Chapter 1: Introduction

1.1. The history, taxonomy, and characteristics of *Escherichia coli*

1.1.1. Introduction

Escherichia coli (*E. coli*) is a bacterial genus and species, belonging to the family *Enterobacteriaceae*. Originally called "Bacterium coli commune" in 1885, it was subsequently renamed *E. coli* after Theodor Escherich, a German paediatric physician, who investigated and characterised the bacteria colonising infant guts, as well as their role in infant digestion [1].

We now know that *E. coli* is a ubiquitous member of microbiota in the lower intestinal tract of reptiles, humans, as well as most other warm-blooded animals [2]. A typical human gut can carry up to 10^9 *E. coli* cells [3]. This compares to the 3.0 x 10^{13} human cells making up the body and the estimated 3.8 x 10^{13} number of bacterial cells that colonise the human body [4]. In the context of the digestive tract, *E. coli* represents about 0.1 - 5% of the gut microbiota; although it is ubiquitous, it represents only a minor component of the gut bacterial population [5, 6]. Since *E. coli* grows rapidly, has relatively simple nutritional requirements, and is highly genetically amenable, it quickly became the main Gram-negative prokaryotic model system [2].

E. coli is not only a benign resident of the human gut, but can also contribute positively to human health. For example, *E. coli* produces Vitamin K and serves as a barrier layer in the gut, helping to outcompete and to outgrow pathogenic bacteria which attempt to colonise that niche [7]. *E. coli* is also known to inhabit niches outside of the warm-blooded gut, including water, sediments and soils [8, 9], it is not restricted to colonising humans or warm-blooded animals. Moreover, along with this ability to colonise a diverse set of niches, *E. coli* is also a diverse organism genetically. The ability of *E. coli* to survive and to colonise many niches, both biological and abiotic environments, as well as to behave as both a commensal and pathogenic bacterium, makes this species an excellent candidate for the study of bacterial evolution and adaptation [10].

1.1.2. Commensal *E. coli*

The population structure of commensal *E. coli* has been reviewed extensively by Tenaillon *et al* [3]. Along with others gut microbiome, *E. coli* is one of the first bacterial species to colonized infant's gastrointestinal tract after natural birth delivery and might originate from the maternal faecal flora [11]. The population of *E. coli* in the human gut is diverse: at any given time point, each person carries one predominant *E. coli* strain/genotypes which accounts for more than 50% of the total *E. coli* in the gut, and simultaneously harbours several other strains/genotypes at different percentages [12]. Longitudinal sampling of stool isolates revealed that there are frequently more than two genotypes of *E. coli* colonising the gut, defined as resident (stay for years) or transient colonisers (only present for a few

days or weeks). *E. coli* are classified into four major phylogroups A, B1, B2 and D [13] and populations of these phylogroups are spatiotemporally diverse. Different human populations are dominated by different phylogroups, and dominant phylogroups in a population may vary over time [14]. For example, the guts of native people in French Guyana are composed of more "commensal*"* *E. coli* phylogroups A and D than people from an urban city like Paris [15]. There was also a shift from phylogroup A to B2 in Parisians between the 1980s and 2000s. Such differences are explained mainly by diet and hygiene [16, 17].

1.1.3. The *E. coli* genome

Bacterial genomes can be sub-divided into core and accessory regions. Core genes are families of genes that are found in all members of a particular species, whereas accessory genes are present in some, but not all, genomes of that species. Together, the core and accessory regions of genes constitute the pan genome. From comparative genomic studies that have analysed the core and pan genomes of *E. coli*, we know that the core genome of *E. coli* consists of nearly 2,000 genes. However, the size of the predicted core depends heavily on the number of genomes and the cutoffs used to define orthologous genes [18, 19].

The pan genome defined by comparative genomic analysis of the first 20 fully sequenced *E. coli* genomes, which represent all *E. coli* pathotypes, contains around 18,000 genes [18]. The number of genes in the *E. coli* pan genome increases linearly with the number of newly added *E. coli* genomes [20]. Compared to more host restricted prokaryotes such as *Chlamydia trachomatis* or *Corynebacterium pseudotuberculosis*, *E. coli is* thought to have an "open" pan genome, whereby accessory genes might provide significant benefits to the *E. coli* host such as substrate utilisation and stress tolerance in harsh environments [21]. A recent analysis of 2,085 *E. coli* genomes estimate that there are 3,188 core gene families across *Escherichia coli* species and approximately 89,000 unique gene families [22, 23].

1.1.4. Pathogenic *E. coli*

Despite being part of the normal microbiota of healthy individuals and a useful laboratory workhorse, variants of *E. coli* are also known to be pathogens which are estimated to cause million infections *per* year [24, 25].

E. coli is one of the four leading causes of moderate to severe diarrheal disease in children under five years of age [26, 27]. Globally, it is the main cause of urinary tract infections (UTIs) and blood stream infections (BSIs) in the elderly, and of meningitis in neonates [28]. These pathogenic groups can be further subdivided into intestinal pathogenic *E. coli* (InPEC) and extraintestinal pathogenic *E. coli* (ExPEC). Of the InPEC, there are five major groups, known as pathotypes, defined based on the specific disease with which they are associated: enteroaggregative *E. coli* (EAEC), enterohaemorrhagic *E. coli* (EHEC), enteropathogenic *E. coli* (EPEC), enterotoxigenic *E. coli* (ETEC), enteroinvasive *E. coli* (EIEC) and diffusively adherent *E. coli* (DAEC). Four main pathotypes exist for ExPEC: uropathogenic

E. coli (UPEC), neonatal meningitis *E. coli* (NMEC), avian-pathogenic *E. coli* (APEC) and sepsis-associated *E. coli* (SePEC). The intestinal *E. coli* pathotype is based on the clinical manifestation of the disease and the virulence genes which they carry [29]. EPEC harbours loci of enterocyte effacement (LEE) and bundle-forming pilus gene (*bfp*), whereas EHEC encodes LEE together with genes encoding the Shiga toxin (*stx1*, *stx2*, or both together). ETEC expresses the LT and ST enterotoxins. EIEC possess *ipaH* genes located on pINV plasmids that enable this pathotype to invade host cells. The genetic determinants of the APEC pathotype remain unclear [30].

Although there are many different *E. coli* pathotypes, the subject of this study are those pathotypes associated with bloodstream infections, sepsis and bacteraemia. These pathotypes will now be considered in more detail.

1.2. *E. coli* associated with bloodstream infections (BSIs), sepsis and bacteraemia

Prior to the year 2000, it is estimated that 10 million people were affected every year by *E. coli* BSIs, with mortality rates of 44 deaths per 100,000 individuals [31]. *E. coli* bacteraemia is defined by the presence of *E. coli* in the blood, whereas sepsis is defined as the host systemic inflammatory response to the *E. coli* bloodstream infection, which leads to multiple organ failure [32]. *E. coli* can enter the blood as a consequence of severe primary infections such as pneumonia or urinary infection [33]. Other routes of infection include the direct inoculation of the bacteria into the blood *via* such routes as chronic urinary catheterisation, mechanical ventilation or prostate biopsy. Bacteraemia can also emerge from the gut through a process called bacterial translocation, whereby bacteria or bacterial products are translocated from the gut to mesenteric lymph nodes, from where they disseminate and cause disease systematically [34, 35]. In leukaemia patients, the bowel appears to be a portal for *E. coli* translocation and repeated bacteraemia episodes, whereby the same genotype has been isolated from blood and stool [36]. Data from mice which have undergone chemotherapy, or from mice bearing tumours, confirm this bacterial translocation hypothesis, although the mechanism underlying this phenomenon remains unknown [37].

In developed countries such as the United States of America, Canada, and the United Kingdom, the annual incidence of invasive *E. coli* bloodstream infections ranges between 30 [38] to 63 cases *per* 100,000 [39]. This increases with age, especially in people over 45 years old, with incidence in this senior population (>45 years old) being up to 150 cases *per* 100,000 [40]. There has also been a general increase in the number of reported of *E. coli* bacteraemia cases [41] in the 21st century; for instance, from 1999 - 2011 and after 2006, the number of reported *E. coli* bacteraemia cases in UK increased from 3.4 to 5.7 *per* 10,000 bedstays [42].

Unfortunately, there are limited data available regarding the incidence and risk factors of *E. coli* bacteremia in developing world. In developing countries within Africa and South Asia, bacteraemia is commonly associated with HIV-positive patients who are co-infected with invasive non-typhoid

Salmonella or *Salmonella* Typhi infections. However, among non-*Salmonella* enterobacterial infections, *E. coli* is still the leading cause of BSIs both in adult and children [43-45] affecting both HIV and non-HIV patients equally [46]. In Cambodia, surveys carried out from 2007 to 2010 showed that *E. coli* accounted for nearly 30% of BSIs cases [47]. In Vietnam, retrospective blood culture data from 1994 to 2008 showed the decline of *S.* Typhi and *S.* Paratyphi A, but an increase in the number of fungi, non-typhoid *Salmonella* and *E. coli* infections that was concurrent with the HIV epidemic [48]. *E. coli* and *Klebsiella pneumoniae* were recently reported as the two most common cause of community and hospital acquired blood stream infection in the North of Vietnam [49]. A recent sepsis study in Southeast Asia encompassing 13 public hospitals in Indonesia, Thailand and Vietnam showed that *E. coli* is one of the most common cause of sepsis in this region [45].

It is hypothesised that those *E. coli* bacteraemia cases which are not related to invasive medical procedures often begin as infections of the urinary tract, gastrointestinal tract, or the lung, and subsequently progress into an invasive blood stream infection [50]. The general risk factors for progressing to a blood stream infection include dialysis, solid organ transplantation and neoplastic disease [38]. In the adult population, aging is an important factor that modulates the risk of developing bacteraemia. Other comorbid medical conditions that alter the immune system, such as infection with HIV, diabetes, cancer, and cirrhosis, also increase the risk of nosocomial bacteraemia [51]. Risk factors for *E. coli* colonisation and infection in neonates include premature birth, low birth weight, prolonged mechanical ventilation, length of hospital stay and antibiotic usage [52].

1.3. Hospital for Tropical Disease (HTD), Vietnam

The retrospective study described in this thesis was conducted at Hospital for Tropical Diseases, a 550-bed tertiary referral hospital for infectious disease in Vietnam. The hospital consists of seven wards, including one intensive care unit (ICU) that treats both adult and children patients and admits 38,000 patients annually. The hospital mainly serves patients resident in Ho Chi Minh City and the southern part of Vietnam, with a catchment population of 42 million people [53]. Patients without infectious diseases, such as cancer, tuberculosis, and non-HIV immunosuppression are referred to other hospitals.

1.4. Antimicrobial resistance in *E. coli*

An antibiotic is defined as a substance able to inhibit the growth of or kill a microbe [54]. Since the discovery of the first antibiotic, penicillin, in 1927 and its initial widespread use as a therapeutic agent in the 1940s, a wide range of antibiotic compounds have saved millions of lives and helped to control infectious disease, as well as to prevent infections during invasive procedures such as surgery and routine biopsies. Whilst the term ‘antibiotic’ was originally used to describe naturally occurring compounds, the term used now is ‘antimicrobial’, in order to encompass semi-synthetic antibiotics [55].

Since the first use of antimicrobials, the phenomenon of bacteria becoming non-susceptible or resistant to these compounds has been documented [54]. Bacterial antimicrobial resistance (AMR) describes the

process by which a micro-organism shows a reduced or non-susceptible phenotype to an antimicrobial compound that previously inhibited or killed it. This is achieved through multiple different strategies such as drug inactivation or modification, alteration of the drug target site, alteration of the metabolic pathway affected by the drug, and reduction of the rate of drug accumulation within the cell *via* reduced uptake or active efflux [56]. Bacterial AMR phenotypes can be due to intrinsic or acquired mechanisms of resistance. Intrinsic resistance is the innate ability of a bacterium to be resistant to antimicrobial compounds due to natural structure or functional characteristics, thereby allowing bacteria to tolerate particular drugs. For example, *Klebsiella* species are naturally resistant to penicillin since they carry core chromosomal genes that encode enzymes which degrade ampicillin as it enters the cell. Acquired resistance however is when drug sensitive bacteria become resistant through point mutations, by acquiring genes from other bacteria through horizontal gene transfer or a combination of these two mechanisms. An example of acquired resistance through the generation of point mutations include mutations within genes encoding the drug target site – these include point mutations in chromosomal genes such as DNA gyrase or the 23S rRNA genes which confer resistance to fluoroquinolones and aminoglycosides, respectively. Examples of acquired resistance through horizontal gene transfer of mobile genetics elements (MGEs) from another bacterium include plasmids that encode genes that confer resistance to carbapenems and most other classes of antimicrobial [56](see Section 1.6).

Although point mutations in chromosomal genes are vertically inherited, and their dissemination throughout a bacterial population requires the spread and expansion of a resistant bacterial clone, MGEs are particularly able to disseminate rapidly and broadly, not only between members of the same species, but also more broadly within bacterial families. For example, *mcr*-1, the gene that confers colistin resistance, was first identified in *E. coli*, but is now found in *Salmonella*, *Klebsiella*, *Enterobacter* [57] and has also been shown to be capable of transfer outside of the *Enterobacteriaceae* into *Pseudomonas aeruginosa* in the laboratory [58].

1.5. AMR *E. coli* in Vietnam

For critically ill patients suffering BSIs, especially in patients in intensive care units, oral or intravenous antibiotics are the only interventions suitable to combat these infections. However, modern treatment options for *E. coli* are very limited, due to the high level of AMR in this species. Multi drug resistant *E. coli* (MDR) is defined as bacteria with reduced susceptibility to at least one antimicrobial agent in three or more antimicrobial classes. In the Hospital for Tropical Diseases (HTD), Vietnam, according to the hospital treatment guideline, patients are treated empirically with fluoroquinolone or third generation cephalosporin if they have a suspected BSI. The reality in HTD, as well as in many other hospital settings, is that blood culture results and antimicrobial susceptibility test data are generally not available until 48 hours after testing. If a patient fails to respond to these antimicrobial treatments because they are infected with a MDR *E. coli*, for example, it is sometimes too late to switch to another antibiotic therapy regime [59]. Such patients commonly develop septic shock leading to a poor

prognosis [60]. Furthermore, successful treatment of patients infected with MDR *E. coli* increases treatment costs because patients need to be treated with newer classes of antibiotic. Infection with MDR *E. coli* BSIs also often recur, and significantly increase length of hospital stay [61]. It is therefore important to identify the prevalence of bacterial infections that show AMR to guide better treatment and patient management. Currently in HTD, all patients with ESBL-producing *E. coli* will be treated with carbapenem or with piperacillin-tazobactam as alternative therapy (personal communication).

The increasing of MDR *E. coli* is not restricted to HTD in Vietnam. The same trend has been observed in other hospitals in low to middle income countries [62, 63]. Indeed, *E. coli* has now been listed by the World Health Organization (WHO) as one of twelve bacteria that pose the greatest threat to human health, highlighting the urgent development of new antibiotic therapies [64]. Since 3rd – generation cephalosporin was used as an empirical treatment for this study we will only focus on the prevalence of ESBL – producing *E. coli* collected between 2010 and 2015 from patients attending HTD in Vietnam.

1.5.1 ESBL – producing *E. coli* and carbepenemase- producing *E. coli*

Extended – spectrum β-lactam (ESBL) is a term that refers to a class of enzyme capable of hydrolysing β-lactam antibiotics. β-lactam antibiotics are a class of broad spectrum antibiotics containing a β-lactam ring, including penicillins, cephalosporins, monobactam and carbapenems. These antibiotics bind to bacterial penicillin binding proteins, making the bacterium unable to synthesise cell walls correctly, leading to a failure of the bacterial cell to divide. Third generation cephalosporins have a broad spectrum of activity against all Gram-negative species; these are favoured by clinicians for use in treating a variety of infections [65, 66]. β-lactamase enzymes, however, can be inhibited by inhibitors such as clavulanic acid or tazobactam.

Prior to the 1990s, class A β-lactamase enzymes, including TEM and SHV enzymes, were typically responsible for conferring an ESBL-producing phenotype upon the *Enterobacteriaceae*. However, the epidemiology of ESBL-producing *E. coli* changed dramatically since the acquisition of CTX-M genes by this species. These genes have been proposed to have originated from *Kluyvera* spp., and are encoded mainly on plasmids [67]. The genetic context of CTX-M-producing *E. coli* is complicated; CTX-M genes are often co-inherited with other resistance genes as part of Insertion Sequences (IS), fragments of DNA capable of being mobilised around bacterial chromosomes and plasmids. IS elements such as IS*Ecp1*, IS*CR1* and IS*26* often coexist with $bla_{\text{CTX-M}}$ [68].

Currently, over 150 $bla_{\text{CTX-M}}$ genes have been described and they are classified into five groups based on their amino acid sequence similarity: CTX-M-1, 2, 8, 9 and 25. CTX-M proteins share >94% amino acid identity within each group and <= 90% similarity between various groups.

1.5.2 Carbapenemase-producing *E. coli*

Carbapenems are also members of the β-lactam class of antibiotics, and include imipenem, ertapenem, meropenem and doripenem. Carbapenems are often considered to be "antibiotics of last resort". Despite this, resistance to carbapenems conferred by carbapenemase enzymes is spreading worldwide [69]. Different type of carbapenemases include $bla_{NDM,}$ bla_{KPC}, bla_{OXA}, bla_{IMP} genes. However, since the majority of *E. coli* remain sensitive to this drug within our hospital, we will not focus on these genes in detail in this thesis.

1.6. Mobile genetic elements (MGEs)

Mobile genetic elements (MGEs) can mobilise resistance genes horizontally within a bacterial population

Horizontal gene transfer (HGT) is the process by which genetic material can be transferred between different strains of a bacterial species (intra-transfer), or between closely related bacteria of different species (inter-transfer). Unlike naturally occurring mutations, which take time to accumulate in a genome and to accumulate in a population during repeated cellular division, bacteria can potentially acquire multiple resistance genes, and therefore phenotypes, in a single acquisition event. The dissemination of drug resistance within a population of bacteria can be potentiated by HGT, such that resistance genes can be spread both laterally and vertically throughout a population.

There are three mechanisms for HGT - transformation, conjugation and transduction.

Transformation is the process by which competent bacteria take up exogenous DNA and thereby incorporate novel genetic material into the genome. This may take place under stressful conditions, or under conditions that affect bacterial growth. In the environment, extracellular DNA is constantly excreted from live bacteria and is released from dead bacteria, serving as a pool of genetic information. Naked chromosomal DNA or plasmids released into soil and water can remain intact for hours or days, whereas DNA that is released into the blood by invasive bacteria will be degraded within several minutes by DNAse enzymes that are components of human serum. However, this relatively short time is sufficient for transformation to occur.

Conjugation is the transfer of genetic material, often plasmids or transposons, through direct contact between bacteria cells using a pili as a bridge. Each plasmid has their own replication system, those with the same replication system cannot coexist stably within the same cell, and are considered to be "incompatible". This incompatibility led to the development of the incompatibility typing scheme (commonly called Inc-typing) for plasmids [70]. Plasmids encode their own replication machinery, separate to that of the chromosome, and also can be maintained vertically using toxin-antitoxin systems [70]. Transposons which contain IS elements and can be mobilised *via* self-encoded transposase enzymes, however, can be incorporated into the host chromosome or plasmids through transposition.

Transduction is the process by which bacteria acquire new DNA upon infection by bacteriophages. Bacteriophages are viruses that infect bacteria and incorporate themselves into the host chromosome as a prophage during the lysogenic stage of their life cycle. A prophage can switch from a lysogenic state to a virulent lytic state, and in doing so may co-package host DNA together with the phage genome. Upon cell lysis, this host DNA can be transduced to a new host cell upon infection by a viral particle, and thus facilitate gene transfer between different bacteria [71].

1.7. Reservoirs of *E. coli* and transmission mechanism

The ability of *E. coli* to colonise and survive in different hosts and environments makes it difficult to identify the source of *E. coli* infections. In the past, *E. coli* infections such as those caused by ExPEC have been described as sporadic infections caused by bacteria that originate from the host's gastrointestinal (GI) tract [72]. In addition, *E. coli* is able to colonise the GI tract of warm blooded animals; therefore, many investigations sought the origin of transmission through the food chain and farm animals. Unfortunately, these studies used low resolution genotyping methods such as pulse-field gel electrophoresis (PFGE), multi locus variable - number tandom repeat analysis (MLVA) and multi locus sequence typing (MLST) [73, 74]. Although detection of the same MLST genotype and same AMR phenotype can occur between animal and human isolates, the conclusions drawn from these studies are potentially misleading and do not provide enough resolution to infer *E. coli* transmission. For example, in a study by Been and colleagues, they show that at the whole genome sequence level, isolates from chicken and humans were not identical and they actually share the same mobile genetic element (plasmid) rather than representing the spill over of same clonal strains between two populations [75].

A recent study in Sweden also supports the notion that the clonal spread of *E. coli* from animals to human was unlikely. This study by Börjesson *et al.* on ESBL-producing *E. coli* from nearly 5,000 samples collected from food, farm animals, healthy humans and BSI patients found that there was limited clonal spread of these *E. coli* strains from animals and food to humans, but *E. coli* from poultry can serve as a reservoir for drug resistance genes and for plasmids that can move to *E. coli* types commonly found in humans [76]. Nevertheless, whether *E. coli* from animals still represents a source for human infections requires further investigation even if it remains controversial [77]. Outside human and animal hosts, *E. coli*, especially strains within phylogroup B2 and D found in coastal sediments, could represent a potential reservoir for commensal and pathogenic *E. coli* [78].

1.8. Aims and objectives of the study

There are no data available regarding the genetics of those *E. coli* that cause bloodstream infections (BSIs) in the Vietnamese population. We aimed to sequence 690 bacterial samples to obtain a high-resolution snapshot of the landscape of *E. coli* causing BSIs. We also screened for virulence genes and AMR genes to determine if these non-core genes are associated with invasive disease.

Research questions

1. Is there a specific lineage or sequence type that is circulating and causing invasive disease in Vietnamese patients, or disease outcome dependent upon host factors alone, such that all *E. coli* are capable of causing invasive disease? (*i.e.*, is there any specific lineage associated with disease outcome?)
2. How is the gene content of invasive *E. coli* different to that of carriage *E. coli* isolates? Are there significant genetic markers that differentiate these invasive bacteria such as AMR genes, colonisation factors, toxins, siderophores, capsule type, *etc*.?