Chapter 3: The accessory genome of *Escherichia coli* of patients attending HTD.

3.1. Introduction

In Chapter 2, we have shown that *E. coli* from rectal swabs and blood samples of patients attending HTD are extraordinarily diverse, with a core genome composed of only ~ 3,000 genes but a pan genome in excess of 40,000 genes, most of which are very specific to a subset of isolates (less than 15% of strains). Within this diversity there are likely to be many different non-pathogenic commensals as well as a range of different pathotypes (see Chapter 1). Since none of our rectal samples came from patients with enteric disease, we were not expecting to find intestinal pathogenic *E. coli* (InPEC). However, data described in Chapter 2 showed that there was a significant association between the phylogenies of the isolates sequenced and sample type, with members of phylogroup B2 significantly associated with blood stream infection at HTD. This is consistent with the literature which shows that extraintestinal pathogenic *E. coli* (ExPEC) dominate this phylogroup clinically [108]. Unlike InPEC which are obligate pathogens and have certain distinctive virulence factors that can be used to define them exclusively, ExPEC and commensal *E. coli* are not easily distinguished from one another [10], not least because ExPEC can also be found in healthy individuals. Nevertheless, there are a set of characteristic virulence genes that were extensively discovered by screening across ExPEC *E. coli* (adapted from [108, 109]; see Table 3.1). Of the ExPECs the most common cause of human disease comes from the uropathogenic *E. coli* (UPEC) subdivision. Whilst UPECs are the major cause of urinary tract infections they are often found in the gut of the infected individuals as well as causing more systemic disease [110].

Table 3.1: Common virulence factors identified in ExPEC

Functional category	Virulence factors	Gene detected
(1) Adhesion	Afimbrial adhesin	*afaABCDEF/draP*
	Type 1 fimbriae (*Fim*)	*fimABCDEFGHI*
	P fimbriae (*Pap*)	*papABCDEFGKX*
	S/F1C fimbriae (*Sfa*, *Foc*)	*sfaABHSY, focACDFGH*
	Common pilus	*yag_ecp*
	Adhesin	*fdeC*
	Dispersin (intestinal adhesin in EAEC)	*aap_aspU*
(2) Iron acquisition	Enterobactin	*entABCDEFS,fepABCD,fes*
	Salmochelin	*iroBCDEN*
	Aerobactin	*iucABCD,iutA*
	Yersiniabactin	*fyuA,ybtAEPQSTUX,irp12*
	Hemin uptake system	*chuASTUVWXY*
(3) Toxin	Enterotoxin	*astA,senB*
	Hemolysin	*hlyABCD*
	Cytotoxic necrotising factor 1 (CNF-1)	*cnf1*
	Serine protease autotransporters *Sat*, *Pic*	*sat, pic*
	Haemoglobin protease	*vat*
(4) Invasin	Invasion of brain endothelium	*ibeA*
(5) Protectins	K1 capsule	*kspDMT*
	Tir domain containing protein	*tcpC*

The genes in Table 3.1 were the main virulence determinants responsible for the mechanisms by which uropathogenic *E. coli* (UPEC) can cause disease, including cystitis and pyelonephritis [111]. These functions can be divided into at least five categories, with many of these factors indeed occuring with high redundancy [112]. In order to establish infection, UPEC first needs to attach to host cells through different types of fimbriae. Among them, Type 1 fimbriae, especially *fimH*, are essential to causing infection in UPEC [113]. Toxins such as *hly* and *vat* are used to lyse host cells, both for bacterial

nutrition and to disseminate to deeper tissues. Iron acquisition systems are used to chelate free iron from host, to enable bacterial proliferation [114]. Immune evasion is achieved by producing capsule (*kspDMT*), to protect bacteria from serum bactericidal activity of human plasma [115, 116].

Apart from virulence genes, antimicrobial resistance (AMR) genes are also an important component of the accessory genome since AMR genes confers selective advantage on members of bacterial populations when under antibiotic selection. MDR *E. coli* have also recently been listed by World Health Organization (WHO) as an urgent threat to humanity, and new antibiotics are urgently needed [64]. Currently, many genes mediating resistance to last resort antibiotics such as bla_{NDM} and *mcr-1* are reported to be located on plasmids [58, 117]. *E. coli* infections in Asia are reported as having the highest prevalence of ESBL-producing isolates, mainly by possessing the $bla_{\text{CTX-M}}$ gene [118]. Tracking and understanding the diversity of AMR genes in *E. coli* causing infections in our collection is required to limit the spread of these bacteria in both community and hospital environments.

Plasmids have long been recognised as the vehicle by which AMR genes and virulence genes spread quickly between bacterial species. Acquisition of virulence and AMR plasmids might change the prevalence of a specific clone in disease-causing bacterial populations [119]. One classic example is that simply by acquiring the 230kb virulence plasmid pINV, which encodes T3SS, both *Shigella spp*. and EIEC can invade cells, live an intracellular lifestyle, and cause diarrheal disease [120]. Other distinct InPEC pathotypes are also mediated by virulence plasmids, such as plasmid pEAF which encodes an adherence factor in enteropathogenic *E. coli* (EPEC); plasmid pAA encoding aggregative adherent fimbria (AAF) in enteroaggrative *E. coli* (EAEC); and colonisation factors, ST and LT toxin can be located on a plasmid in enterotoxigeneic *E. coli* (ETEC) [121]. Plasmids in *Enterobacteriaceae* are diverse in terms of their size and the number of genes they carry. A *rep* gene PCR based method has been developed to classify plasmids as belonging to different Incompatibility (Inc) classes, called replicon typing scheme [122].

However, information about virulence, AMR and plasmid replicon genes are often missing from the core genome analysis because by definition these are not core genes. Thus, it is essential to screen for different components of accessory genomes and to relate them back to the core genome phylogenetic tree to look for genes that differently distributed between them.

Aim:

We hypothesised that specific virulence factors might explain the virulence potential of ExPEC relative to carriage *E. coli*, as well as explaining the successful expansion of several common lineages such as ST131, ST1193, ST73 and ST95. In order to test this hypothesis, we used a systematic approach to scan the *E. coli* genomes in our study for the presence of known and well-defined virulence genes, as well as for genes known to confer antimicrobial resistance. We then compared the distribution of these genes between the randomly sampled isolates cultured from blood and rectal swabs (see Chapter 2 Section

2.2). We used these data to look for any associations between genes and sample site, and to study differences between different lineages and STs.

Specific questions to be answered:

1) Are there any genes significantly associated with disease phenotype (blood/carriage)?

2) How are virulence genes distributed between different lineages?

3) Do AMR genes confer a selective advantage on population of *E. coli* causing invasive disease?

3.2. Methods

3.2.1. *In silico* virulence, AMR, and plasmid replicon type gene detection

ARIBA [123] was used to screen genomes for the presence of antimicrobial resistance (AMR) genes. Briefly, this algorithm mapped raw reads to reference gene sequences, used mapped reads to perform local assemblies, and to report gene presence or absence as well as whether the gene in question existed as a new variant. We also use ARIBA to screen for virulence genes and replicon types. AMR gene sequences were obtained from the ResFinder database [124], virulence gene sequences were taken from the VFDB database [125], and plasmid replicon types from the PlasmidFinder database [119]. PlasmidSPADES [126] was used to identify contigs likely to be plasmids from SPADES assembly graphs [127] and plasmid contigs were visualised using Bandage [128].

3.2.2. Antimicrobial phenotype testing

All isolates were tested for their antimicrobial susceptibility phenotype using 13 antimicrobials including amikacin (30 μg) (AMK), amoxicillin (20 μg) / clavulanic acid (10 μg) (AMC), ceftazidime (30 μg) (CAZ), ceftriaxone (30 μg) (CRO), cefepime (30 μg) (FEP), ciprofloxacin (5 μg) (CIP), ofloxacin (5 μg) (OFX), trimethoprim (1.25 μg) /sulfamethoxazole (23.75 μg) (SXT), imipenem (10 μg) (IPM), ertapenem (10 μg) (ETP), meropenem(10 μg) (MEM), tazocine or piperacillin (100 μg) / tazobactam (10 μg) (TZP), ticarcillin (75 μg) / clavulanic acid (10 μg) (TCC). Antimicrobial susceptibility was determined using the disk diffusion method following Clinical and Laboratory Standard Institute (CLSI) guidelines [129]. An EBSL-producing phenotype was determined by the double disk diffusion method [130], using a combination of third-generation cephalosporin discs, *i.e.* ceftazidime, ceftriaxone, cefepime and amoxicillin (20 μg) / clavulanic acid (10 μg). All phenotypic tests were performed by Microbiology staff at HTD as part of clinical routine laboratory work.

3.2.3. Statistical analysis

Logistic regression was used to calculate the association between a gene and an outcome of being invasive or carriage phenotype in Stata v.13 (StataCorp). The *P* value was adjusted for multiple comparisons using the Bonferroni correction method. Values of $P < 0.05$ after Bonferroni correction were considered to be significant. Differences in gene distribution in each group were calculated using

Chi squared or Fisher exact test (when observation in contingency table <5). Distribution of total AMR genes and replicon genes between STs/BAPS lineage or blood/carriage were computed using the Wilcoxon test. Gene presence/absence data were visualised in Phandango [99] and plots were produced by ggplot2 package [131] in R (R Foundation).

3.3. Results

3.3.1. The distribution of virulence factors between blood and rectal swab isolates

Using the virulence database 'VFDB' [125], containing 2,600 reference genes across Gram-negative and Gram-positive bacteria, we identified a total of 205 virulence genes across our strain collection, classified into five groups: adhesion, iron acquisition, toxins, invasins and protectins (Table 3.1). By comparing the distribution of these genes between blood and rectal swab isolates our analysis, we identified several genes that might serve as predictors for an invasive phenotype (OR $\geq$ 2, p $\leq$ 0.05). These include loci related to adhesion (*pap* cluster, *fdeC*), iron acquisition system (yersiniabactin, salmochelin, aerobactin), toxins (*hly*, *vat*) and protectin (*kspD*) (Figure 3.1). One of the most significant associations was between P fimbriae and invasive disease (*papBCDEFGIJK*; OR $\geq$ 3, p $\leq 10^{-8}$). P fimbriae mediate the ability to adhere to kidney epithelial cells and are strongly associated with ExPEC and invasive disease [132]. Of the siderophores showing a strong odds ratio for blood stream infections, the highest score was seen for yersiniabactin (*ybt*, *irp*, *fyuA*) (OR 3.3, 95% CI 2.2-5, p = 4.26 x 10^{-8}) followed by salmochelin (*iro*) (OR 2.7, 95% CI 1.64-4.5, p = 3.5 x 10^{-4}) and aerobactin (*iuc*) (OR 2.2, 95% CI 1.6-3.3, p = 4 x 10^{-5}). Enterobactin, an archetypal siderophore system, was identified across all *E. coli* isolates included in this study, and therefore is not considered to be characteristic of ExPEC, consistent with previous studies [133].

We also detect the association between member of serine protease autotransporters (SPATE) family including *sat*, *pic* and *vat*, which were also more likely to be present in invasive isolates than carriage. Each autotransporter has been shown to cause vacuolating of host tissues [112]. Other detected virulence factors have also been well-studied, including hemolysin (*hlyABCD*), *kspMD* which mediates K1 capsule, and *ibeA*, all of which increase the odds of causing invasive disease.

Apart from genes that are enriched in blood isolates, we also found genes/loci that are more likely to be associated with intestinal pathogenic *E. coli* (InPEC) isolates. These included *esp* genes which encode type three secretion system effector proteins and are encoded on the locus of enterocyte effacement (LEE), a pathogeneicity island, characteristic of enterohemorrhagic *E. coli* (EHEC) and EPEC pathotypes, both members of the InPEC (see section 3.1). Included within this locus we detected *espX* which is only known to be found on the LEE in EHEC isolates [134] (Figure 3.2). In addition, we also found many other virulence genes characteristic of O157:H7 such as *cesABDLT*, *etgA*, *nleBCEGH*, *sepDLQ*, *escCDEFGIJLNMPRSTUV* in one patient with a fatal outcome, suggesting that EHEC

O157:H7 infection (ST752; data not shown) is rare in Vietnam, but can subsequently gain access to the bloodstream.

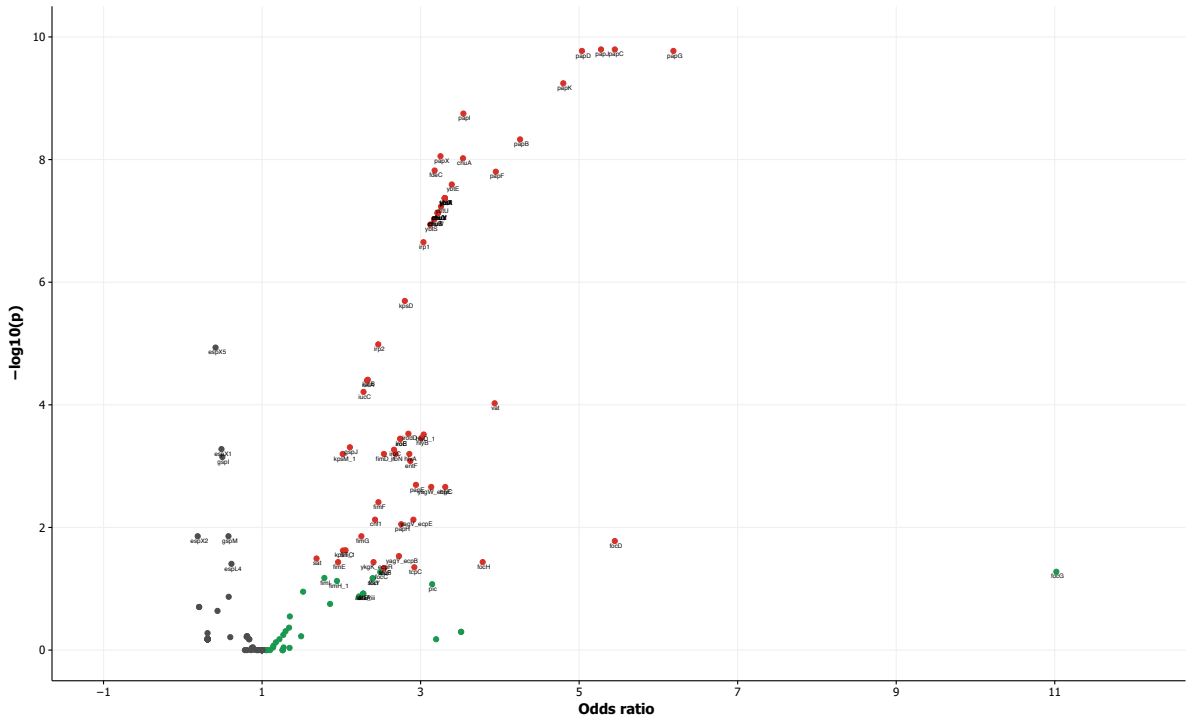

Figure 3.1: Volcano plot for odds ratio versus $-\log_{10}$ p-value of gene distribution between invasive and carriage isolates. The graph shows the odds ratio (x-axis) versus the $-\log_{10}$ of p-value (y-axis) from the results of the logistic regression model. Each dot represents a single gene. OR< 1 indicates that the gene is more likely to be associated with rectal swab isolates (coloured in grey) whereas genes with OR> 1 are more likely to be associated with blood isolates. Genes in red represent the association is statistically significant with $-\log_{10}$(p-value) ≥ 1.3 (*i.e.*, $p < 0.05$). Genes in green indicate that there is an association between tested gene and invasive disease but it is not statistically significant. All analysis was performed in Stata v.13 (StataCorp) and results were plotted using ggplot2 package in R (R computing).

3.3.2. The distribution of virulence factors between different BAPS lineages/ phylogroups/STs

When studying the distribution of virulence genes across lineages, we noticed that genes were clustered according to ECOR phylogroups (Figure 3.2). For examples the *chu* and *pap* genes are absent from phylogroup A (L14)/B1 (L11), whereas *esp* genes are mostly present in phylogroup D (L1,2,3,4,9) and phylogroup A/B1. A certain number of virulence genes are present in all isolates regardless of their lineages or potential pathogenic outcome. These include genes that mediate enterobactin, outer membrane protein A, type 1 fimbriae (*fim*), general secretion pathway (*gsp*) on T3SS, and common pilus (*yag_ecp*).

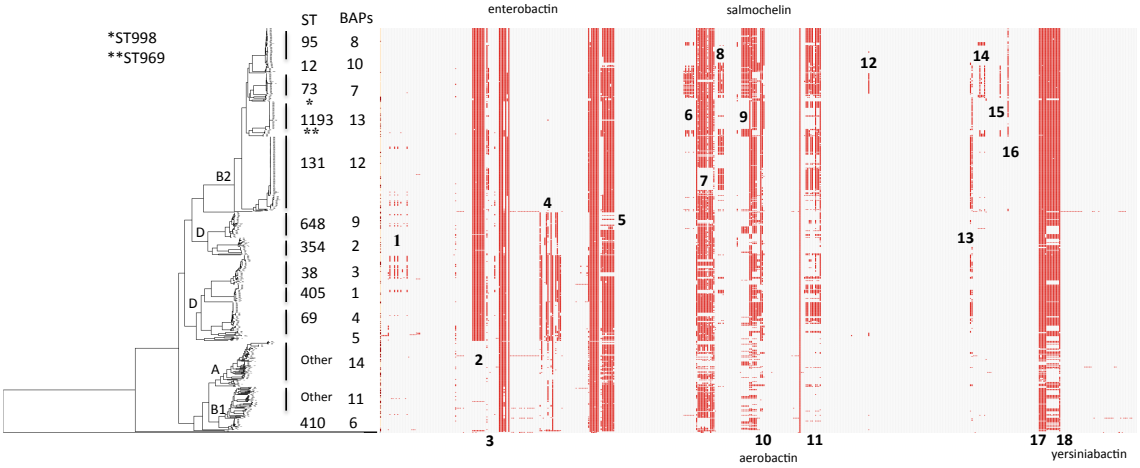

Figure 3.2: Phylogeny of *E. coli* lineages and virulence genes found in each isolate. Red dots represent gene presence; white indicates gene absence. The tree, rooted on *E. albertii*, includes STs and BAPS lineage information. Key gene differences within and between STs were highlighted (1 to 18). (1) Afimbrial adhesion (*afa*/*draP*), (2) Hemin uptake system (*chu*), (3) *papB*, (4) T3SS effector (*esp*), (5) adhesion (*fim*), (6) adhesion (*foc*), (7) T2SS general secretion pathway protein (*gsp*), (8) hemolysin (*hly*), (9) salmochelin (*iro*), (10) aerobactin (*iuc*, *iut*), (11) adhesion (*pap*), (12) autotransporter (*pic*), (13) toxin (*sat*, *senB*), (14) adhesion (*sfa*), (15) protectin (*tcp*), (16) toxin (*vat*), (17) adhesion (*yag*-Ecp), (18) yersiniabactin (*ybt*).

Breaking the phylogroups down into ST or BAPS lineages, we looked at the distribution of the virulence genes in more detail across the twelve most prevalent STs in our collection (Table 4.2). Within phylogroup B2, the four dominant STs are ST95 (BAPS L8), 73 (L7), 1193 (L13) and 131(L12). Although ST131 is a globally successful pathogenic clone of *E. coli,* ST73 carried eight more virulence genes than ST131, ST95 and ST1193 (Figure 3.2), and several of these genes were exclusive to ST73 in our collection. These included the *foc* gene cluster (*focA* to *focI*), *sfa* cluster (*sfaA*, *sfaF*), *hly* (*hlyA* to *hlyC*), *cnf1*, *pic*, *sat*, *vat* and *tcpC* (Table 4.2). ST73 in our study had more *Pap* genes than ST131, as well as having additional fimbriae such as S1 fimbriae (*sfa* operon) and F1C fimbriae (*foc* operon). Interestingly other rare STs in phylogroup B2, including ST12/ST969/ST998 (L10), appear to possess the same virulence profile as ST73 and so likely to be capable of causing invasive disease.

Even within a specific ST, we observed the gain and loss of several specific genes. For example, within ST131, not all isolates (62%) possess P fimbriae, and those that lack P fimbriae may harbour *afa/draP* (9.4%) and to a lesser extent *sfa/foc* (1.7%). In addition to having genes related to colonization, 50% to 80% of our ST131 isolates carry gene that encode toxins including *hly*, *sat*, *senB* and *cnf*-1 (Figure 4.2, Table 4.2). Of the other STs carrying *pap* genes 12% of the member of ST38 in our collection harbour

papG. In those that lack *papG*, 63.4% utilise afimbrial adhesin (*afa*) and Dr fimbrae (*draP*) which can bind to uroepithelial cells in the same way as P fimbriae. 12% of ST38 also use dispersin (*aap*/*aspU*), a protein mediated colonization factor characteristic of enteroaggrative *E. coli* pathotypes (EAEC) for binding.

Table 3.2: The percentage of isolates carrying specific virulence genes in the 12 most dominant sequence types of *E. coli* found in this study. Total represents the percentage of all isolates among the 665 isolates that possess the gene/locus.

Functional category	Gene cluster	Genes	Phylogroup												
			ST												
			10	**1193**	**12**	**131**	**354**	**38**	**405**	**410**	**648**	**69**	**73**	**95**	**other**	**Total (%)**
Adhesion	*afaABCDEF/draP*	*afaACDF*	0	0	0	9.4	0	63.4	14.3	0	13.9	2.6	0	0	3.8	8.3
		afaE1	0	0	0	2.6	0	2.4	14.3	0	8.3	0	0	0	0.5	1.8
		afaE2	0	0	0	0	0	12.2	0	0	0	0	0	0	0	0.8
		draP	0	0	0	9.4	0	63.4	14.3	0	13.9	2.6	0	0	3.8	8.3
	Type 1 fimbriae (Fim)	*fimA*	25	0	50	0.9	0	73.2	0	71.4	25	0	97.1	0	31.1	23.8
	fimABCDEFGHI	*fimB1*	100	95.7	83.3	18.8	73.3	82.9	100	100	30.6	100	100	82.8	79.2	71.6
		fimC	100	95.7	83.3	100	80	82.9	100	100	30.6	100	100	100	85.4	89.5
		fimD1	91.7	95.7	66.7	100	80	75.6	100	100	30.6	100	100	100	82.1	87.5
		fimE	100	95.7	83.3	99.1	80	82.9	100	100	30.6	100	100	100	86.3	89.6
		fimF	100	93.6	75	100	86.7	82.9	100	100	44.4	100	100	100	89.2	91.3
		fimG	100	95.7	83.3	100	86.7	82.9	100	100	44.4	100	85.3	100	89.2	90.8
		fimH	91.7	95.7	83.3	99.1	86.7	82.9	100	100	44.4	100	100	100	92	92.2
		fimI	100	95.7	83.3	100	80	82.9	100	92.9	30.6	100	100	100	81.6	88.1
	P fimbriae (Pap)	*papA*	0	0	0	0.9	0	7.3	21.4	0	0	0	2.9	0	0.5	1.8
	papABCDEFGKX	*papB*	0	23.4	91.7	45.3	0	17.1	71.4	21.4	8.3	79.5	85.3	94.8	21.7	40.5

		papC	0	27.7	100	60.7	0	12.2	50	7.1	2.8	79.5	97.1	94.8	17.5	41.1
		papD	0	27.7	100	62.4	0	12.2	53.6	7.1	2.8	79.5	97.1	94.8	17.9	41.7
		papE	0	4.3	25	1.7	0	19.5	0	0	19.4	5.1	44.1	93.1	11.8	17.7
		papF	0	27.7	75	62.4	0	22	53.6	7.1	16.7	82.1	94.1	94.8	9.9	40
		papG	0	27.7	83.3	62.4	0	12.2	53.6	7.1	0	76.9	97.1	93.1	7.5	37.6
		papH	0	4.3	66.7	4.3	0	7.3	25	0	2.8	74.4	79.4	3.4	6.1	14.6
		papI	16.7	93.6	91.7	91.5	0	48.8	57.1	7.1	22.2	76.9	85.3	94.8	23.1	55.9
		papJ	0	27.7	100	62.4	0	12.2	53.6	7.1	2.8	79.5	97.1	94.8	17.5	41.5
		papK	0	27.7	91.7	61.5	0	12.2	53.6	7.1	2.8	82.1	97.1	94.8	15.6	40.8
		papX	16.7	100	75	82.1	60	78	89.3	7.1	75	84.6	82.4	98.3	25	63
	S/F1C fimbriae (Sfa,Foc)	*sfaA*	0	0	0	0	0	0	0	0	2.8	0	0	8.6	1.9	1.5
	sfaABHSY	*sfaB*	0	0	91.7	1.7	0	0	0	0	2.8	0	85.3	8.6	9.4	10.2
		sfaH	0	0	0	0	0	0	0	0	2.8	0	14.7	8.6	4.7	3.2
		sfaS	0	0	0	0	0	0	0	0	2.8	0	0	8.6	4.2	2.3
		sfaY	0	0	91.7	1.7	0	0	0	0	2.8	0	79.4	8.6	9	9.8
	focACDFGH	*focA*	0	0	16.7	0	0	0	0	0	0	0	76.5	0	1.4	4.7
		focC	0	0	91.7	1.7	0	0	0	0	2.8	0	82.4	8.6	9.4	10.1
		focD	0	0	16.7	0.9	0	0	0	0	2.8	0	85.3	8.6	6.1	7.7
		focF	0	0	91.7	1.7	0	0	0	0	2.8	0	85.3	8.6	9.4	10.2
		focG	0	0	16.7	0	0	0	0	0	0	0	82.4	0	1.9	5.1

		focH	0	0	91.7	0	0	0	0	0	0	0	79.4	0	5.2	7.4
		focI	0	0	91.7	1.7	0	0	0	0	2.8	0	76.5	8.6	9.4	9.8
	Common pilus	*yagv_ecpE*	83.3	100	100	100	93.3	100	100	92.9	100	100	97.1	100	84.9	94.4
	yag_ecp	*yagw_ecpD*	66.7	100	100	100	93.3	100	100	92.9	100	100	100	100	84	94
		yagx_ecpC	83.3	100	100	100	93.3	100	100	92.9	100	100	100	100	83.5	94.1
		yagy_ecpB	83.3	100	100	100	93.3	100	100	92.9	100	100	100	100	88.2	95.6
		yagz_ecpA	83.3	100	100	100	93.3	100	100	92.9	100	100	100	100	85.4	94.7
	Adhesin	*fdeC*	50	97.9	100	97.4	0	100	100	64.3	0	0	100	100	41	65.4
	Dispersin (intestinal adhesin in EAEC)	*aap_aspU*	0	0	0	0	0	12.2	0	0	0	0	0	0	6.1	2.7

Functional category	Gene cluster	Genes	ST												
			10	**1193**	**12**	**131**	**354**	**38**	**405**	**410**	**648**	**69**	**73**	**95**	**other**	**Total (%)**
Iron acquisition	Enterobactin (*entABCDEFS*)	*entABCES*	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		entD	100	0	100	0	100	100	100	100	100	97.4	100	100	74.1	66.9
		entF	0	0	0	82.1	0	0	0	0	0	0	100	0	4.2	20.9
	fepABCDG	*fepA*	100	100	100	100	100	100	100	100	100	100	100	55.2	97.6	95.3
		fepBCDG	100	100	100	100	100	100	100	100	100	100	100	100	99.5	99.8
	Salmochelin (*iroBCDEN*)	*iroBE*	16.7	0	91.7	0	0	2.4	0	35.7	33.3	12.8	82.4	96.6	23.6	25.6
		iroCD	16.7	0	91.7	0	0	2.4	0	35.7	30.6	12.8	82.4	94.8	23.1	25.1
		iroN	16.7	0	75	0	0	0	0	35.7	33.3	12.8	82.4	89.7	22.6	24.2
	Aerobactin (*iucABCD, iutA*)	*iucABC*	66.7	97.9	91.7	97.4	13.3	58.5	71.4	92.9	80.6	92.3	58.8	96.6	33.5	67.7
		iucD	8.3	0	83.3	0.9	0	19.5	25	0	11.1	69.2	2.9	0	11.8	12.6
		iutA	8.3	0	83.3	0.9	0	19.5	21.4	0	13.9	69.2	2.9	0	11.8	12.6
	Yersiniabactin (*fyuA, ybtAEPQSTUX, irp12*)	*fyuA*	50	100	100	100	40	87.8	100	71.4	58.3	94.9	100	100	51.4	78.3
		ybtAEPQSTUX	50	100	100	100	40	87.8	100	71.4	58.3	94.9	100	100	51.4	78.3
		irp1	50	100	91.7	100	40	87.8	100	71.4	58.3	94.9	100	96.6	50	77.4
		irp2	50	0	91.7	94	40	87.8	100	71.4	52.8	92.3	94.1	94.8	48.6	68
	Hemin uptake system (*chuASTUVWXY*)	*chuASTUVWXY*	0	100	100	100	100	100	100	0	100	100	100	100	44.8	78.5

Invasin	Invasion of brain endothelium	*ibeA*	0	0	0	0.9	73.3	0	0	0	0	0	0	12.1	7.5	5.3
Protectins	K1 capsule	*kspD*	16.7	100	91.7	98.3	100	100	64.3	0	91.7	82.1	97.1	100	51.9	77.4
		kspM	16.7	100	91.7	59.8	33.3	80.5	42.9	0	86.1	64.1	85.3	98.3	42.5	62
		kspT	16.7	100	0	0.9	0	0	0	0	0	0	2.9	100	11.8	20.2
	Tir domain containing protein	*tcpC*	0	0	83.3	0	0	0	0	0	0	0	97.1	0	7.1	8.7

Functional category	**Gene cluster**	**Genes**	**ST**												
			10	**1193**	**12**	**131**	**354**	**38**	**405**	**410**	**648**	**69**	**73**	**95**	**other**	**Total (%)**
Toxin	Enterotoxin	*astA*	16.7	0	8.3	2.6	13.3	14.6	0	7.1	5.6	0	5.9	0	20.3	9.3
		senB	25	85.1	58.3	80.3	0	34.1	32.1	21.4	0	66.7	44.1	0	13.2	35.9
	Hemolysin	*hlyAB*	0	4.3	91.7	55.6	0	7.3	21.4	0	0	5.1	91.2	0	12.3	22
		hlyC	0	4.3	75	54.7	0	7.3	21.4	0	0	2.6	44.1	0	0.9	15.3
	Cytotoxic necrotizing factor 1 (CNF-1)	*cnf1*	0	4.3	66.7	53.8	0	0	3.6	0	0	2.6	88.2	1.7	8.5	18.6
	Serine protease autotransporters Sat, Pic	*sat*	16.7	0	0	96.6	0	53.7	32.1	0	13.9	76.9	61.8	0	12.3	34.3
		pic	0	0	0	0	0	0	0	0	0	0	94.1	3.4	3.8	6.3
	Haemoglobin protease	*vat*	0	66	83.3	0	0	0	0	0	0	0	85.3	63.8	13.2	20.3

3.3.3. Antimicrobial resistance phenotype between blood/rectal swab isolates

We tested for phenotypic resistance to 13 antimicrobials belonging to five classes, including aminoglycosides, 3rd and 4th – generation cephalosporins, β-lactam plus inhibitor, carbapenem, trimethoprim and fluoroquinolones. Isolates exhibited the highest resistance to cotrimoxazole (STX) with 70% and 75% isolates in blood and carriage isolates respectively. Ninety-nine percent of isolates were sensitive to carbapenem (ertapenem, imipenem, meropenem) and aminoglycosides (amikacin). Resistance to 3rd – generation cephalosporins, which are used as empirical treatments against suspected BSIs in HTD, was ~ 50% (Figure 3.3).

Looking at the dominant STs, ST131 was the most prevalent ST in our blood collection (100/506, 19.76%). Its global spread has been partially attributed to its drug resistance profile that includes resistance to fluoroquinolones and third-generation β-lactams [135]. Of note the newly-emerging ST1193 clone seen in HTD, which accounted for 7.11% of cases (36/506), has the same resistance profile as ST131. Nevertheless, ST95 (57/506, 11.26%) and ST73 (34/506, 6.72%), the second and fifth prevalence STs, are generally susceptible to most antibiotics (See supplementary Tables S1 and S2 for complete list of resistance phenotypes).

Looking at all samples we observed no statistical significance between AMR phenotypes across blood and rectal swab isolates, except for combinations of β-lactam antimicrobials that include an inhibitor (amoxicillin-clavulanic acid, ticarcillin-clavulanic acid and piperacillin-tazobactam; highlighted in Figure 3.3). The most dramatic difference was seen in the case of piperacillin-tazobactam (TZP), a drug combination of extended spectrum penicillin (piperacillin) and tazobactam; tazobactam acts as an inhibitor of the β-lactamase enzyme. Fewer than 20% of BSI isolates showed reduced susceptibility to TZP and were classed as resistant, whereas >50% of all rectal swab isolates were classed as phenotypically resistant (Figure 3.3). This is important because TZP and carbapenem are used as an alternative treatment for ESBL-producing *E. coli* BSI at HTD.

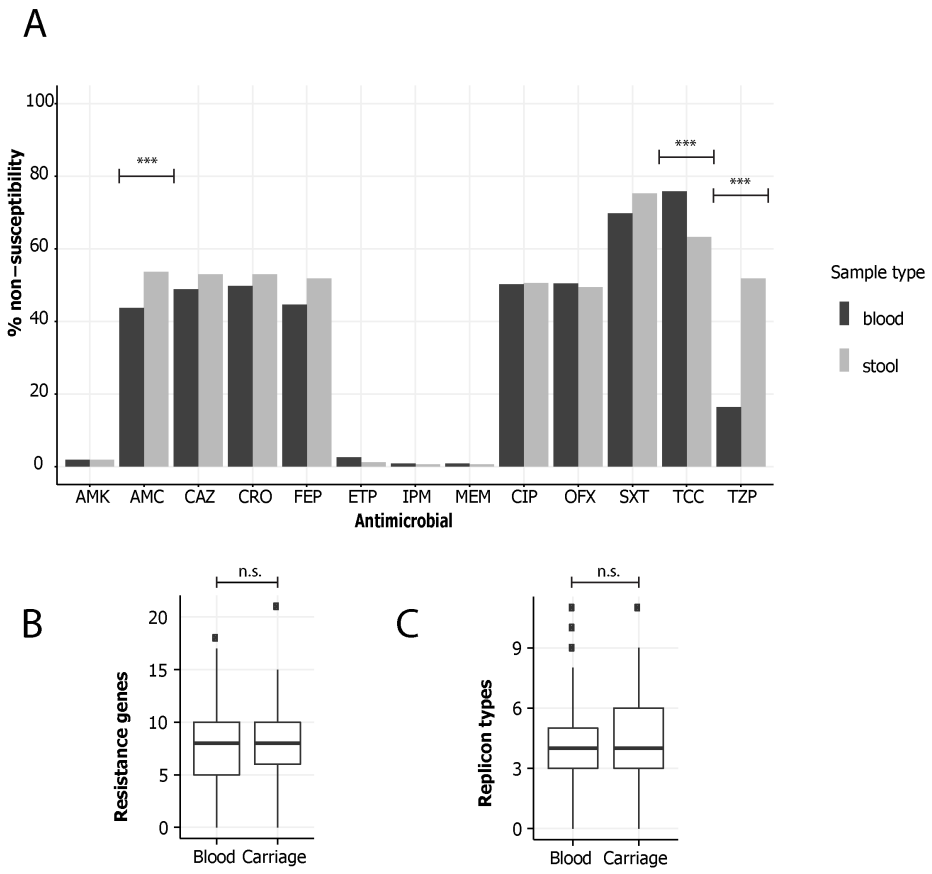

Figure 3.3: (A) The distribution of AMR resistance between blood and carriage isolate. Asterisks highlight the three antimicrobials showing differences in the levels of resistance between BSI isolates and rectal swab. (B) Total resistant genes and (C) total replicon genes found in two groups. n.s: not significant, *** $p < 0.001$; p-values were calculated using Wilcoxon test for continuous data and Chi square test for category data. Antimicrobial abbreviations: amikacin (AMK), amoxicillin-clavulanic acid (AMC), ceftazidime (CAZ), ceftriaxone (CRO), cefepime (FEP), ertapenem (ETP), imipenem (IPM), meropenem (MEM), ciprofloxacin (CIP), ofloxacin (OFX), trimethoprim-sulfamethoxazole (SXT), ticarcillin-clavulanic acid (TCC), piperacillin-tazobactam (TZP).

3.3.4. Antimicrobial resistance genes distribution in blood/ rectal swab

We screened *E. coli* genomes in this study for resistance genes and plasmid replicons that are associated with AMR to evaluate the diversity of AMR genes and plasmids in our *E. coli* collection. We found a total of 54 genes that have the potential to confer resistance to not only drugs against which these isolates exhibit a resistant phenotype (β -lactam, fluoroquinolone, *etc.*), but also to members of other classes including rifampicin, chloramphenicol, macrolide, fosfomycin, lincosamide, tetracyclin, streptomycin and colistin (Figure 3.5). The general lack of difference in phenotypic AMR profile between blood and carriage isolates was supported by comparing the total number of AMR genes and

plasmid replicon genes in BSI and rectal swab isolates: overall there was no statistical significant difference (median 8 vs 8, Wilcoxon test) (Figure 3.3).

Consistent with the phenotypic data showing that the majority of our isolates remained sensitive to carbapenem - we only found three isolates harbouring $bla_{\text{NDM-4}}$ or $bla_{\text{NDM-5}}$, and the presence of these genes corresponded to the resistance phenotype of the specific isolate (Supplementary Table S6). Although there was no difference in the profile of genes when comparing blood and rectal swab derived isolates overall, the total number of AMR genes and replicon types are substantially different between STs/BAPS lineage (Figures 3.4 & 3.5). ST131 and ST1193 contained more AMR genes than "susceptible" STs such as ST73 and ST95 (median 9 *versus* 6, p= 4.07 x 10^{-06}, Wilcoxon test). However, ST131 contain less replicon types than ST1193 (median 4 *versus* 6, p= 4.463 x 10^{-11}, Wilcoxon test).

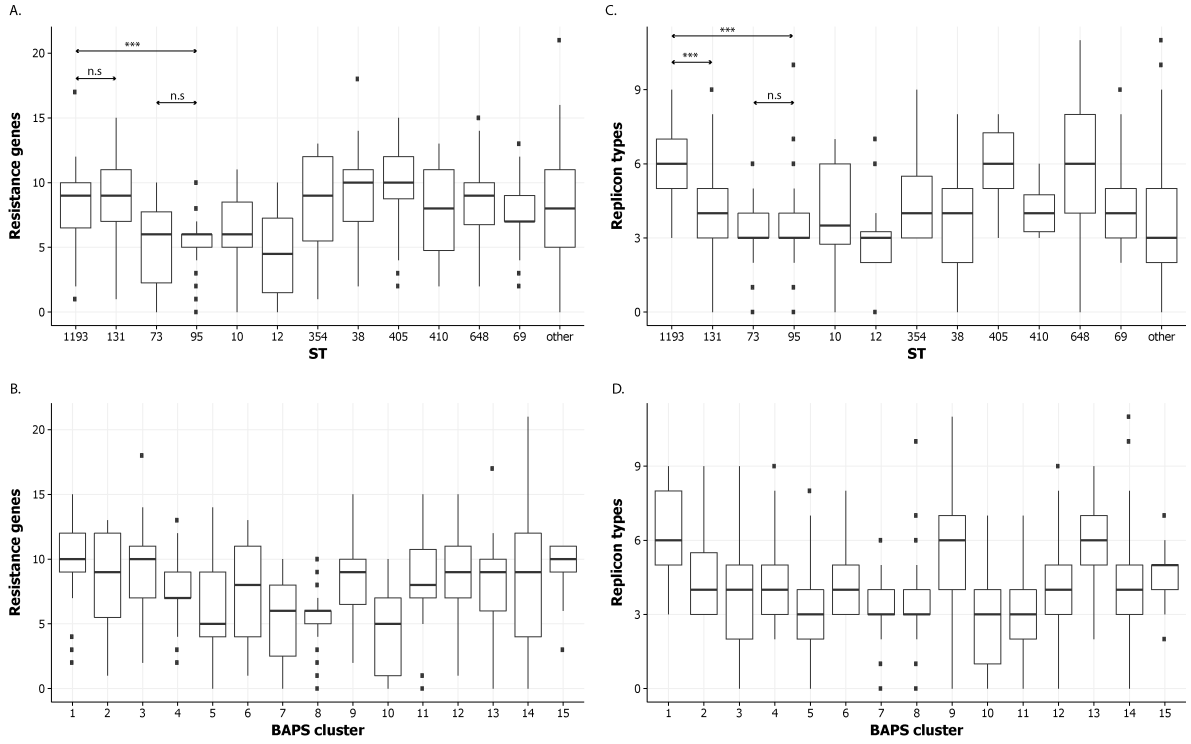

Figure 3.4: Total number of resistance genes and replicon types distributed across the most prevalent STs and among all 15 BAPS lineages in this dataset. Resistance genes found in different STs (A) and BAPs lineages (B). Total replicon types found in STs (C) and BAPs lineages (D). n.s: not significant, *** p<0.001 (Wilcoxon test)

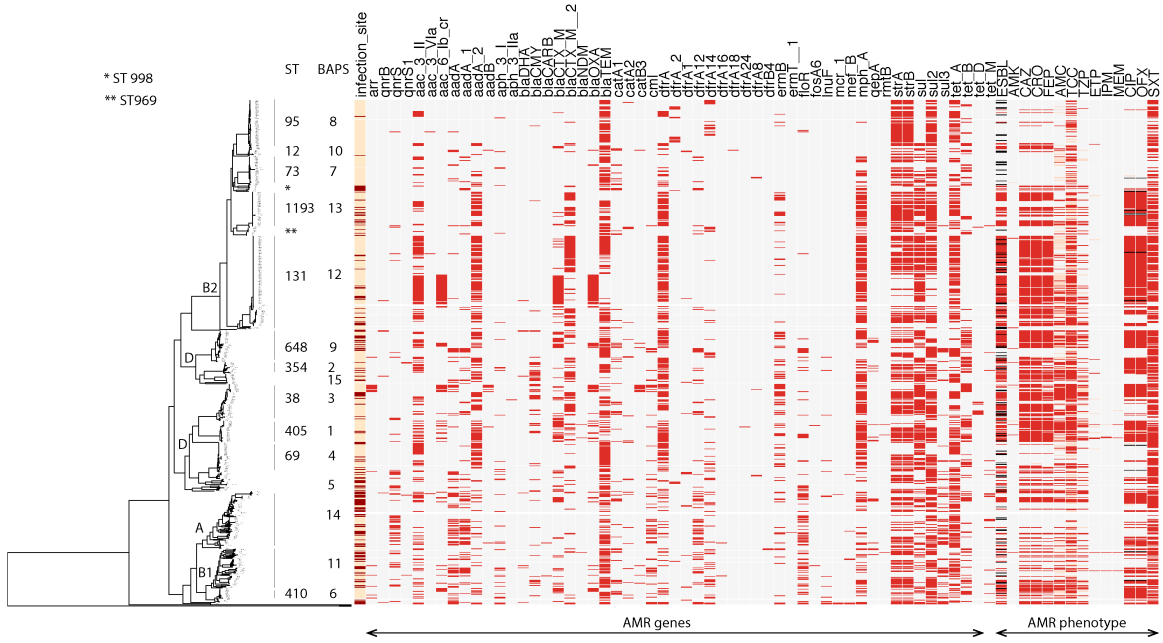

Figure 3.5: AMR genotypes and phenotypes across all *E. coli* isolates in this collection. Columns report infection site (blood/carriage, see key in Fig 2.7), red is present while white is absent, black were missing phenotype data (ESBL, CIP, OFX) since ESBL phenotypes were not assessed for isolates from 2010. Phylogroups are denoted on the branches of the phylogenetic tree.

Since extended spectrum β-lactams are one of the most important classes of antimicrobial at HTD we looked for ESBL-genes across all lineages (Table 3.3). We found that the majority of ESBL-producing isolates belonging to ST1193 carried *bla*$_{\text{CTX-M-27}}$ (27/31, 87%), whereas ST131 isolates carried different *bla*$_{\text{CTX-M}}$ gene classes such as *bla*$_{\text{CTX-M-14}}$ (33/106, 31%), *bla*$_{\text{CTX-M-15}}$ (32/106, 32%) or *bla*$_{\text{CTX-M-27}}$ (22/106, 20.8%). However, a small number of ST131 isolates contained multiple *bla*$_{\text{CTX-M}}$ genes and ST131 isolates carried the greatest range of ESBL genes, suggesting that this genetic background has the ability to acquire and maintain multiple plasmids containing multiple *bla*$_{\text{CTX-M}}$ genes. This contrasted with ST1193 which almost exclusively carried *bla*$_{\text{CTX-M-27}}$. The absence of multiple classes A (cefotaximase (CTX-M), temoneira (TEM) and sulfhydryl variable (SHV); Figure 3.5) ESBLs was also noticeable in other lineages such as the STs within BAPS lineages L10, L8 and L7. These lineages carried the *bla*$_{\text{TEM}}$ ESBL gene but, apart from essentially sporadic isolates, lacked other class A ESBL genes (Figure 3.5 and Figure 2.7 for the BAPS lineages). Importantly it is noticeable that these lineages were also largely susceptible to TZP as well as many other front-line antimicrobials (Figure 3.5).

We also detected the newly discovered *mcr*-1 in three isolates from patients attending HTD - one from an invasive isolate (ST617) and two from consecutive rectal swabs of day 0 and day 5 in one patient in ICU (Supplementary Table S6). Looking at the assembled genome data it was clear that the *mcr*-1 gene in the rectal swab isolates were located on an IncI plasmid, with the same structure as plasmid pHNSHP45 by Liu *et al* [136] (data was confirmed by visualizing *mcr-1* genes on contigs in Bandage

[128]). However, in the blood isolate, this gene was flanked by *IS*Apl1transposon and the *tnpA* gene on a large chromosomal contig, suggesting that the gene might be chromosomally integrated rather than plasmid-borne (Figure 3.6). Chromosomally integrated *mcr*-1 has been reported recently in clinical *E. coli* isolates from Vietnam and *E. coli* isolated from retail chicken meat in Switzerland - these reports highlight the easy mobilisation of *mcr-1* across different genetic locations [137, 138].

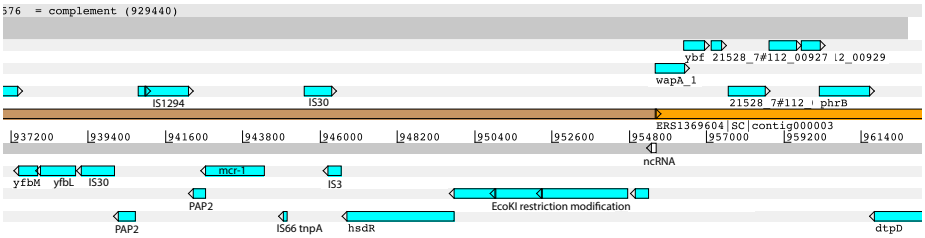

Figure 3.6: The genetic environment of *mcr-1* on a chromosomal contig from isolate 221214_63810, visualized using Artemis [139]

Table 3.3: *bla* CTX-M distribution among different STs (n = number of isolates)

CTX-M genes	ST (n)												
	10	**1193**	**12**	**131**	**354**	**38**	**405**	**410**	**648**	**69**	**95**	**other**	**Total**
*bla*CTX_M_15 + *bla*CTX_M_14	0	0	0	7	0	0	0	0	0	0	0	0	7
*bla*CTX_M_15 + *bla*CTX_M_24	0	0	0	0	0	0	1	0	0	0	0	0	1
*bla*CTX_M_15 + *bla*CTX_M_27	0	1	0	5	0	0	1	0	1	0	0	2	10
*bla*CTX_M_15 + *bla*CTX_M_45	0	0	0	2	0	0	0	0	0	0	0	0	2
*bla*CTX_M_15	1	1	2	34	1	4	15	5	8	1	1	24	97
*bla*CTX_M_3	0	0	0	0	0	0	0	0	1	0	0	0	1
*bla*CTX_M_55 + *bla*CTX_M_24	0	0	0	1	0	0	0	0	0	0	0	0	1
*bla*CTX_M_55	0	2	0	1	0	2	1	1	5	1	0	10	23
*bla*CTX_M_14	0	0	1	33	2	3	1	2	4	2	0	7	55
*bla*CTX_M_24	0	0	0	1	3	2	2	0	0	0	0	1	9
*bla*CTX_M_27	0	27	0	22	1	13	4	0	2	4	1	25	99
Total isolates	1	31	3	106	7	24	25	8	21	8	2	69	305

Table 3.4: Plasmid replicon type identified in *E. coli* collection. (n = number of isolates)

	ST (n)													
Replicon	**10**	**1193**	**12**	**131**	**354**	**38**	**405**	**410**	**648**	**69**	**73**	**95**	**other**	**Total**
col156	3	46	0	97	1	21	13	4	19	28	15	4	34	285
col8282	2	8	1	15	2	2	3	0	13	7	1	4	13	71
colE10	0	0	0	0	0	0	0	0	0	0	0	0	3	3
colRNAI	6	14	1	14	4	2	1	8	5	4	14	31	51	155
col_BS512	2	44	1	10	3	3	15	0	10	7	3	3	21	122
col_BS512_1	2	44	1	10	3	4	18	0	10	6	2	1	21	122
col_kphs6	0	0	0	0	0	0	0	0	0	0	0	0	4	4
col_mg828	0	9	1	10	2	1	12	1	23	7	6	9	12	93
col_mp18	1	1	0	1	1	0	0	0	0	0	1	0	2	7
colPVC	1	0	0	0	0	0	1	0	4	2	0	0	5	13
incA/C	0	0	0	0	0	0	0	0	0	0	0	0	2	2
incB/O/K/Z	2	2	2	13	2	0	2	0	5	4	0	3	6	41
incFI	8	3	9	69	2	31	24	14	21	26	18	55	127	407
incFIA	3	47	3	69	14	18	20	6	29	11	9	19	43	291
incFIA_hi1	0	0	0	0	0	0	1	0	1	0	0	0	17	19

incFIB	0	0	0	1	11	3	1	0	5	0	2	0	2	25
incFIB_ap001918	8	45	11	108	4	33	25	13	26	38	22	56	120	509
incFIB_K	0	0	0	0	0	0	0	0	0	0	1	1	19	21
incFIB_phcm2	0	0	0	0	0	0	0	0	0	0	0	0	2	2
incFIB_plf82	0	0	0	1	0	1	0	0	0	0	0	0	2	4
incFII_p	0	0	3	53	0	9	9	0	12	1	6	1	29	123
incHI1a	0	0	0	0	0	0	0	0	1	0	0	0	6	7
incHI1b	0	0	0	0	0	0	0	0	1	0	0	0	4	5
incHI2	0	0	0	0	0	1	0	0	1	1	0	0	4	7
incHI2a	0	0	0	0	0	1	0	0	1	1	0	0	5	8
incI1	2	8	2	7	7	10	8	6	4	0	0	1	42	97
incI2	0	0	0	1	0	1	0	0	1	0	0	0	5	8
incL_M	0	0	0	0	0	0	1	0	1	0	0	0	1	3
incL_M_pmu407	0	0	0	0	0	0	0	0	1	0	0	0	1	2
incN	0	0	0	2	0	1	2	1	0	1	0	0	6	13
incN1	0	0	0	0	0	0	0	0	0	0	0	0	1	1
incQ1	0	6	3	3	8	0	1	1	4	9	2	14	25	76
incR	0	1	0	0	0	0	0	0	0	1	0	0	12	14

incX1	3	2	0	0	0	0	0	0	5	3	0	2	30	45
incX1_1	3	1	0	0	0	0	0	0	2	0	0	1	27	34
incX2	0	0	0	0	0	0	0	0	0	0	0	0	1	1
incX3_pec14	1	0	0	0	0	0	0	0	0	0	0	0	0	1
incX4	0	0	0	2	0	3	0	0	0	0	0	0	3	8
incX4_1	0	0	0	0	0	0	0	0	0	0	0	0	2	2
incY	0	0	0	2	3	4	5	2	3	2	2	3	17	43
p0111	1	1	0	6	0	1	11	1	3	0	0	2	15	41
psl483	0	0	0	0	0	0	0	0	1	0	0	0	1	2
Total	12	47	12	117	15	41	28	14	36	39	34	58	212	665

3.4. Discussion

In this chapter, we compared the distribution of different virulence related and antimicrobial resistance gene contents for the randomly sampled BSI and rectal swab isolates sequenced in this study. The most significant association between its presence and the ability to cause invasive disease was for P fimbriae (OR > 3, $p \leq 10^{-8}$). P fimbriae is an important colonization factor that enhances early establishment of *E. coli* in the urinary tract. P fimbriae are often associated with UPEC causing pyelonephritis. The reciprocal regulation of P fimbriae, type 1 fimbriae and flagella help UPEC to ascend from bladder to kidneys [132]. However, the correlation between P fimbriae and virulence remains inconclusive in literature [30].

After attachment to host tissue and following colonisation, in order to grow and proliferate, *E. coli* requires a siderophore system to acquire iron from the host, where free iron is maintained in an incredibly small concentration (10^{-24} M) [133]. Four siderophore systems have been studied in *E. coli*: enterobactin (*fepABCDG*, *entDFCEBA*, *fes*, *entS*), salmochelin (*iroNEDCB*), yersiniabactin (*ybt*, *fyuA*, *irp*) and aerobactin (*iucABCD*, *iutA*). The human host has many strategies to prevent heme acquisition from pathogenic bacteria, such as Lipocalin-2 which can bind and inactivate bacterial enterobactin [142] and IL-22 which induces the production of the heme scavenger haemopexin, further limiting host iron availability [143]. Therefore, possessing more than one siderophore system simultaneously may give an *E. coli* isolate increased fitness and a competitive advantage relative to other *E. coli* strains.

Several virulence genes showed lineage restriction. Phylogroup B2 isolates which include ST131, ST73, ST95 and ST1193 carried the highest number of genes characteristic of ExPECs, as would be expected from previous studies. This is consistent with data shown in Chapter 2 which showed an association between Phylogroup B2 and gender as well as BSIs. Women are known to be highly susceptible to UTI infections by uropathogenic members UPEC members of the ExPEC. However, it is clear that different lineages employ different fimbrial systems for attaching and colonising the host, and may not necessarily use the same pili or retain more than one due to possible functional redundancy. ST73 in our study possessed more *Pap* genes than ST131, as well as having additional fimbriae such as S1 fimbriae (*sfa* operon) and F1C fimbriae (*foc* operon). This might enhance its ability to bind to human bladder and kidney epithelium cells [144] as well as to form biofilms [145]. The most striking observation was that *tcpC* can only found in ST73 (97%) and in some less prevalent STs such as ST12, ST969 and ST998. TcpC is a secreted protein that binds to myeloid differentiation factor 88 (MyD88) and impedes Toll-like receptor (TLR4) signalling pathways in innate immunity, therefore promoting intracellular bacterial survival of pathogen [146]. The ability to colonise different cell types, together with having a sophisticated molecular strategy such as *tcpC* to subvert host defence, and harbouring a toxin such as *cnf1* to cause cell apoptosis [147], might explain why ST73 was able to emerge as a UPEC clone, with potential to translocate into the blood.

The high prevalence of AMR genes in both BSI and carriage isolates highlights the urgent threat posed by multidrug resistant *E. coli* that are resistant to all major antibiotics, and the challenge that the clinical management of community acquired *E. coli* BSIs presents. Luckily genes conferring resistance to last line antimicrobials, bla_{NDM} and *mcr*-1, were rare and only sporadically distributed in our *E. coli* collection at the time of testing.

One other important observation made here was for TZP. TZP is an intravenous combination antibiotic that is not available in the community. TZP, like clavulanic acid, is a combination therapy including a β-lactam inhibitor (tazobactam), inhibiting the β -lactamase produced by ESBL-producing bacteria, so that the effectiveness of ESBL antibiotic combined with it, in this case piperacillin, is preserved. However, TZP is more efficient than clavulanic acid at inhibiting a wider range of ESBLs. Resistance mechanisms for TZP are not well defined but the efficiency of the inhibition is thought to be affected by the type of the ESBL produced, since the inhibitor binds to and blocks the enzyme active site [148]. Sugumar *et al*., (2014) have suggested that possession of bla_{OXA} in combination with other *bla* genes leads to high resistance to the Tazobactam treatment. Consistent with this from looking at the carriage isolates there was a significantly higher proportion of bla_{OXA} and bla_{CMY} genes present in TZP resistant isolates than TZP sensitive isolates (25.1% *versus* 8%; 20.4% *versus* 5.2%, $p < 0.001$, Chi square test) suggesting that Tazobactam is less efficient at inhibiting the products of bla_{OXA} and bla_{CMY} than other genes such as bla_{TEM} (48.5% *versus* 61.8%, p=0.002, Chi square test). Figure 3.5 also shows that phylogroup B2, BAPS lineages L10, L8 and L7 (ST95, ST73 and ST1193) are dominated by isolates carrying bla_{TEM} genes with only sporadic possession of other *bla* genes. Phenotypically these isolates were largely susceptible to TZP (Figure 3.5) and these susceptible isolates represent a significant proportion (28%) of all BSI isolates sampled in this study. This may explain why the proportion of blood isolates showing resistance to TZP was significantly lower than rectal swab isolates (Figure 3.3), which are more strongly associated with more diverse, less frequently sampled STs from lineages possessing multiple types of *bla* gene. It is not clear if this could be associated with the intravenously delivery with blood isolates exposed to high therapeutic doses compared to gut isolates due to differential drug penetration. Although as we have shown common BSI STs are present in rectal swab samples.

If the TZP-sensitive ST131 isolate numbers (87/100) from blood (the most prevalent ST causing BSI at HTD) are combined with ST95, ST73 and ST1193 then these 4 STs represent 48% of all BSIs sampled in this study. What's difficult to explain from the above is that a high proportion of ST131 isolates carry bla_{OXA} in combination with other *bla* genes. Within the ST131 TZP-sensitive isolates from blood 61 and 26 are bla_{OXA} negative or positive, respectively. Of the ST131 TZP-resistant isolates 6 and 7 are bla_{OXA} negative or positive, respectively. In addition, all of the bla_{OXA} positive isolates also possess either bla_{CTX} or bla_{TEM}. What is more, by performing a genome wide association study using SCOARY [149] which takes the gene presence/absence data from ROARY (see 2.2), then uses the phenotype to

work out gene distribution between 2 groups (TZP resistant or sensitive), in addition to *bla* genes several other genes showed a positive association, including *acrAB* (multidrug efflux pump), *mphA* (multidrug transporter), *ompC* (outermembrane protein linked to resistance) and the multiple antibiotic resistance protein MarB. Hence, TZP resistance appears to also be associated with general mechanisms of resistance and likely to be multifactorial like macrolide resistance. This may imply genetic background is as important as the possession of specific *bla* genes for resistance.

What is clear is that there is that the B2 lineages that were dominant in our BSI isolate collection remain significantly more susceptible to TZP than other phylogroups more associated with rectal swab. We have shown that once established in a community the relative proportion of different STs is stable over time. If we can use this data to identify the less well understood mechanisms of resistance, such as to TZP, we may be able to reach a more proactive state matching any changes in disease patterns in a hospital with flux in the bacterial population and then acquisition of different mechanisms of resistance and change interventions accordingly. Unfortunately, there was insufficient time to pursue this line of investigation further.