votintseva A, Everitt RG, Street T, Cule M, Ip CLC, Didelot X, Peto TEA, Harding RM, Wilson DJ, Crook DW, Bowden R. 2013.** Within-host evolution of *Staphylococcus aureus* during asymptomatic carriage. *PLOS ONE* **8**:e61319 DOI 10.1371/journal.pone.0061319.

**González-González A, Sánchez-Reyes LL, Sapien GDelgado, Eguiarte LE, Souza V. 2013.** Hierarchical clustering of genetic diversity associated to different levels of mutation and recombination in *Escherichia coli*: a study based on Mexican isolates. *Infections, Genetics and Evolution* **13**:187–197 DOI 10.1016/j.meegid.2012.09.003.

**Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. 2017.** The dynamics of molecular evolution over 60,000 generations. *Nature* **551**:45–50 DOI 10.1038/nature24287.

**Gordon DM, O'Brien CL, Pavli P. 2015.** *Escherichia coli* diversity in the lower intestinal tract of humans. *Environmental Microbiology Reports* **7**:642–648 DOI 10.1111/1758-2229.12300.

**Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013.** QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**:1072–1075 DOI 10.1093/bioinformatics/btt086.

**Hartl DL, Dykhuizen DE. 1984.** The population genetics of *Escherichia coli*. *Annual Reviews of Genetics* **18**:31–68 DOI 10.1146/annurev.ge.18.120184.000335.

**Hobman JL, Penn CW, Pallen MJ. 2007.** Laboratory strains of *Escherichia coli*: model citizens or deceitful delinquents growing old disgracefully? *Molecular Microbiology* **64**:881–885 DOI 10.1111/j.1365-2958.2007.05710.x.

**Kotloff KL, Blackwelder WC, Nasrin D, Nataro JP, Farag TH, Van Eijk A, Adegbola RA, Alonso PL, Breiman RF, Faruque AS, Saha D, Sow SO, Sur D, Zaidi AK, Biswas K, Panchalingam S, Clemens JD, Cohen D, Glass RI, Mintz ED, Sommerfelt H, Levine MM. 2012.** The Global Enteric Multicenter Study (GEMS) of diarrheal disease in infants and young children in developing countries: epidemiologic and clinical methods of the case/control study. *Clinical Infectious Diseases* **55(Suppl 4)**:S232–245 DOI 10.1093/cid/cis753.

**Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam S, Wu Y, Sow SO, Sur D, Breiman RF, Faruque ASG, Zaidi AKM, Saha D, Alonso PL, Tamboura B, Sanogo D, Onwuchekwa U, Manna B, Ramamurthy T, Kanungo S, Ochieng JB, Omore R, Oundo JO, Hossain A, Das SK, Ahmed S, Qureshi S, Quadri F, Adegbola RA, Antonio M, Hossain MJ, Akinsola A, Mandomando I, Nhampossa T, Acácio S, Biswas K, O'Reilly CE, Mintz ED, Berkeley LY, Muhsen K, Sommerfelt H, Robins-Browne RM, Levine MM. 2013.** Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global

Enteric Multicenter Study, GEMS): a prospective, case-control study. *The Lancet* **382**:209–222 DOI 10.1016/S0140-6736(13)60844-2.

**Laxminarayan R, Duse A, Wattal C, Zaidi AK, Wertheim HF, Sumpradit N, Vlieghe E, Hara GL, Gould IM, Goossens H, Greko C, So AD, Bigdeli M, Tomson G, Woodhouse W, Ombaka E, Peralta AQ, Qamar FN, Mir F, Kariuki S, Bhutta ZA, Coates A, Bergstrom R, Wright GD, Brown ED, Cars O. 2013.** Antibiotic resistance-the need for global solutions. *The Lancet Infectious Diseases* **13**:1057–1098 DOI 10.1016/s1473-3099(13)70318-9.

**Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. 2009.** The sequence alignment/map format and SAMtools. *Bioinformatics* **25**:2078–2079 DOI 10.1093/bioinformatics/btp352.

**Lidin-Janson G, Kaijser B, Lincoln K, Olling S, Wedel H. 1978.** The homogeneity of the faecal coliform flora of normal school-girls, characterized by serological and biochemical properties. *Medical Microbiology and Immunology* **164**:247–253 DOI 10.1007/BF02125493.

**Liu J, Platts-Mills JA, Juma J, Kabir F, Nkeze J, Okoi C, Operario DJ, Uddin J, Ahmed S, Alonso PL, Antonio M, Becker SM, Blackwelder WC, Breiman RF, Faruque AS, Fields B, Gratz J, Haque R, Hossain A, Hossain MJ, Jarju S, Qamar F, Iqbal NT, Kwambana B, Mandomando I, McMurry TL, Ochieng C, Ochieng JB, Ochieng M, Onyango C, Panchalingam S, Kalam A, Aziz F, Qureshi S, Ramamurthy T, Roberts JH, Saha D, Sow SO, Stroup SE, Sur D, Tamboura B, Taniuchi M, Tennant SM, Toema D, Wu Y, Zaidi A, Nataro JP, Kotloff KL, Levine MM, Houpt ER. 2016.** Use of quantitative molecular diagnostic methods to identify causes of diarrhoea in children: a reanalysis of the GEMS case-control study. *Lancet* **388**:1291–1301 DOI 10.1016/S0140-6736(16)31529-X.

**Massot M, Daubié AS, Clermont O, Jauréguy F, Couffignal C, Dahbi G, Mora A, Blanco J, Branger C, Mentré F, Eddi A, Picard B, Denamur E, The Coliville Group. 2016.** Phylogenetic, virulence and antibiotic resistance characteristics of commensal strain populations of *Escherichia coli* from community subjects in the Paris area in 2010 and evolution over 30 years. *Microbiology* **162**:642–650 DOI 10.1099/mic.0.000242.

**McNally A, Alhashash F, Collins M, Alqasim A, Paszckiewicz K, Weston V, Diggle M. 2013.** Genomic analysis of extra-intestinal pathogenic *Escherichia coli* urosepsis. *Clinical Microbiology and Infection* **19**:E328–E334 DOI 10.1111/1469-0691.12202.

**Mellata M. 2013.** Human and avian extraintestinal pathogenic *Escherichia coli*: infections, zoonotic risks, and antibiotic resistance trends. *Foodborne Pathogens and Disease* **10**:916–932 DOI 10.1089/fpd.2013.1533.

**Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D. 2013.** Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* **14**:193–202 DOI 10.1093/bib/bbs012.

**Moremi N, Claus H, Vogel U, Mshana SE. 2017.** Faecal carriage of CTX-M extended-spectrum beta-lactamase-producing *Enterobacteriaceae* among street children dwelling in Mwanza city, Tanzania. *PLOS ONE* **12**:e0184592 DOI 10.1371/journal.pone.0184592.

**Nielsen KL, Stegger M, Godfrey PA, Feldgarden M, Andersen PS, Frimodt-Møller N. 2016.** Adaptation of *Escherichia coli* traversing from the faecal environment to the urinary tract. *International Journal of Medical Microbiology* **306**:595–603 DOI 10.1016/j.ijmm.2016.10.005.

**Oshima K, Toh H, Ogura Y, Sasamoto H, Morita H, Park SH, Ooka T, Iyoda S, Taylor TD, Hayashi T, Itoh K, Hattori M. 2008.** Complete genome sequence and comparative analysis of the wild-type commensal *Escherichia coli* strain SE11 isolated from a healthy adult. *DNA Research* **15**:375–386 DOI 10.1093/dnares/dsn026.

**Pokharel P, Habouria H, Bessaiah H, Dozois CM. 2019.** Serine Protease Autotransporters of the *Enterobacteriaceae* (SPATEs): out and about and chopping it up. *Microorganisms* **7**:594 DOI 10.3390/microorganisms7120594.

**Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel J. 2008.** The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *Journal of Bacteriology* **190**:6881–6893 DOI 10.1128/JB.00619-08.

**Reeves PR, Liu B, Zhou Z, Li D, Guo D, Ren Y, Clabots C, Lan R, Johnson JR, Wang L. 2011.** Rates of mutation and host transmission for an *Escherichia coli* clone over 3 years. *PLOS ONE* **6**:e26907–e26907 DOI 10.1371/journal.pone.0026907.

**Richter TKS, Hazen TH, Lam D, Coles CL, Seidman JC, You Y, Silbergeld EK, Fraser CM, Rasko DA. 2018.** Temporal variability of *Escherichia coli* diversity in the gastrointestinal tracts of Tanzanian children with and without exposure to antibiotics. *mSphere* **3**:e00558-18 DOI 10.1128/mSphere.00558-18.

**Rodríguez-Baño J, Picón E, Gijón P, Hernández JR, Cisneros JM, Peña C, Almela M, Almirante B, Grill F, Colomina J, Molinos S, Oliver A, Fernández-Mazarrasa C, Navarro G, Coloma A, López-Cerero L, Pascual A. 2010.** Risk factors and prognosis of nosocomial bloodstream infections caused by extended-spectrum-beta-lactamase-producing *Escherichia coli*. *Journal of Clinical Microbiology* **48**:1726–1731 DOI 10.1128/JCM.02353-09.

**Russo TA, Johnson JR. 2003.** Medical and economic impact of extraintestinal infections due to *Escherichia coli*: focus on an increasingly important endemic problem. *Microbes and Infection* **5**:449–456 DOI 10.1016/s1286-4579(03)00049-2.

**Schlager TA, Hendley JO, Bell AL, Whittam TS. 2002.** Clonal diversity of *Escherichia coli* colonizing stools and urinary tracts of young girls. *Infection and Immunity* **70**:1225–1229 DOI 10.1128/iai.70.3.1225-1229.2002.

**Sears HJ, Brownlee I, Uchiyama JK. 1950.** Persistence of individual strains of *Escherichia coli* in the intestinal tract of man. *Journal of Bacteriology* **59(2)**:293–301.

**Sears HJ, Brownlee I. 1952.** Further observations on the persistence of individual strains of *Escherichia coli* in the intestinal tract of man. *Journal of Bacteriology* **63(1)**:47–57.

**Shooter RA, Bettleheim KA, Lennox-King SM, O'Farrell S. 1977.** *Escherichia coli* serotypes in the faeces of healthy adults over a period of several months. *Journal of Hygiene* **78**:95–98 DOI 10.1017/s0022172400055972.

**Skurnik D, Bonnet D, Bernède-Bauduin C, Michel R, Guette C, Becker JM, Balaire C, Chau F, Mohler J, Jarlier V, Boutin JP, Moreau B, Guillemot D, Denamur E, Andremont A, Ruimy R. 2008.** Characteristics of human intestinal *Escherichia coli* with changing environments. *Environmental Microbiology* **10**:2132–2137 DOI 10.1111/j.1462-2920.2008.01636.x.

**Stamatakis A. 2006.** RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**:2688–2690 DOI 10.1093/bioinformatics/btl446.

**Stegger M, Leihof RF, Baig S, Sieber RN, Thingholm KR, Marvig RL, Frimodt-Møller N, Nielsen KL. 2020.** A snapshot of diversity: intraclonal variation of *Escherichia coli* clones as commensals and pathogens. *International Journal of Medical Microbiology* **310**:151401–151407 DOI 10.1016/j.ijmm.2020.151401.

**Stoesser N, Sheppard AE, Moore CE, Golubchik T, Parry CM, Nget P, Saroeun M, Day NP, Giess A, Johnson JR, Peto TE, Crook DW, Walker AS, Group MMMI. 2015.** Extensive within-host diversity in fecally carried extended-spectrum-beta-lactamase-producing *Escherichia coli* isolates: implications for transmission analyses. *Journal of Clinical Microbiology* **53**:2122–2131 DOI 10.1128/JCM.00378-15.

**Stoppe NC, Silva JS, Carlos C, Sato MIZ, Saraiva AM, Ottoboni LMM, Torres TT. 2017.** Worldwide phylogenetic group patterns of *Escherichia coli* from commensal human and wastewater treatment plant isolates. *Frontiers in Microbiology* **8**:2512–2532 DOI 10.3389/fmicb.2017.02512.

**Tenaillon O, Skurnik D, Picard B, Denamur E. 2010.** The population genetics of commensal *Escherichia coli*. *Nature Reviews Microbiology* **8**:207–217 DOI 10.1038/nrmicro2298.

**Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguénec C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Médigue C, Rocha EP, Denamur E. 2009.** Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLOS Genetics* **5**:e1000344 DOI 10.1371/journal.pgen.1000344.

**Walk ST, Alm EW, Gordon DM, Ram JL, Toranzos GA, Tiedje JM, Whittam TS. 2009.** Cryptic lineages of the genus *Escherichia*. *Applied and Environmental Microbioogy* **75**:6534–6544 DOI 10.1128/aem.01262-09.

**Wheeler TJ. 2009.** *Large-scale neighbor-joining with NINJA in algorithms in bioinformatics*. Berlin: Springer Berlin Heidelberg.

**Wingett SW, Andrews S. 2018.** FastQ screen: a tool for multi-genome mapping and quality control. *F1000Research* **7**:1338–1350 DOI 10.12688/f1000research.15931.2.

**Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MCJ, Ochman H, Achtman M. 2006.** Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Molecular Microbiology* **60**:1136–1151 DOI 10.1111/j.1365-2958.2006.05172.x.

**Wold AE, Caugant DA, Lidin-Janson G, De Man P, Svanborg C. 1992.** Resident colonic *Escherichia coli* strains frequently display uropathogenic characteristics. *Journal of Infectious Diseases* **165**:46–52 DOI 10.1093/infdis/165.1.46.

**Zhou Z, Alikhan NF, Mohamed K, Fan Y, Achtman M, Group AS. 2020.** The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Research* **30**:138–152 DOI 10.1101/gr.251678.119.