Chapter 2: Comparative Genomics of BSIs caused by *E. coli* in the Hospital for Tropical Diseases in Vietnam

2.1. Introduction

Hospital acquired infections (HAIs) are a major threat to patient safety, which in locations with poor surveillance and infection control can lead to high rates of infection and significant mortality. The problem of HAIs is further exacerbated by the emergence of antimicrobial resistance, particularly extended spectrum β-lactamases (ESBLs) and carbapenemases in Gram-negative bacteria.

Healthcare settings across Asia are under increasing pressure to control the spread of multidrug-resistant (MDR) HAIs and to reduce the overall levels of MDR in bacteria. Recent work has specifically highlighted nosocomial bloodstream infections (BSIs) as a major cause of mortality in high dependency units. Notably, two outbreaks of *Klebsiella pneumoniae* in 2012 occurred in the neonatal intensive care unit (NICU) in a tertiary hospital in Nepal, which had a mortality rate of 75%. This study found, using whole genome sequencing, that there were regular peaks of BSIs across the hospital, associated with MDR HAIs that recurred even after deep cleaning of the hospital. This upsurge in MDR BSIs with high mortality is a growing trend echoed in other healthcare settings across Asia [79]. These peaks in infection are driven by an increase in the rate of isolation (from blood) of bacteria such as *Escherichia coli* that are resistant to most available antimicrobials. Currently, in 2015, 24% of all BSIs in Hospital for Tropical Disease (HTD), Ho Chi Minh city, Vietnam are attributable to *E. coli* (unpublished data). Understanding the nature of HAIs across Asia is required to enhance infection control measures, to optimise therapeutic approaches, to minimise the further emergence of antimicrobial resistance, and to reduce morbidity and mortality due to BSIs caused by MDR HAIs.

The hypothesis:

We hypothesise that bacterial isolates causing BSIs are more closely related to one other than to isolates taken from the gut of patients attending our hospital. This is because many BSI *E. coli* are hospital acquired isolates which are likely to be members of a restricted number of genetic lineages, and to carry genetic loci (such as genes encoding antimicrobial resistance and siderophores) which are more strongly associated with an invasive phenotype in comparison to non-invasive phenotypes, including UTIs and respiratory infections.

Specific Aim:

The aim of this chapter was to determine the genetic epidemiology of *E. coli* causing BSIs in the Hospital for Tropical Diseases (HTD) in Ho Chi Minh city, Vietnam. To address this hypothesis, we set out to compare the genomes of *E. coli* associated with BSIs to those of carriage *E. coli* isolates taken from patients attending this hospital. For a limited number of samples, it was possible to collect bacterial

isolates from both the rectal swab samples and bloodstream infections of the same patient for comparison.

2.2. Methods

2.2.1. Strain collection

Our strain collection consisted of 690 *E. coli* isolates derived from two different studies called 15EN and 06EI. Of these 496 represented a longitudinal collection of isolates that were cultured from patients with BSIs admitted to Hospital for Tropical Diseases (HTD) in Vietnam between 2010 and 2014 (15EN). These isolates were selected randomly from the approximately 100 isolates collected per year during this period. Also included were 94 randomly selected *E. coli* isolates taken from rectal swabs from patients included in the 06EI study looking for HAIs in the intensive care unit (ICU) in HTD during 2015. All of the above randomly selected samples were collected from patients that had been sampled on or after admission. All isolates are described in detail in Supplementary Tables S1 and S2.

The remaining 100 isolates included in this study were derived from 31 patients with matched rectal swab (69) and BSI (31) isolates. These patients were enrolled in an extension of the 15EN study during 2015 at HTD. These matched samples are included in the general analysis below but will also be described in more detail in Chapter 4. All rectal isolates were considered to be carriage isolates as they were not linked to any clinically recorded signs of enteric disease.

2.2.2 Ethics

Both studies were approved by the ethical committee at Hospital for Tropical Disease (HTD) and by the University of Oxford Tropical Research Ethics Committee (OxTREC).

2.2.3 Bacterial isolation

All microbiological culture work was performed in laboratories at HTD and OUCRU. Routine care at HTD dictates that patients admitted with febrile illness and with a suspected bacteraemia will have their blood cultured as part of routine diagnosis of the hospital. Briefly, for these samples, 5 mL of patient's blood was incubated at 37 ˚C in a BACTEC bottle. Positive BACTEC bottles were then sub-cultured onto selective MacConkey agar (MC) for further microbiology identification either by matrix assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI TOF, Bruker) or using API20E strips (BioMerieux, France) following the manufacturer's guidelines. Commensal and carriage *E. coli* were cultured from rectal swabs of patients participating in the 06EI study. Swabs were taken twice weekly from admission day until discharge. Microbiology identification procedure was as for blood isolates, except that for each rectal swab, if *E. coli* colonies with different morphologies or AMR phenotypes (lactose fermenting, non-lactose fermenting and β-haemolytic) were identified, multiple colonies were collected. The number of isolates collected per patient ranged from one to four isolates,

except for one patient from which a total of 20 isolates were collected during a long-term hospitalisation (described in detail in Chapter 4).

All isolates included in this study were tested for their antimicrobial susceptibility phenotype using 13 antimicrobials, also described in detail in Chapter 3.

2.2.4. Metadata

Patient demographic data were extracted from electronic hospital records in both studies. This included information about patients including age and gender (see Supplementary Tables S1). We inferred patient comorbidities, such as their HIV status or chronic hepatitis/liver cirrhosis, based either on the ward into which the patient was admitted, or from initial diagnoses from doctors when blood culture requests were submitted. These samples were denoted "HIV/liver disease" samples. Length of hospital stay was calculated by subtracting the date of discharge from the date of admission. The clinical outcomes for all patients included in this study were available. However, data on recent hospital admission and antibiotic treatment were only available for isolates that were collected in the 06EI study.

In an effort to look for the origin of infection, or coinfection at different bodily sites, the patient identifiers for all patients for whom we had associated blood-borne isolates were used to screen the urinary and peritoneal fluid (ascites) positive patient database to determine whether there were also corresponding records of *E. coli* isolates cultured from the liver abscesses or urinary tract infections from that patient too. However, we were only able to collect information about antimicrobial susceptibility phenotype for these *E. coli* isolates. Unfortunately, we could not include them in this study for sequencing because the samples were not being kept.

We define community acquired (CA) infections as being patients who test positive for *E. coli* from blood cultures collected during the first 48 hours of the patients admission to HTD and define hospital acquired (HA) infections as isolates cultured from blood collected at least 48 hours after admission.

2.2.5. Whole genome sequencing

Genomic DNA was extracted in OUCRU using Wizard Genomic DNA Purification kit (Promega, USA) and sent to the Sanger Institute. Approximately 2 μg of DNA was subjected to sequencing using Illumina HiSeq 2500 platform to generate 125 bp paired end reads.

2.2.6. Accessions

Short read sequence data from this study are available in the European Read Archive (ERA) under project accession ERP017161. All samples and their accession numbers are detailed in Supplementary Table S1 and S2.

2.2.7. Genome assembly and annotation

Annotated assemblies were produced using the pipeline described in detail in Page *et al* [80]. For each

sample, sequence reads were used to create multiple assemblies using VelvetOptimiser v2.2.5 [81] and Velvet v1.2 [82]. An assembly improvement step was applied to the assembly with the best N_{50}, and SSPACE was used to scaffold the contigs [83]. Sequence gaps were filled using GapFiller [84]. Automated annotation was performed using PROKKA v1.11 [85] and *Escherichia* databases from RefSeq [86].

2.2.8. Quality control of sequencing data

In order to confirm the quality of the sequenced bacterial genomes, a range of different tools were used to determine the quality of the short-read Illumina sequence data. Short reads data quality score was accessed using FastQC v0.11.4 [87]. FastQC compiles sequence quality into one report which details basic statistics of all reads in each sample, such as GC content percentage (GC%), total number of reads, and sequence quality score. All reports were aggregated using MultiQC [88] for a complete output and was manually assessed. Samples showing a GC% different from that expected for *E. coli* (49-52%) were excluded from the analysis. Sequence quality was further checked by mapping the reads to a reference genome *E. coli* K-12 strain MG1655 (EMBL accession number U00096). Samples were excluded if number of contigs was higher than 300 or the total reference genome coverage was less than 80%.

Kraken was used to assign taxonomic labels to short DNA sequences to confirm the genus and species of each sample [89]. CheckM was also used to test for genome heterogeneity and completeness [90].

2.2.9. *In silico* MLST and phylogroup classification

Multilocus sequence typing (MLST) is a traditional typing method that uses the combined sequence of seven housekeeping genes in the genome to assign the bacteria to a sequence type (ST). STs were determined from WGS data using MLST check [91], which compares the assembled genome against the *E. coli* MLST database (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli) to assign a specific allelic profile.

Originally *E. coli* was classified into phylogroups A, B1, B2 and D: phylogroups were defined as group of isolates representing distinct clades in a phylogeny built from the ECOR collection [92]. Different phylogroups were believed to have differences in gene content that reflected different niches and lifestyle [93]. *E. coli* phylogroup of each isolate was determined by blasting *de novo* assemblies against three marker genes, *chuA*, *yjaA* and *tspE4.C2* proposed by Clement *et al* [94]. Those from our collection that could not be classified into these four phylogroups were label as “other phylogroups”.

2.2.10. Construction of core genome phylogeny

The core and pan genome of *Escherichia coli* was determined using ROARY [95], in which core genes are defined as those which produce products that share a minimum BLASTp percentage identity of 95% and are present in at least 95% of all isolates. The sequences of all core genes were concatenated into a multiple sequence alignment file. We also included 25 reference genomes from annotated reference

strains from different phylogroups, plus *E. fergusonii* and *E. albertii* (see Supplementary Table S3).

Approximate maximum likelihood phylogenetic trees, based on core gene alignments, were constructed using FastTree v2.1 [96]. SNP alignments of core genes was used to subdivide the population of the sequenced *E. coli* isolates by hierarchical clustering using hierBAPS [97]. Phylogenetic trees and associated metadata were visualized in iTOL [98] or Phandango [99].

2.3. Results

2.3.1. Sequencing and QC

DNA was prepared for whole genome sequencing from a total of 690 isolates (see Methods). Following library preparation and quality control steps, which took into account the combined results from FastQC, Kraken, CheckM, genome size as well as number of contigs (see Methods), 665 were taken forward for further analysis (see Figure 2.1). This included 506 genomes from isolates cultured from blood and 159 from isolates cultured from rectal swabs (described below; Supplementary Table S1 and S2).

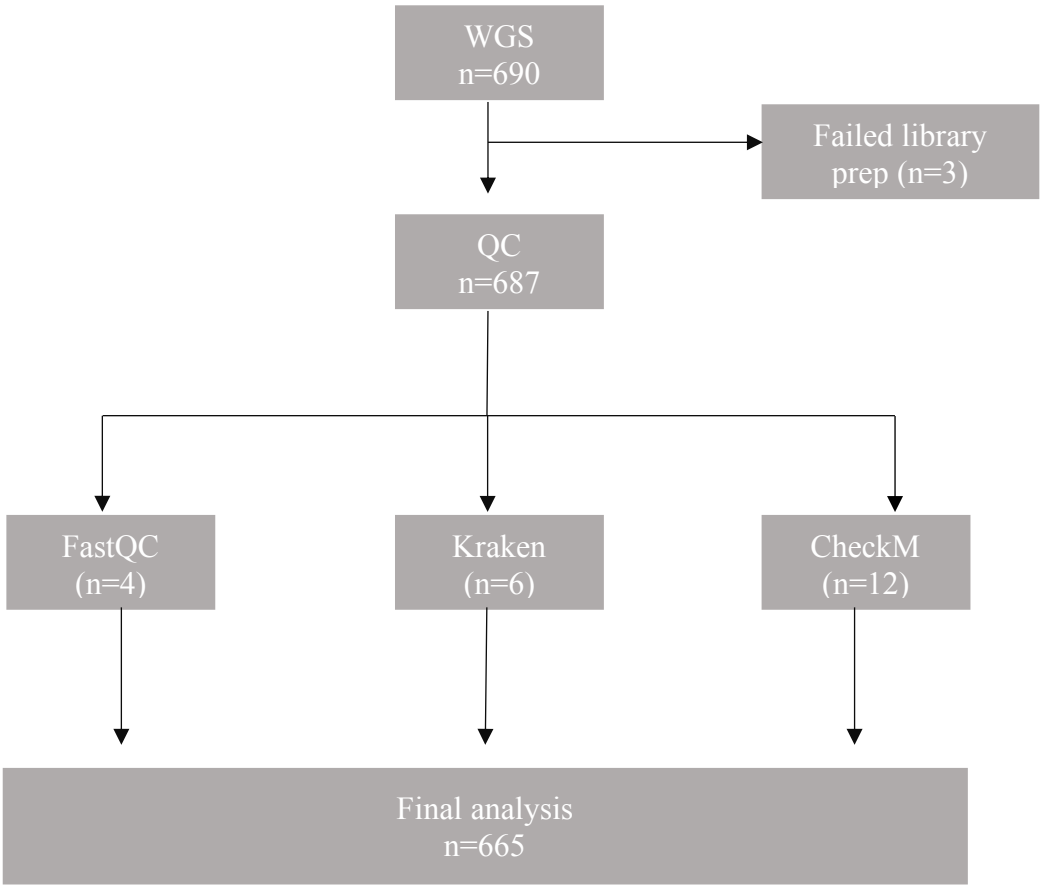

Figure 2.1. Sequencing and QC work flow for the samples included for whole genome sequencing (WGS) in this study. Indicated are the number of samples that passed through each stage of sequencing and QC. The numbers in brackets indicate the number of samples that failed at each sequencing stage or QC step.

2.3.2. Basic patient demographic and clinical outcome of hospitalised patients with BSIs.

The patient characteristics linked to the 506 *E. coli* genomes derived from BSIs are summarised in Table 2.1. Looking at the demographic data, 54% of all *E. coli* BSI patients were female and the median age of all patients with BSIs was 53 (interquartile range (IQR) 38 – 65 years old). Based on the ward into which the patient was admitted and/or the hospital records, immunocompromised patients infected with HIV accounted for 63/506 (12%) of all BSI patients while 109/506 (22%) BSI isolates were derived from patients with chronic hepatitis B/C or cirrhosis. If age was taken into account then for patients under 40 years of age, 37% of patients with BSI were also HIV positive, while in the BSI patients between 40 and 60 years of age, 7% of them were HIV positive and 32% had liver disease (Figure 2.2).

Table 2.1: Clinical characteristic of *E. coli* BSIs patients

Patient characteristic		**n=506**	**% - Median(IQR)**
Gender	female	270	53.36
	male	236	46.64
Age		53	38-65
Acquisition	CA (<48h)	445	87.94
	HA(>48h)	57	11.26
	unknown	4	0.79
Outcome	recover	368	72.8
	worsen*	56	11.1
	no change*	69	13.7
	death*	12	2.4
Hospital length		11	8-15
Comorbidity	Liver disease	109	21.54
	HIV	63	12.45

* “Death” means patients died at HTD. “worsen” means clinical symptom progress toward a severe infection manifesting as increased breathing rate, increased white cell counts, respiratory failure, coma. This is likely to equate to death since Vietnamese traditional practice is to be discharge patients home to die. “No change" means the patients clinical condition did not improved after treatment or they were transfered to another hospital and therefore lost to follow up.

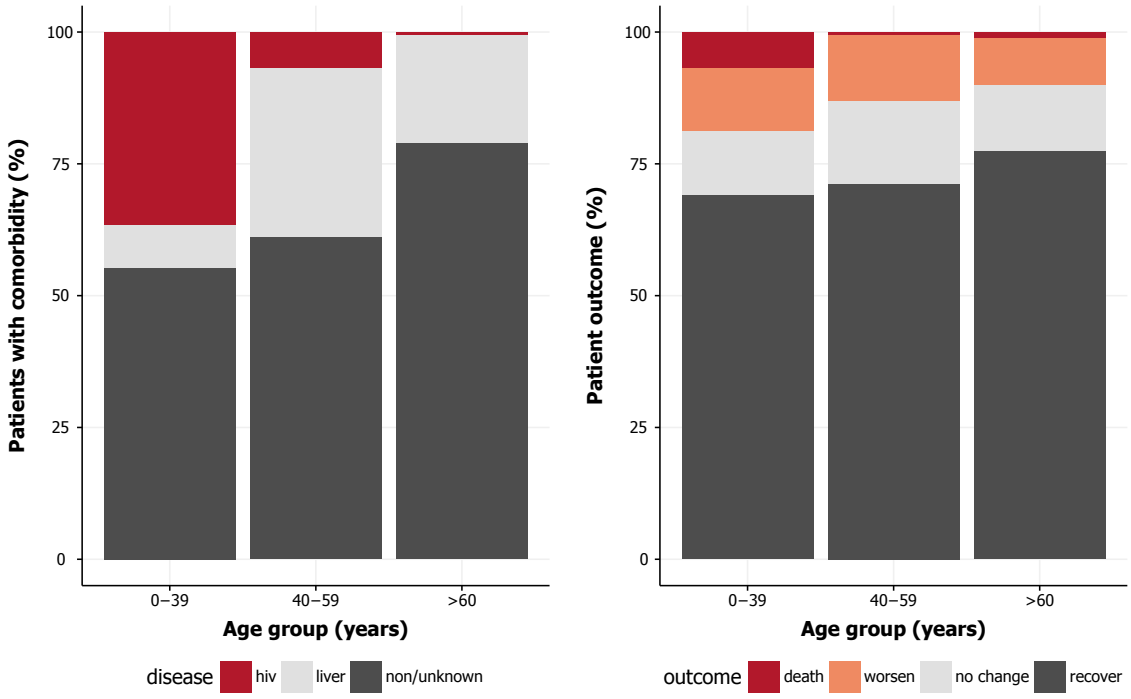

Figure 2.2: Comorbidity and outcome in patients with *E. coli* BSIs. Other comorbidity but not prevalent including diabetes, severe pneumonia, severe Dengue fever, cholecystitis and meningitis.

Overall, 368/506 (72.8%) patient conditions were improved or recovered when discharged. 56/506 (11%) was discharged to die at home, while 12/506 (2.4%) died at the hospital. Sixty-nine patients (69/506;13.6%) showed no improvement, either because they were transferred to another hospital at discharge or clinical symptoms had not improved after treatment and they had been discharged so there was no means to confirm clinical outcome. Mortality rates could range from 14% up to 27% (summarised in Figure 2.2). If gender is considered mortality rates were 58% and 42% in men and women, respectively. However, mortality was highest in HIV positive (8/12; 66.7%) patients of which the majority were men (6/8 patients; 75%), explaining the anomaly. The mean time to death at the hospital was 2 days (IQR 1-3 days) after admission, while patients, if they survived spent an average of 12-14 days recovering from BSIs.

2.3.3. The genomic diversity and population structure of BSI and carriage *E. coli* from Vietnamese patients attending HTD.

In order to construct an accurate phylogeny, we determined the core genome for the entire sequenced dataset and 22 reference *E. coli* genomes plus three *Escherichia* outgroups: one *E. fergusonii* and two *E. albertii* (Supplementary table S3). The core and pan genome of the *E. coli* was determined using ROARY (see Methods). The pan genome of all isolates analysed in this study, amounted to 42,856

genes (Figure 2.3). In total, 2,796 genes were conserved in more than 95% of all isolates. Since the average *E. coli* genome contains approximately 4,500 ~ 5,000 genes, this indicates that the core genes comprised approximately half of the average genome gene content. Interestingly, from the accessory genes, the number of genes present in less than 15% (103 isolates) of total genomes comprised 36,531 genes. This underlines the extreme level of diversity seen in these genomes and indicated that *E. coli* has an open genome (Figure 2.3), consistent with previous studies [100]. This also underlines the ability of *E. coli* to adapt to new environments and indicates that there is a high level of genetic diversity in *E. coli* present in patients attending HTD. Our pan genome prediction includes all mobile elements such as prophage and insertion sequences.

To determine the population structure of both BSIs and carriage *E. coli* included in this study, and including the reference sequences, we constructed a phylogenetic tree from a multiple sequence alignment of the core genes. We identified 443,135 single nucleotide polymorphisms (SNPs) over 2,700,577 nucleotides of core genes. Due to the high genetic diversity of the alignment, maximum likelihood phylogenetic trees were constructed using FastTree v2.1 [96]. The SNP alignments of core genes were used to subdivide the population of the sequenced *E. coli* isolates by hierarchical clustering using hierBAPS [97]. This analysis identified 15 primary BAPS clusters denoted as L1-L15 (Figure 2.4). The MLST ST and ECOR phypogroups was also predicted for all isolates in order to link the genomic data to the common terms of reference used in hospitals, like HTD, to describe *E. coli*.

Looking at MLST STs across our dataset, we identified a total of 117 STs amongst the 665 isolates analysed, of which 12/117 (10.2%) were new STs with a single locus variant (Supplementary table S1and S2). The ST distribution showed significant diversity but also highlighted the circulation of 12 dominant STs (ST131, ST95, ST1193, ST38, ST69, ST648, ST73, ST405, ST354, ST410, ST12, ST10), which account for nearly 70% (453/665) of all STs identified in this study (Figure 2.4). Among these, ST131 and ST95 were the most common in our collection, comprising 17.6% (117/665) and 8.7% (58/665) of all STs identified in this study, respectively. Apart from the 12 dominant STs, the remaining 105 STs, were each only defined in between 1 and 9 isolates. The latter are referred to as 'Others' in Figure 2.4. It is evident from Figure 2.4 that the 15 HierBAPS phylogenetic groupings are largely congruent with the MLST ST subdivisions apart from BAPS cluster 10 (L10), 11 (L11) and 14 (L14), which encompassed multiple different STs.

When comparing the WGS phylogeny with the ECOR phylogroup assignment for isolates collected in HTD, it was apparent that isolates were represented in all four historical phylogroups A, B1, B2 and D. We confirmed previous observations that phylogroup A and B1 are closely related lineages [101]. However, within each lineage A and B1, there is significant genetic diversity between isolates. For example, in phylogroup A, the median number of pairwise SNPs between isolates was 31,768 (range 0 – 72,133), whereas in phylogroup B1, the median was 28,791 SNPs (range 0 – 37,189). Genetic diversity is lower between isolates in phylogroup B2 (median 193, range 0 – 32,164) and phylogroup

D (median 3,417, range 0-51,318). It is also clear from Figure 2.4 that the ECOR phylogroups are broad subdivisions that include multiple STs and BAPS groupings. Although important to relate these findings to standard terms of reference, it is clear they do not offer sufficiently high resolution for fine scaled analysis.

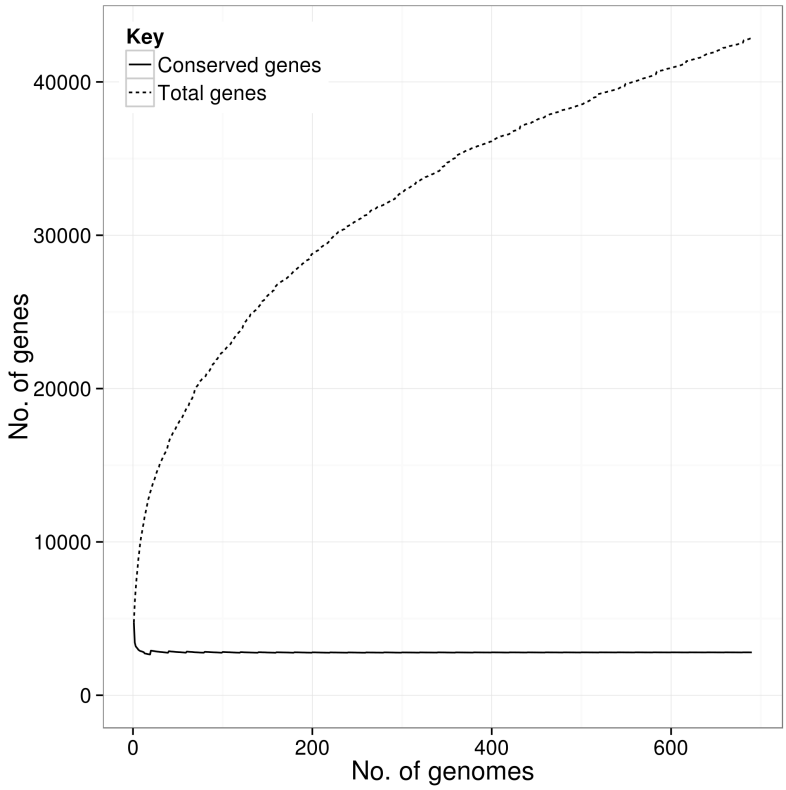

Figure 2.3: The pan genome of 687 *Escherichia coli*, two *Escherichia albertii* and one *Escherichia fergusonii* comprised 42,846 genes in total with 2,796 genes core genes present in more than 95% of all isolates. 33 genes were defined as soft core genes, which were genes present from 94 – 95% isolates. Finally, shell genes were 3,496 genes that appeared in more than 15% but less than 94% isolates.

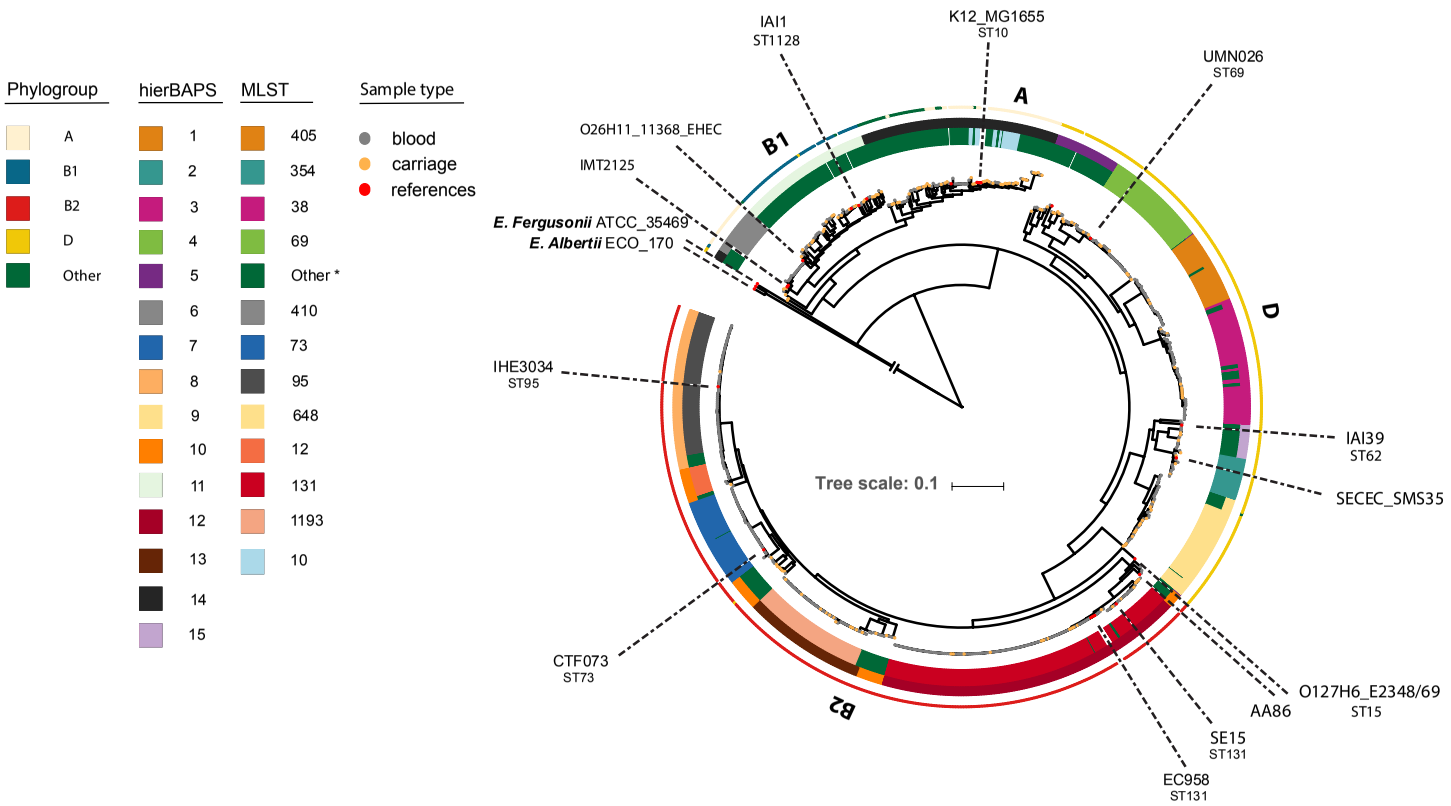

Figure 2.4: Population structure of 665 *E. coli* isolates analysed in this study. Rooted maximum-likelihood tree of *E. coli* from a core gene alignment consisting of 2,796 genes. Tree was rooted using *E. albertii* (ECO 170, KF1) and *E. fergusonii* (ATCC 35469) as outgroups. The tree includes data from 22 reference genomes representing four phylogroups, represent as red dots at the branch tips. The coloured ring from inside to outside includes MLST sequence type (ST) for the dominant 12 STs (See Figure 2.7 for representation of all STs), all hierBAPS clusters and the four ECOR phylogroups. Grey and yellow dots at the branch tips represent BSI isolates and rectal isolates, respectively. Other * denotes allelic variants of specific STs and other rare STs.

Relating whole genome phylogeny, BAPS groups, ECOR phylogroupings and MLST STs to patient clinical data.

The WGS phylogeny of the *E. coli* isolates was then used as a framework on to which MLST, ECOR phylogroup and the clinical metadata was mapped. When relating the phylogeny to clinical data including patient gender, comorbidities as well as linking patient BSIs samples to corresponding isolates reported from other sterile sites (urine/ peritoneal fluid), we identify several patterns. First, although isolates taken from female and male patients were distributed equally across the phylogeny (Figure 2.5), in certain BAPS clusters and STs there was a skew in the gender distribution (Figure 2.5). For example, in ST69(L4), 95(L8), 410(L6), 1193(L13) of phylogroup B2 and D (>60% patients were female) while in other STs of phylogroup B1 and A (L11/L14), majority of patients were male (>60%). These STs 69, 95 and 410 are known to be associated with urinary tract infections (UTIs) in women [102], and this was supported by having the corresponding isolates with the same antimicrobial resistant (AMR) profile from urine isolates listed in the HTD patient database. Secondly, looking at the patient demographic and clinical data for lineages included within ECOR phylogroups B1, A and D showed *E. coli* isolates more associated with carriage and male patients with liver disease as well as patients that also tested positive by culture for *E. coli* from peritoneal fluid (Figure 2.5). Interestingly, isolates taken from patients with HIV were more likely to be associated with phylogroup D and B2, but not exclusively.

Comparing ST and sample type, it was evident that STs within the B2 phylogroup were more likely to have been isolated from blood (52.17%) than stool (25.16%) (264/506 *versus* 40/159, $p<0.001$, Chi square test, Figure 2.4). In contrast, rectal swab isolates were enriched in phylogroups A and B1: 16.3% (26/159) phylogroup A in rectal swab *versus* 7.31% (37/506) in blood ($p<0.001$) and 12.58% (20/159) phylogroup B1 in rectal swab *versus* 5.53% (28/506) in blood ($p=0.003$). However, blood and rectal swab derived isolates were almost equally represented amongst isolates belonging to phylogroup D (31% in blood *versus* 36% in rectal swabs, $p=0.275$).

From looking at all 117 STs, there were a total of 42 STs only found in isolates taken from blood samples, 42 STs only found only in rectal swab derived isolates, and the remaining 33 STs were evenly distributed between blood and rectal swab *E. coli* (Figure 2.6). In our study, the most globally prevalent ST, ST131, was the most common ST isolated from both blood (19.7%, 100/506 isolates) and rectal swab samples (6.7%, 10/159 isolates) possibly suggesting that carriage in the gut is linked to pathogenic *E. coli* establishing BSIs in an otherwise healthy individual. Conversely, none of ST73 and only one of ST95 were found in rectal swabs (Figure 2.6). This suggests that if these two STs entered to blood after colonizing a different bodily site, then they either lack the ability to colonize the gut in sufficient number to be detected in our clinical screen, or that they may be more prone to colonizing bladder and kidney epithelial cells instead of gastrointestinal tract.

E. coli carriage isolates, taken from rectal swabs, showed many different STs and long branches consistent a highly diverse community of *E. coli* living together in our gastrointestinal tracts (Figure 2.7). However, none of our isolates belonged to the other *Escherichia* species, *E. fergusonii* or *E. albertii,* even when we sample a large collection of *E. coli* longitudinally over a 5-year period.

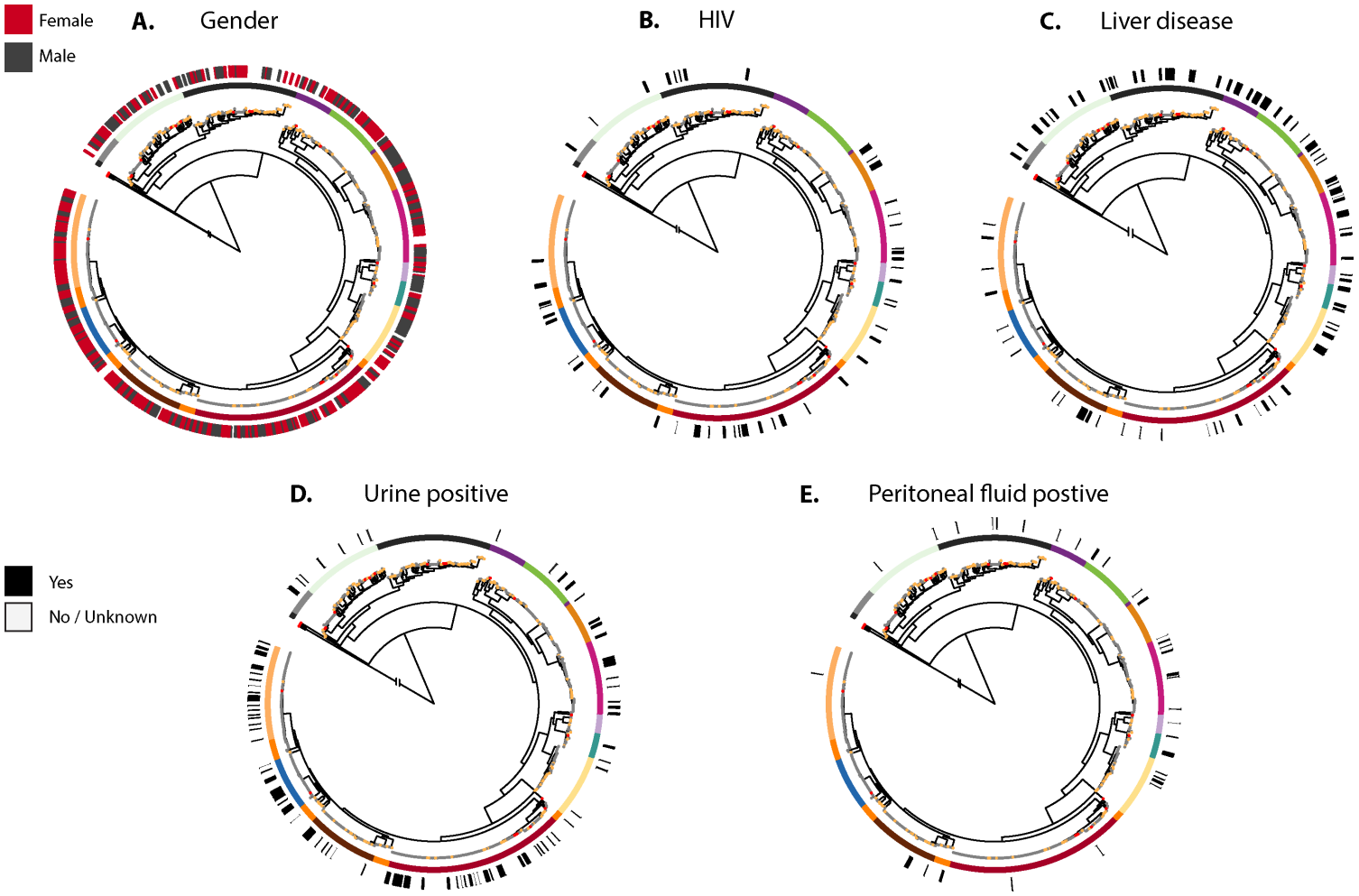

Figure 2.5: The phylogenies of *E. coli* taken from HTD, including reference strains, reported alongside clinical metadata for *E. coli* causing BSI. Inner rings denote BAPS clusters (see key of Figure 2.4), outer rings denote comorbidities or the presence of a corresponding isolate. See key for gender (A) or presence/absence (B-E).

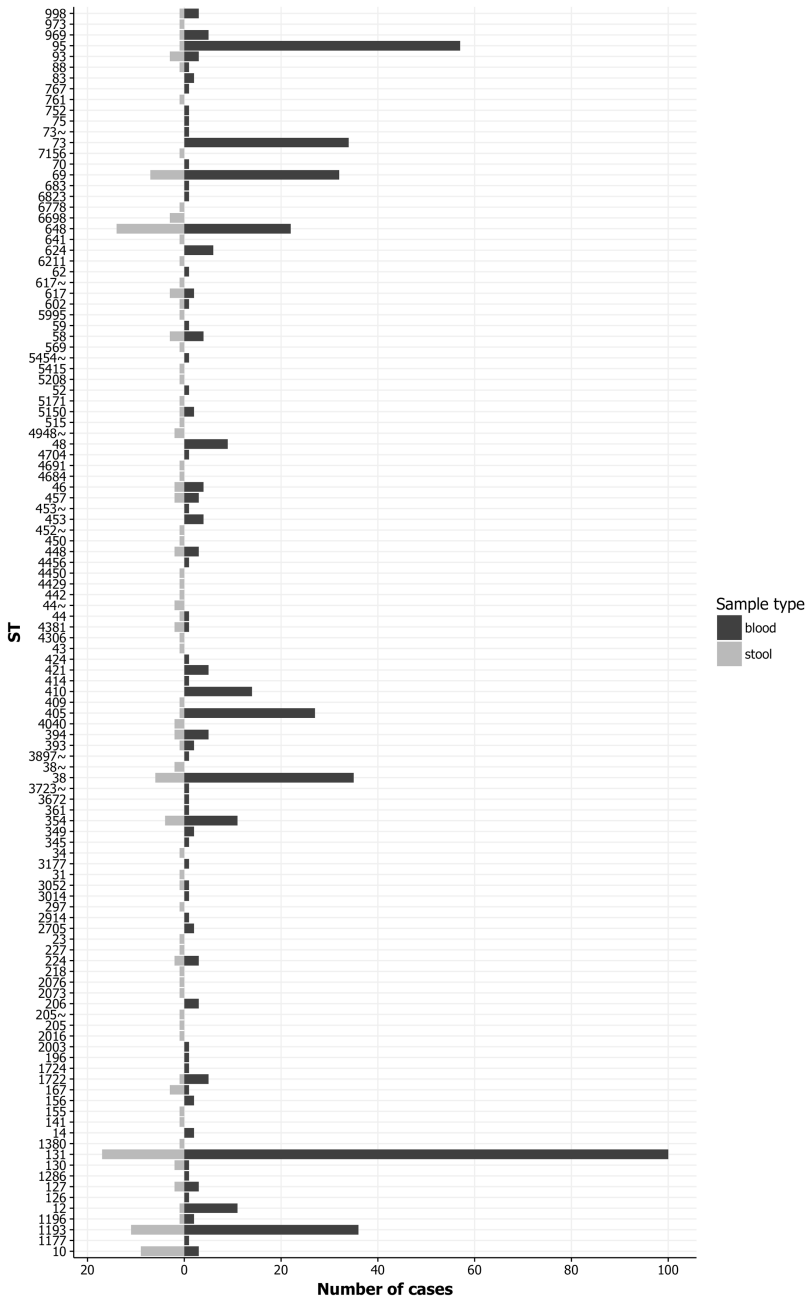

Figure 2.6: STs distribution of blood and rectal swab isolates. For colour see key.

In addition to differences in the relative abundance of STs across the whole collection we observed minor fluctuations in the relative proportion of different STs/BAPS clusters that comprised the total *E. coli* for different years (Figure 2.7), the exception being the proportion of ST131 which remained high throughout the whole 5-year period ranging from 18% to 22% of isolates sequenced/year. The most striking increase in numbers observed was for ST1193 which accounted for 3.7% *E. coli* BSIs in 2011, but then increased to 10% in 2014. We observed a reduction in the proportion of ST69, ST73 and other less common STs over the same time period.

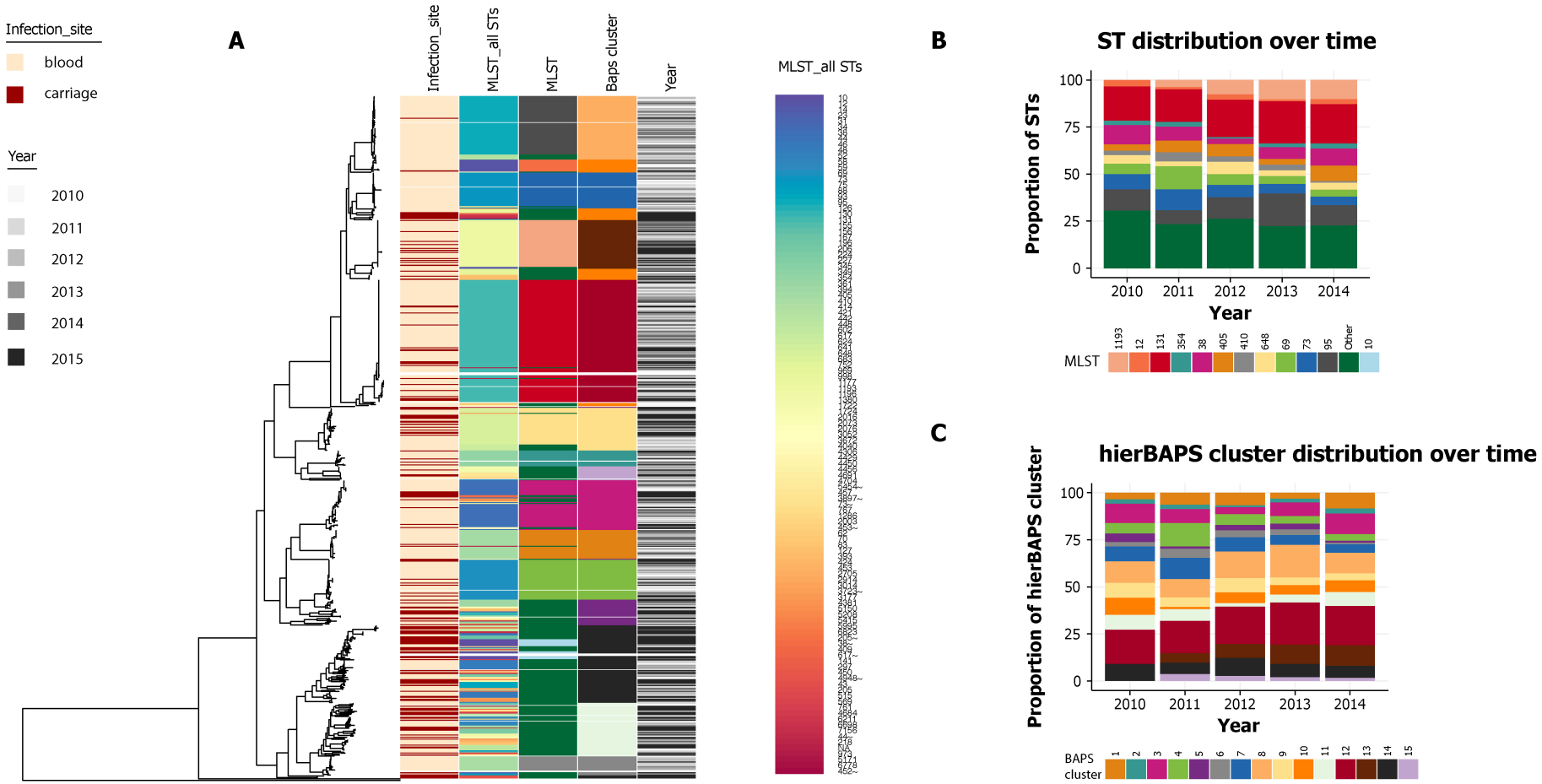

Figure 2.7: A core genome tree visualised in Phandango for blood and rectal swab isolates (A). The distribution of STs (B) and hierBAPS clusters (C) are reported for BSI *E. coli* over a five-year period from 2010 to 2014. Blood and rectal swab samples collected in 2015 were not included in panels B and C because they were not sampled randomly.

2.4. Discussion

Among 38, 000 hospital admissions to HTD every year, *E. coli* causes more BSIs than any other Gram-negative bacteria or fungi, accounting for around 20% of all BSI cases (HTD unpublished hospital records). Unlike seen in other nosocomial bacteria such as *Klebsiella pneumoniae* or *Staphylococcus aureus* [103, 104], the majority of BSI in our study (88%) were from patients with BSIs diagnosed on or within 2 days of admission and so deemed to have been community acquired. The remaining patients were considered to have HAI *E. coli* infections because they acquired the BSI >2 days after admission (this is discussed further in Chapter 4 for the longitudinally sampled patients). All of these patients had severe chronic diseases such as AIDS, cirrhosis, hepatitis B/C or tetanus which required long term hospitalization, and sometimes were seen to develop *E. coli* BSIs as secondary infections as described in detail in Chapter 4. Positively, in terms of patient care, of the 506 patients that had blood stream infections over the period of study 70% recovered and so likely responded to antimicrobial treatment. However, treatment took on average between 12-14 days representing a considerable burden on space and resource within HTD. It was also apparent from the basic patient meta data that was a slight gender bias in BSI towards females (53%) in the sample collection sequenced in this study. In contrast, BSI at HTD caused by all pathogens are generally biased towards occurring in men (65% of all BSI cases, unpublished data).

Aiming to understand the nature of *E. coli* causing BSIs in HTD, Vietnam as well to see if the trend of *E. coli* BSIs seen here mimics patterns of *E. coli* causing disease elsewhere in the world, we used WGS to investigate the population structure of isolates taken from patients with *E. coli* BSIs from HTD between Jan 2010 to Dec 2014. Also included were patients with matched blood stream and rectal swab isolates from 2015. The WGS data for all isolates was used to construct an accurate whole genome phylogeny. This was then used to predict BAPs clusters, ECOR phylogroups and MLST sequence types for all isolates. Not only is this the first study of its type at HTD, this is also the largest collection WGS of *E. coli* from one hospital to date.

We used these data to determine the composition of different STs with potential to carriage and causing BSIs in the HTD patients and showed that there was high diversity in isolates affecting patients at HTD including 15 primary BAPS clusters all the ECOR phypogroups and a total of 117 STs by in silico MLST. From these data, it was clear that the four dominant STs affecting patients attending HTD are the same in many studies internationally [102, 105]. Study by Adams-Sapper *et al* [105] found the same 4 STs dominant from 2007-2010 when investigating the nature of *E. coli* causing BSIs in United States. What was also apparent from these data was that these STs remained largely stable over the time of the study. The exception being ST 1193. We observed the introduction of ST1193 to patients at HTD in 2011, since then ST1193 has increased steadily to account for ~10% BSIs cases every year. Interestingly, there was little documentation of ST1193 in the literature until it was reported recently as

the second most common ST after ST131 in China in 2011 [106]. We hypothesised that the introduction and successful establishment in Vietnamese population was linked to antimicrobial resistant genes they carried. We will discuss in more detailed about ST1193 in final discussion in Chapter 5.

Here we also found strong associations between patient gender and comorbidity with the lineage of *E. coli* causing BSIs: lineages (STs and BAPS groups) within the ECOR phylogroup B1 and A were significantly associated with rectal swab samples, men and liver abscesses. Isolates in these phylogroup groups showed high pairwise differences in SNP content between isolates. Lineages within the ECOR phylogroup D were equally represented in blood and rectal swab samples. But perhaps most importantly our data showed that there was a significant association between members of phylogroup B2 and invasive disease at HTD. In addition, isolates belonging to B2 (as well phylogroup D) were also associated with female patients. For HIV patients (most of whom were male patients in our study), these data suggested that there was no significant association between the phylogenies of the isolates causing BSI and HIV occurrence. These data are consistent with previous studies which have shown that extraintestinal pathogenic *E. coli* (ExPEC; Chapter 1) commonly belong to phylogroup B2, and less frequently phylogroup D, whereas the more diverse commensal and intestinal pathogenic *E. coli* are often members of phylogroups A and B1 [13, 107, 108].

In summary, here we determined the full diversity, and flux in that diversity, of isolates from patients attending HTD. The following chapters will further investigate the possible link between clinical outcome and bacterial lineage by focussing on the presence or absence of known virulence genes and those associated with antimicrobial resistance.