Within-host evolution and immigration of *Escherichia coli* in the human gastrointestinal tract

Ojas V. A. Dixit

A thesis submitted for the degree of Master of Philosophy

Research School of Biology
The Australian National University
February 2017
DECLARATION

This is to state that the works included in this thesis titled “Within-host evolution and immigration of *Escherichia coli* in the human gastrointestinal tract” is my original work. As per my knowledge, this thesis has not been previously submitted at the Australian National University or any other University. To the best of my knowledge and belief, this thesis contains no material previously published or written by anyone else, except where due reference is made in the text.

Ojas V. A. Dixit

February, 2017
ACKNOWLEDGEMENTS

I would like to express my gratefulness to my supervisor, David Gordon for guiding me throughout my research. I have learned a lot from his critical analysis and constant feedback. Without his support and constant feedback, it would not be possible to finish this thesis efficiently in such a short duration. I would also like to thank him for his patience in explaining the many questions I had.

I am very much indebted to my supervisory committee member Claire O’Brien for always being very helpful regarding my research progress and thesis writing. Furthermore, I would sincerely like to thank Paul Pavli who always showed his interest in my research and gave me feedback.

My lab work would not have been possible without the excellent technical support of Samantha Burn. I appreciate her for helping me in finishing my lab work on time and also for teaching instrument handling technique and timely guidance in the lab. Special thanks to Belinda as I learned a lot from the numerous discussions with you. I have been blessed with a friendly and cheerful fellows Judith, Buddhie, Mahboobeh and Angelingifta. I would also like I forget to extend a special vote of thanks to all my friends, especially Amit, Neel and Shrey for their helping hand and for keeping my smile alive throughout the research.

Last but not the least important, I would thank my family members for their encouragement throughout my life.
Several studies have shown that when a human host harbours two or more strains of *E. coli*, the second strain is significantly more likely to be a member of the same phylogroup than to be a member of a different phylogroup. Such an outcome may be the consequence of a within host evolution event or due an independent immigration/establishment event. To determine the relative importance of these two types of events in determining *E. coli* diversity in a host, a large collection of *E. coli* that consisted of up to 100 isolates recovered from each of 69 patients undergoing colonoscopies was used. Whole genome sequence data was available for 174 isolates selected to represent one example of every REP-fingerprint type identified in a patient.

Sequence type characterisation and single-nucleotide polymorphism analysis revealed that 83% of the strains observed in the host population were a consequence of immigration/establishment events. Restricting the analysis to those hosts harbouring two or strains belonging to the same phylogroup revealed that in about half of these cases the presence of a second strain belonging to the same phylogroup was the consequence of an independent immigration/establishment event. Single nucleotide polymorphism analysis coupled with the assumption of a mutation rate of 1.1 nucleotides/year indicated a residence time for those strains inferred to have evolved within the host of 1.2 years and a maximum estimated residence time of 11 years.

This study has shown that when a host harbours two strains of the same phylogroup, then in about half of such cases, this is due to the immigration and establishment of strains and not within host evolution. Thus, the results of this study show that despite hosts being regularly exposed to a diversity of *E. coli* through the food that they eat, factors related to the host, at least in part, determines what *E. coli* strains succeed in establishing.
# TABLE OF CONTENTS

ACKNOWLEDGEMENT.............................................................................................................IV
ABSTRACT...............................................................................................................................VI
TABLE OF CONTENTS...........................................................................................................VII
LIST OF FIGURES................................................................................................................IX
LIST OF TABLES..................................................................................................................X

## CHAPTER 1: INTRODUCTION

1.1 Gut microbiome of humans.............................................................................................2
1.2 *Escherichia coli* in the gastrointestinal tract.................................................................3
1.3 Clinical importance .........................................................................................................4
1.4 Genetic structure of *E. coli*.............................................................................................4
1.5 Phylogroups of *E. coli*....................................................................................................5
1.6 Typing of *E. coli*............................................................................................................5
  1.6.1 Phenotypic methods ....................................................................................................5
  1.6.2 DNA-based methods ..................................................................................................6
1.7 Factors influencing *E. coli* diversity: Among host variation...........................................8
1.8 Factors influencing *E. coli* diversity in humans: Within host variation .........................9
1.9 Aim of the study...............................................................................................................11

## CHAPTER 2: MATERIALS and METHODS

2.1 Strain collection...............................................................................................................13
2.2 Whole-genome sequencing ...........................................................................................13
2.3 Microbial genome assembly.........................................................................................13
2.4 Basic characterization........................................................................................................14
2.5 CSIPhylogeny - Variant finding.....................................................................................14
2.6 Annotation of variants.....................................................................................................14

## CHAPTER 3: RESULTS

3.1 Host characteristics........................................................................................................16
3.2 SNP detection.................................................................................................................16
3.3 Defining isolates as distinct..........................................................................................18
3.4 Strain diversity within hosts..........................................................................................19
3.5 Immigration/Establishment \textit{versus} within-host evolution.........................................21
3.6 Genetic changes and within host evolution...................................................................24

CHAPTER 4: DISCUSSION

4.1 Caveats..........................................................................................................................29
4.2 The host as a determinant of \textit{E. coli} clonal diversity ...............................................31
4.3 Within-host evolution....................................................................................................32
4.4 Measuring evolution of \textit{E. coli} strains.................................................................33

REFERENCES.........................................................................................................................35

APPENDIX.............................................................................................................................43
LIST OF FIGURES

Figure 3.1. Number of SNPs detected using the detection method CSIphylogeny (Kaas et al., 2014) as a function of base call and read mapping quality, and coverage.............17

Figure 3.2. Distribution of the number of SNP differences between strains isolate from the same host and belonging to the same ST.................................................................22

Figure 3.3. Phylogenetic relationships among all the sequence type 73 strains observed in this study. The first number in the strain name denotes the patient from which the isolates were taken. Patient 23 harboured 3 ST73 strains representing two different immigration/establishment events, while one of the strains gave rise to a within host evolution event..............................................................................................................23
Table 3.1. Phylogroup and sequence type (ST) of the distinct strains detected in each patient. A migration event denotes a case when the patient harbours two or more strains representing different STs or belonging to the same ST but differing by more than 1000 SNPs. A within-host evolution event denotes two or more strains of the same ST than differ by more than 2 but less than 50 SNPs.

Table 3.2. The average number of SNPs detected among strains belonging to the same sequence type (ST) but isolated from different patients.

Table 3.3. Fraction of cases where the presence of two or more strains within a host was a consequence of within-host evolution versus migration/establishment as a function of the phylogroup membership of the strain.

Table 3.4. Mutations occurring in strains inferred to have been derived as a result of within host evolution.

Table 4.1. Comparison of relative abundance of phylogroups distribution between Gordon et al. (2015) and this study.
Chapter 1

INTRODUCTION
1.1 Gut microbiome of humans

The gut microbiome consists of a diverse assemblage of bacteria and bacteriophage (Gorbach, S. L., 1996; Blotiere et al., 2013; Candela et al., 2010). The gut microbiota consists of more than 500 bacterial species and these diverse populations of bacteria play a significant role in human health (Eckburg et al., 2005; Dore and Blotiere, 2015; Zhang et al., 2015). For example, gut bacteria produce butyrate, the main energy source of colonocytes, and synthesize vitamin K, which we can’t accomplish on our own. The gut bacteria may also promote barrier integrity, and prevent microbial pathogens from entering the mucosal layer (Zhang et al., 2015). As well, the gut microbiota has the ability to digest substrates such as cellulose, which humans are incapable of digesting on their own, and as a consequence, the gut microbiota plays important role in determining the host’s energy intake. The gut microbiome has been implicated in obesity (Turnbaugh et al., 2006), as well as many other conditions, including diabetes (Paun et al., 2017), inflammatory bowel disease (Kostic et al., 2014), and colon cancer (Vogtmann et al., 2016; Sobhani et al., 2013).

Metagenomic analysis of total faecal DNA showed that bacterial genes make up 99% of the normal gut microbiome, with genes from viruses and fungi making up the remaining 1% (Qin et al., 2010; Zhang et al., 2015). Bacteria are taxonomically classified into four phyla: Bacteroidetes, Firmicutes, Actinobacteria and Proteobacteria. Cells of species belonging to Bacteroidetes and Firmicutes represent 95% of the microbial biomass. Microbial diversity and density increase along the gastrointestinal tract, from the stomach to colon (Zoetendal et al., 2008), with each region containing different species and concentrations of bacterial cells (Tiihonen et al., 2010). Few bacteria (Helicobacter pylori etc.) can survive the acidic conditions of the stomach and upper small intestine, and bacterial cell densities are relatively low (10^3-10^4 bacterial cells/g of contents) compared to the lower small intestine (10^3-10^4 bacterial cells/g), and colon (10^{10}-10^{12} bacterial cells/g) which contains the most diverse community of bacteria (Zhang et al., 2015; Tiihonen et al., 2010). The inability of many bacteria to survive in the stomach and upper small intestine is likely due to the more favourable pH in the lower small intestine (pH 6-7.5) and colon (pH 5-7) compared to these regions (pH
1-3), and the absence of large amounts of bicarbonate, digestive enzymes, proteases, lipases, and bile, which are secreted in the upper small gut. The colon also has a slower transit time than the small intestine (Gordon et al., 2015). All of these factors result in different gut regions harbouring distinct bacterial species.

1.2 *Escherichia coli* in the gastrointestinal tract

*E. coli* belongs to the phylum Proteobacteria, family *Enterobacteriaceae*, and comprises approximately 0.1% of the microbial mass in healthy humans. *E. coli* are gram-negative, non-sporulating, facultative anaerobes that inhabit the lower gastrointestinal tract and faeces of warm-blooded animals, birds, and reptiles (Tenaillon et al., 2010; Gordon and Cowling, 2003; Smati et al., 2015).

*E. coli* is the most studied prokaryotic organism. The species is easily grown in the laboratory and is often used for recombinant DNA work. *E. coli* cells do not survive long outside a host making them an ideal faecal contamination indicator (Boehm and Sassoubre, 2014). In 1884, the German-Austrian pediatrician, Theodor Escherich, discovered *E. coli* in the stool of healthy patients. He suggested that *E. coli* colonises the gastrointestinal tract of humans 4-18 hours after birth (Alm et al., 2011).

*E. coli* resides in the mucus layer covering the intestinal epithelial barrier, which provides it with nutrients to grow and divide (Tenaillon et al., 2010). The diversity, population dynamics, and cell densities of *E. coli* within the gastrointestinal tract depends on host size, gut morphology, gut dynamics, diet, the wider microbial community (O’Brien & Gordon, 2011), and the dominant strain of *E. coli* present (Gordon et al., 2015; Ley et al., 2008; Tenaillon et al., 2010). *E. coli* is more likely to be isolated from mammals than from birds, as well as from larger rather than from smaller hosts, and is less likely to be isolated from carnivores compared to herbivores or omnivores (Gordon and Cowling, 2003).
1.3 Clinical importance

Although *E. coli* comprises a small percentage of the total gut microbiota, it has a significant role in human health (Donnenberg, 2002). *E. coli* is the ideal candidate to study commensalism and pathogenicity, because while most strains are commensal, some strains cause disease, resulting in approximately two million deaths per year (Tenaillon *et al.*, 2010; Massot *et al.*, 2016). Pathogenic strains of *E. coli* can cause serious food poisoning in humans through the release of toxins, while others result in extra-intestinal manifestations. *E. coli* are responsible for the majority (~90%) of urinary tract infections (UTI) (Johnson *et al.*, 2003), and may also cause peritonitis, neonatal meningitis, mastitis, pneumonia, septicemia, and other infections (Johnson and Russo, 2002). The ability of a strain to cause disease depends mainly on its virulence properties, which elicit a host immune response, and include antigens (O, K, and H antigen), toxins, and adhesins.

*E. coli* has distinct pathovars which cause disease. Those pathovars are broadly classified into diarrhoeagenic or extraintestinal *E. coli* (ExPEC) (Croxan and Finlay, 2010). These pathovars are sub-classified into enteropathogenic *E. coli* (EPEC), enterohaemorrhagic *E. coli* (EHEC), enteroinvasive *E. coli* (EIEC), enterotoxigenic *E. coli* (ETEC), enteroaggregative *E. coli* (EAEC) and diffusely adherent *E. coli* (DAEC). While other strains are uropathogenic *E. coli* (UPEC) or cause neonatal meningitis (NMEC) (Croxan and Finlay, 2010; Baker K. S., 2015)

1.4 Genetic structure of *E. coli*

*E. coli* is a diverse species, and commensal and pathogenic strains differ in their genetic background and content. The average number of genes in an *E. coli* strain is ~4700. Over 18,000 different orthologous genes have been detected in the species, but only ~2,000 genes are common to all strains (core genes) of *E. coli* (Touchon *et al.*, 2009). The mutation rate of *E. coli* (~1 mutations/year/genome) is much lower compared to species like *Klebsiella pneumoniae* (~10 mutations/year/genome), or *Staphylococcus aureus* (~8 mutations/year/genome) (Didelot *et al.*, 2016).
1.5 Phylogroups of *E. coli*

*E. coli* are classified into phylogroups based on their genomic content. There are four main phylogroups, designated A, B1, B2, and D, and four relatively rare phylogroups: C, E, F and Clade I (Tenaillon *et al.*, 2010). Strains belonging to the different phylogroups vary in the size of their genomes, propensity to cause disease, ecological niche, and life-history traits (Archer *et al.*, 2011). B2, F and D strains are more likely to cause extra-intestinal infections and A, B1, C or E strains, but most diarrheal causing *E. coli* are derived from phylogroups A, B1, and E.

1.6 Typing of *E. coli*

Numerous phenotypic and DNA-based typing methods have been developed for *E. coli*. Although, phenotypic methods cannot be used to determine phylogenetic relationships, some of the DNA-based methods can be used infer the phylogenetic relationships among different strains (Gordon D. M., 2010).

1.6.1 Phenotypic methods

Typical phenotypic methods like serotyping, antibiotic resistance and biochemical profiling have the potential to discriminate the *E. coli* strains. Serotyping of *E. coli* is determined by somatic (O), capsular (K) and flagellar (H) antigens. Particular O antigens are associated with pathogenic strains, therefore O typing *E. coli* enabled these strains to be identified (Orskov, F. and Orskov I., 1992). Strains of the same serotype do not always share a common ancestor. For instance, Mora *et al.* (2009) found that isolates with an O1: H7 serotype are from two different phylogroups, where 85% were from phylogroup D and 15% were from B2. However, O-typing is expensive, complex, and time consuming. Therefore, allele-specific PCR methods were developed by Clermont *et al.* (2007) enabling isolates to be cost effectively screened for the most clinically significant O-types. WGS-based in-silico serotyping is now the more common method to determine serotypes, as the method is reproducible and has a better power of discrimination (Joenson *et al.*, 2015).

*E. coli* is the phenotypically diverse bacterial species, and while collectively the species can exploit many different substrates as carbon and nitrogen sources, there is
substantial among strain variation. For example, biochemical profiling of 692 *E. coli* isolates from humans, mammals and the environment revealed 322 distinct profiles for 31 biochemical reactions (Walk et al., 2009). Antibiotics have been used widely to treat diseases in recent decades and this has given rise to many antimicrobial resistant species, including *E. coli*. To identify these resistant *E. coli* strains, antimicrobial resistance profiling can be helpful to find the level of resistance of a particular strain. For example, human (n=118) and Australian vertebrate (n=229) isolates yielded 25 and 13 resistant profiles for 15 antibiotics respectively (Gordon D. M., 2010).

### 1.6.2 DNA-based methods

Currently, many DNA-based methods are in use: randomly amplified polymorphic DNA (RAPD), repetitive extra-genic palindromic PCR (REP-PCR), multi-locus sequence typing (MLST), pulsed-field gel electrophoresis (PFGE), single nucleotide polymorphism (SNP) genotyping, and others.

RAPD primers are a random sequence of, usually, ten nucleotides and RAPD assays are a cost effect for discriminating *E. coli* isolates, with the degree of discrimination achieved, generally increasing with the number of primers used. Rep-PCR (primers containing repetitive sequence) is a method that can also achieve a high degree of discrimination. However, comparison among different REP-PCR primers suggests that not all REP primers achieve the same levels of discrimination. Mohapatra and Mazumder (2008) found that primers such as the enterobacterial repetitive intergenic consensus primer (ERIC) has less discrimination ability than (GTG)$_5$ primer. Although best for among strain differentiation, REP-PCR results also yield some information concerning the phylogenetic relationships of the strains being compared.

MLST was for many years considered the gold standard for typing bacteria (Maiden et al., 1998). The method is based on Sanger sequencing a 400-500 portion of 7 to 8 housekeeping genes. Each distinct sequence for a gene is defined as a different allele, and the particular combination of alleles found for an isolate defines its sequence type (ST). An extension of MLST is multi-locus sequence analysis (MLSA), where by the nucleotide sequence data for each gene is concatenated and used to infer a phylogenetic
tree (Larsen et al., 2012). There are currently three MLST schemes in use for *E. coli*: First, EcMLST database for pathogenic *E. coli* MLST available at http://www.shigatox.net/ecmlst/cgi-bin/index (Qi et al., 2004). EcMLST database was initially created to characterize Shiga-toxin producing *E. coli*. Second, a 7 gene scheme was developed by Wirth et al. (2006) and is available at http://mlst.warwick.ac.uk/mlst/dbs/Ecoli and the recently developed 8 gene scheme is described by Jaureguy et al. (2008) which is available at http://bigsdb.pasteur.fr/perl/bigsdb/bigsdb.pl?db=pubmlst_ecoli_seqdef_public. By far the largest database is the Warwick Enterobase scheme, which currently has MLST data for over 50,000 isolates (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli).

Pulse field gel electrophoresis (PFGE), is a method whereby DNA is isolated from a culture and cut with a restriction enzyme. The resulting fragments are resolved on an agarose gel where the voltage is switched among three directions, thereby allowing better resolution of large fragments. PFGE has more discrimination power than the MLST and is easier to interpret than typing PCR-based typing methods. However, PFGE is labor intensive and expensive, and the method is falling into disuse.

SNP genotyping is an excellent method for typing bacterial strains. For example, 13 SNPs from the 5 genes (*trpA, trpB, putP, icdA*, and *polB*) of *E. coli* (n=30) strains were selected as a reference for grouping and subgrouping of *E. coli* (n=65) strains (Hommais et al., 2005). The SNP data was used to construct a phylogenetic tree (parsimony) that demonstrated that SNP genotyping could be suitable for phylogenetic analysis because the outcome was concordant with MLST typing (Hommais et al., 2005). This approach was successfully used by Zhang et al. (2006), who illustrated the evolutionary relationship between Shiga toxin-producing *Escherichia coli* (STEC) that are responsible for bloody diarrhea and hemolytic uremic syndrome.

The advent of high throughput sequencing methods and rapidly declining library preparation and sequencing costs has meant that whole genome sequencing (WGS) is replacing many previous typing methods. There are numerous web-based tools and software available to analyze WGS data (Clermont et al., 2015). For example, the Center
for Genomic Epidemiology (CGE) developed a web-based service which is publically available at [http://www.genomicepidemiology.org/](http://www.genomicepidemiology.org/). CGE platform provides tools for detecting virulence and antimicrobial resistance genes, sequence type determination, serotyping, and phylogenetic analysis.

### 1.7 Factors influencing *E. coli* diversity: Among host variation

The probability of detecting *E. coli* in a mammal varies with the host body size and diet. *E. coli* is rare in insectivorous bats; and among the carnivorous marsupials, the probability of detecting *E. coli* increases with host body mass. Among non-human vertebrates, B1 strains are most prevalent in fish, reptiles, birds and mammalian carnivores, while B2 strains are most prevalent in hosts with a hindgut fermentation chamber (Gordon & Cowling, 2003). Phylogroup D strains are more likely to be isolated from birds and mammals than from other vertebrates. Among humans, B2 strains are most likely to be observed in people residing in developed countries in the temperate regions of the world, while A and B1 strain predominate in people residing in tropical regions and developing countries (Escobar – Páramo et al., 2004; Skurnik et al., 2008). The rate at which the material moves through the gut varies with host diet and body mass. Transit times tend to increase with host body mass. While animals having a largely plant-based diet tend to have slower gut transit times, largely due morphological changes associated with a plant-based diet: the need for a fore or hind-gut fermentation chamber. The results of the survey studies indicate that the interactions among hosts diet, gut morphology and dynamics are important determinants of the genetic structure of *E. coli* populations in different host species. Experimental studies support the notion that diet and gut dynamics influences the diversity of *E. coli* present in the host.

O’Brien and Gordon (2011) showed that in rats fed on diets varying in their amount of non-fermentable fibre *E. coli* cell densities varied with the turnover rate of the gut. In addition, in animals on high fibre diets, animals where B2 strains dominated had lower cell densities than when an animal’s *E. coli* population was dominated by a B1 strain (O’Brien and Gordon, 2011). In another experiment, it was found that as the difference between liquid and particle retention times increased, *E. coli* faecal cell densities
decreased, while the likelihood that an animal’s dominant *E. coli* strain possessed a gene involved in adhesion (agn43) increased (Blyton et al. 2015).

### 1.8 Factors influencing *E. coli* diversity in humans: within host variation

Studies examining human faecal isolates of *E. coli* typically detect 1-3 genotypes per host although as many as 9 or more genotypes have been detected (Alm *et al.*, 2011). Studies of pigs and humans indicate that *E. coli* strain diversity based on faecal isolates may underestimate the actual number of strains present in the host. The data suggest that there are strains present in the small intestine that are not detected in faeces. There is also data from *E. coli* isolated from pigs that indicates the genotypic characteristics of strains isolated from different regions of the gut may also vary (Dixit *et al.*, 2004; Abraham *et al.*, 2012). Although tissue tropism is well-recognized phenomena among diarrheal *E. coli* (Fitzhenry *et al.*, 2002), with some diarrheal types targeting cells are in the terminal ileum and others colon cells (Donnenberg, M., 2013). O157:H7 is a diarrheal pathogen of humans, but a commensal of cattle. In cattle, O157:H7 is more likely to be detected in the rectum than in other regions of the gut or in the faeces (Croxan and Finlay, 2010). It is not clear how common tissue tropism is among non-diarrheal strains of *E. coli*. In a study that isolated *E. coli* from biopsies taken from 5 regions of the human gut (terminal ileum, ascending, transverse and descending colon, and rectum) there was no indication that strain diversity changed with the biopsy location (Gordon *et al.*, 2015). However, there were strains detected in either the terminal ileum or rectum that were not detected in any other gut region, suggesting that tissue tropism might exist for strains inhabiting the human gut (Gordon *et al.*, 2015).

Some studies have also indicated that human host age or sex plays a role in determining the diversity and genetic structure of *E. coli* in a person. The phylogenetic membership of the dominant strain has been shown to change with host age (Gordon *et al.*, 2005; Vollmerhausen *et al.*, 2011), while the number of strains detected per host has been shown to increase with host age (Vollmerhausen *et al.*, 2011). These observations accord with the non-human mammal survey studies and experimental studies with regard to the importance of gut dynamics. In human females, small intestine and colon transit times are 50% slower than males. Moreover, transit times change as humans
age, with small intestine transit times decreasing and large intestine transit times increasing with age (Graff et al., 2000). However, not all studies have been able to demonstrate the effect of host age on the composition of human E. coli populations (Escobar-Páramo et al., 2004; Li et al., 2010).

Within-host competition also appears to play a role in determining the diversity of strains found in a host. Several studies have observed that when a B2 strain is numerically dominant fewer strains will be detected in the host than in hosts where a strain belonging to phylogroups A, B1 or D is numerically dominant (Moreno et al., 2009). That is, the presence of a B2 strain in a host appears to inhibit the establishment of other strains.

Humans are exposed to E. coli on a daily basis, with estimates suggesting humans ingest $10^6$ cells of E. coli with every gram of food eaten (Hartl & Dykhuizen, 1984). However, as the survey and experimental studies show, the strains which succeed in establishing in a host are a non-random subset of the strains to which a host is exposed. Although the precise factors determining the type of strain establishing in a host are unknown.

Studies examining multiple isolates from a single host provide additional evidence supporting the conclusion that only particular strains will succeed in getting established inside a host. Several studies have shown that when two or more strains are present in a host, the second most abundant strain is significantly more likely to belong to the same phylogroup as the numerically dominant strain (Smati et al., 2014; Blyton et al., 2014; Gordon et al., 2015).

There are non-mutually exclusive explanations for the observation that when two or more strains are found in a host they are more likely to be members of the same phylogroup, rather than different phylogroups. One explanation is that unknown factors related to the gastrointestinal environment, determine the kind of strains that can be established inside a host. That is, if a host exhibits a gut environment conducive to the establishment of, for example, a B2 strain rather than strains of the other phylogroups, then when a host harbours more than one strain they will be members of
the same phylogroup. The other explanation is that the presence of two or more strains belonging to the same phylogroup in a host is a result of within-host evolution events. That is, a strain of a particular phylogroup succeeds in getting established inside a host and persists long enough for variants of the original strain to appear. Although there is relatively little data, there are examples of *E. coli* strains persisting in a host for from 2-5 years (Clermont et al., 2008). Long-term *E. coli* evolution studies in mice, have shown that new variants will arise and be detected in strains persisting in a host for a year (Lee et al., 2010).

1.9 Aim of the study

The aim of this study is to determine, in hosts harbouring multiple strains belonging to the same phylogroup, what fraction represent independent immigration and establishment events *versus* what fraction represent diversity arising as a consequence of within-host evolution.

To investigate this question, I will exploit a large collection of strains isolated from biopsies taken from different gut regions of 67 patients undergoing the colonoscopy and characterized using REP-PCR fingerprinting (Gordon et al., 2015). One example of every fingerprint type available from a patient was whole genome sequenced and these data were used to address the aims of the study.
Chapter 2

MATERIALS and METHODS
2.1 Strain Collection

This study was based on a collection of strains reported by Gordon et al. (2015). In this study, biopsies were collected from 69 patients undergoing colonoscopies at the Canberra Hospital, Canberra, Australia in 2010 and 2011. Biopsies were collected from up to five gut regions in each patient: terminal ileum, ascending colon, transverse colon, descending colon and rectum. All the details of disease condition of each patient are presented in appendix 1. Up to 20 isolates of *E. coli* were recovered from each biopsy. All isolates were REP-PCR fingerprinted using two primers (CCG)\(_5\) and ERIC (Adamus-Bialek et al., 2009) and assigned to a phylogroup using the Quadruplex method (Clermont et al., 2013).

2.2 Whole genome sequencing

Some isolates from this study were previously sequenced as a part of study of adherent invasive *E. coli* (O’Brien et al., 2016) using the Illumina HiSeq 2000 platform. In addition to these isolates, an attempt was made to whole genome sequence one example of each fingerprint type detected in every host. However, -80°C cultures were not available for every fingerprint type from each patient, as some had been lost while the Gordon et al. (2015) study was being carried out.

DNA from every available isolate was extracted from a 100 µL aliquot of an overnight lysogeny broth culture using the ISOLATE II Genomic DNA Kit – For Bacteria (Bioline) according to manufacturer’s protocol. Qubit\textsuperscript{®} dsDNA BR quantification was used to determine the concentration of DNA using a Qubit\textsuperscript{®} Fluorometer (Invitrogen\textsuperscript{™}). After quantification DNA samples were diluted to achieve a final concentration of 0.2 ng/µL. DNA libraries were made using the Nextera\textsuperscript{®} XT DNA library preparation kit (Illumina\textsuperscript{®}) and the Nextera XT Index Kit (Illumina\textsuperscript{®}) according to the manufacturer’s protocols. Libraries were sequenced using the MiSeq\textsuperscript{®} system (Illumina\textsuperscript{®}) together with the MiSeq Reagent Kit v3 (600-cycles). Forty eighty strain were sequenced per MiSeq run.

2.3 Microbial genome assembly

The A5MiSeq (Andrew And Aaron’s Awesome Assembly pipeline)-MiSeq assembly software was used to assemble the read data for each isolate (Coil et al., 2015). The
WGS data for every strain was submitted to Enterobase (http://enterobase.warwick.ac.uk/) and is publically available.

2.4 Basic characterization
The strains were assigned to sequence types (STs) using the University of Warwick MLST scheme (http://enterobase.warwick.ac.uk/). The Centre for Genomic Epidemiology (CGE) website (http://www.genomicepidemiology.org/) tool SeroTypeFinder (Joensen et al., 2015) was used to determine the serotype of all isolates.

2.5 CSIPhylogeny - Variant finding
In this study, comparisons between related strains were done using the CSIPhylogeny tool (Kaas et al., 2014). Variant calling or Single Nucleotide Polymorphism (SNP) calling is the simplest method to find the difference between strains. This tool is publically available at http://cge.cbs.dtu.dk/services/CSIPhylogeny/.

2.6 Annotation of variants
A tool from Pathosystems Resource Integrated Center (PATRIC) was used to identify and annotate the variants. The PATRIC analysis tool is available on http://www.beta.patricbrc.org/app/variation. This tool maps read data against a reference genome and provides a detailed description of the variants detected. The first stage in variant detection is annotation of the reference genome. PATRIC uses Rapid Annotations using Subsystems Technology (RAST) for genome annotation (Wattam et al., 2014). The differences between each combination of strains observed in this study that were inferred to represent within host evolution events were investigated. For each pair of strains compared, one strain was used as the reference strain and the raw read data from the other strains mapped to the reference, the process was then repeated by switching the role of each strain.
Chapter 3

RESULTS
3.1 Host characteristics

Biopsies were collected from 67 patients out of which 32 were female and 35 male. Individuals ranged from 18 to 73 years of age. The disease state of each individual was determined based on an individual’s clinical history and the appearance of the gastro-intestinal tract during the colonoscopy: 32 individuals were diagnosed with Crohn’s disease, 17 with ulcerative colitis, and 18 of the individuals had gastro-intestinal tracts that appeared to be disease free. Age, sex and disease state for each individual is presented in Appendix 1.

3.2 SNP detection

Strains belonging to different phylogroups or sequence types (STs) isolated from the same host were considered to be a consequence of independent exposure and establishment events. However, a number of individuals harboured isolates belonging to the same ST, but which were found to represent different REP types based on previous REP-PCR fingerprinting (Gordon et al., 2015).

In order to determine if isolates from the same individual and belonging to the same ST could be considered distinct strains, it was necessary to first determine how many SNPs could be observed between the assemblies of the same isolate that had been whole genome sequenced on two separate occasions. If the number of SNPs observed between two isolates exceeded the number of SNPs observed between the same isolate sequenced twice then the two isolates were defined as representing different strains, if not, then the two isolates were considered to be multiple examples of the same strain.

In order to determine the number of SNPs differentiating two isolates, the SNP detection method Call SNPs and Infer Phylogeny (CSIPhylogeny) (Kaas et al., 2014) implemented at the Centre for Genomic Epidemiology web site http://cge.cbs.dtu.dk/services/CSIPhylogeny was used. CSIPhylogeny maps the raw sequence data for an isolate against a reference strain. The method implements a number of quality control steps. The minimum amount of coverage at a SNP position can be varied. Base call quality and read quality can also be varied. Base quality is a measure of the accuracy with which a base has been called, while read quality is a
measure of the accuracy with which a single read has been mapped to the reference. In CSIPhylogeny two other quality control steps have been implemented. One is the minimum relative depth at a SNP position, this step represents an attempt to control situations where the level of coverage varies extensively for the WGS data for different samples (e.g. 20x vs 1000x coverage). As the great majority of the WGS data in the present study had broadly similar levels of coverage this option was disabled in CSIPhylogeny. In CSIPhylogeny, there is also the ability to control the minimum distance allowed between SNPs, but this option was disabled as the goal was to detect all SNPs.

In order to select the appropriate cut-offs for the various quality control parameters the WGS data for the ST73 strain 23-1-TC4 was used together with the assembly for the ST73 reference strain CFT073. First all quality control parameters, except for the one being investigated were disabled. Then a value was chosen for the remaining quality control parameter (e.g. depth at SNP position) and the number of SNPs detected was recorded. This procedure was repeated using different values for the quality control parameter. The same procedure was used to investigate the impact of the other quality control parameters on the number of SNPs detected.

Figure 3.1. Number of SNPs detected using the detection method CSIPhylogeny (Kaas et al., 2014) as a function of base call and read mapping quality, and coverage.
Coverage at a SNP was found to be the most important parameter in determining the number of SNPs detected. When coverage at a SNP position was disabled significantly more SNPs were detected regardless of the score used for base call quality or read mapping quality (Fig. 3.1). For example, a quality score of 20 (99% probability of being correct) for read mapping quality yielded 76 SNPs while 70 SNPs were detected when a quality score of 20 was used for base calling accuracy. By contrast, when a coverage value of 20 was used only 4 SNPs were detected. As a consequence, it was decided that all SNP detections would be determined using a read mapping quality score of 25, a SNP quality score of 30, a SNP depth score of 10, and all other options were disabled.

3.3 Defining isolates as distinct
The first step in determining if two isolates belonging to the same ST were distinct was to determine the number of SNPs detected when comparing the WGS data for the same strain that had been independently sequenced twice. Two independent culture of the same strain of 19 samples were used for DNA extraction and for WGS. However, two sets of WGS data was available for a total of 19 strains (Appendix C). To determine the number of SNP differences between the two sets of WGS data the following approach was used. CSIPhylogeny was used to detect SNPs using the parameter setting described previously. Each set of read data was assembled using the A5MiSeq pipeline and the raw read data from one sequence run was mapped against the assembly for the other sequence run of the strain and the number of and position of all SNPs determined. The procedure was then repeated using the other set of sequence data as the assembled reference and the raw read data. The predictions of the two analyses were compared and the true SNPs were considered to be those SNPs detected in both sets of SNP data.

The result of this analysis revealed that in 17 of the 19 cases of the same isolate being sequenced twice there were no SNP differences detected between the two assemblies. In two cases (10.5%) a single SNP difference was detected. Therefore, if the WGS data of two isolates representing the same ST differed by less than two SNPs then these two isolates were considered to represent two examples of the same strain. If two or more SNPs were detected, then these two isolates were considered to represent different strains.
3.4 Strain diversity within hosts

WGS data was available for 174 isolates from 67 hosts. Based on REP-PCR fingerprinting, these were predicted to represent 158 distinct genotypes (on average, 2.3 genotypes per individual). The balance of the 174 isolates represented 16 cases where isolates with the same REP-PCR fingerprint type, but from different gut regions of the same patient were sequenced.

The ST of all isolates was determined. Isolates representing the same ST from the same patient were compared in a pairwise fashion in order to determine if the isolates did represent distinct strains, that is, the pair of isolates being compared differed by 2 or more SNPs. There was a total of 113 isolates representing the same ST from the same patient and 34 of these actually represented distinct strains. Therefore, while strains belonging to different STs always had different REP-PCR fingerprints, there were cases where isolates belonging to the same ST from the same host but with different REP-PCR fingerprints did not in fact represent different strains.

The diversity of strains within each of the 67 hosts is presented in Table 3.1. In several cases WGS data was not available for an isolate as no culture existed. However, the phylogroup membership of the missing isolate was known as was its virulence factor profile. There were two cases (Hosts 6 and 8) where the virulence factor profile clearly indicated that there were two very different strains of the same phylogroup although WGS data was not available for one of the strains. Overall, there was an average of 2.1 strains per host.
Table 3.1. Phylogroup and sequence type (ST) of the distinct strains detected in each patient. A migration event denotes a case when the patient harbours two or more strains representing different STs or belonging to the same ST but differing by more than 1000 SNPs. A within-host evolution event denotes two or more strains of the same ST than differ by more than 2 but less than 50 SNPs. (See text for further details)

<table>
<thead>
<tr>
<th>Host</th>
<th>Genotype Number</th>
<th>Migration Events</th>
<th>Within-Host Evolution Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B2 (91)</td>
<td>D (349)</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>B2</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>D (69)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>B2 (95)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>B2 (95)</td>
<td>B2</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>D (349)</td>
<td>B2</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>A (10)</td>
<td>B1 (270)</td>
<td>D (3300)</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>B2 (131)</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>B2 (131)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>D (69)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>B2 (127)</td>
<td>B2 (127)</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>B2 (420)</td>
<td>B2 (420)</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>B2</td>
<td>D (3056)</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>A</td>
<td>B2 (91)</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>A (216)</td>
<td>A (216)</td>
<td>B2</td>
</tr>
<tr>
<td>17</td>
<td>B1</td>
<td>F (59)</td>
<td>F (59)</td>
</tr>
<tr>
<td>18</td>
<td>B1 (906)</td>
<td>B2 (73)</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>B1 (1304)</td>
<td>B1 (1304)</td>
<td>B2 (73)</td>
</tr>
<tr>
<td>21</td>
<td>B2 (73)</td>
<td>B2 (73)</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>D (69)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>B2 (73)</td>
<td>B2 (73)</td>
<td>B2 (73)</td>
</tr>
<tr>
<td>24</td>
<td>A</td>
<td>B1 (58)</td>
<td>B2 (28)</td>
</tr>
<tr>
<td>25</td>
<td>D (132)</td>
<td>D (362)</td>
<td>2</td>
</tr>
<tr>
<td>26</td>
<td>D (973)</td>
<td>F (1674)</td>
<td>2</td>
</tr>
<tr>
<td>27</td>
<td>B2 (95)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>D</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>29</td>
<td>B1 (1049)</td>
<td>B2 (589)</td>
<td>D (362)</td>
</tr>
<tr>
<td>30</td>
<td>F (648)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>B2 (131)</td>
<td>B2 (420)</td>
<td>2</td>
</tr>
<tr>
<td>32</td>
<td>B2 (420)</td>
<td>B2 (131)</td>
<td>2</td>
</tr>
<tr>
<td>33</td>
<td>B1 (1727)</td>
<td>B2 (131)</td>
<td>2</td>
</tr>
<tr>
<td>34</td>
<td>B2 (95)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>35</td>
<td>B2 (429)</td>
<td>B2 (429)</td>
<td>1</td>
</tr>
<tr>
<td>36</td>
<td>B1 (6171)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>37</td>
<td>B2 (95)</td>
<td>B2 (95)</td>
<td>1</td>
</tr>
<tr>
<td>38</td>
<td>B2 (95)</td>
<td>B2 (95)</td>
<td>1</td>
</tr>
<tr>
<td>39</td>
<td>D (720)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>40</td>
<td>A (48)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>41</td>
<td>B1 (1642)</td>
<td>B1 (2773)</td>
<td>2</td>
</tr>
<tr>
<td>42</td>
<td>B1 (58)</td>
<td>B1 (58)</td>
<td>1</td>
</tr>
<tr>
<td>43</td>
<td>A (398)</td>
<td>B2 (95)</td>
<td>B2 (144)</td>
</tr>
<tr>
<td>44</td>
<td>B2 (95)</td>
<td>D (69)</td>
<td>2</td>
</tr>
<tr>
<td>Host</td>
<td>Genotype Number</td>
<td>Migration Events</td>
<td>Within-Host Evolution Events</td>
</tr>
<tr>
<td>------</td>
<td>----------------</td>
<td>-----------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>47</td>
<td>B1 (160)</td>
<td>B2 (12)</td>
<td>D (110)</td>
</tr>
<tr>
<td>48</td>
<td>B1 (1727)</td>
<td>D (108)</td>
<td>2</td>
</tr>
<tr>
<td>50</td>
<td>A (6174)</td>
<td>B1 (154)</td>
<td>B1 (6169)</td>
</tr>
<tr>
<td>51</td>
<td>B2 (95)</td>
<td>B2 (550)</td>
<td>2</td>
</tr>
<tr>
<td>53</td>
<td>B1 (58)</td>
<td>B1 (58)</td>
<td>1</td>
</tr>
<tr>
<td>54</td>
<td>B2 (1919)</td>
<td>B2 (1919)</td>
<td>1</td>
</tr>
<tr>
<td>55</td>
<td>B2 (131)</td>
<td>B2 (131)</td>
<td>1</td>
</tr>
<tr>
<td>56</td>
<td>B2 (95)</td>
<td>D (6170)</td>
<td>D (6170)</td>
</tr>
<tr>
<td>57</td>
<td>A (409)</td>
<td>B2 (537)</td>
<td>2</td>
</tr>
<tr>
<td>58</td>
<td>A (607)</td>
<td>B2 (28)</td>
<td>2</td>
</tr>
<tr>
<td>59</td>
<td>A (653)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>60</td>
<td>B2 (80)</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>61</td>
<td>B2 (95)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>62</td>
<td>B1 (223)</td>
<td>B2 (12)</td>
<td>B2 (95)</td>
</tr>
<tr>
<td>63</td>
<td>B2 (80)</td>
<td>D (963)</td>
<td>2</td>
</tr>
<tr>
<td>65</td>
<td>A (10)</td>
<td>D (1177)</td>
<td>D (1177)</td>
</tr>
<tr>
<td>66</td>
<td>B2</td>
<td>D (938)</td>
<td>2</td>
</tr>
<tr>
<td>67</td>
<td>A (216)</td>
<td>D (132)</td>
<td>D (362)</td>
</tr>
<tr>
<td>68</td>
<td>B1 (517)</td>
<td>D (362)</td>
<td>2</td>
</tr>
<tr>
<td>69</td>
<td>B2 (569)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>70</td>
<td>B2 (95)</td>
<td>B2 (569)</td>
<td>D (69)</td>
</tr>
<tr>
<td>71</td>
<td>D (362)</td>
<td>F (1674)</td>
<td>2</td>
</tr>
<tr>
<td>73</td>
<td>B1 (2005)</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

3.5 Immigration/Establishment versus within-host evolution

The presence of a one strain in a host represents an immigration/establishment event. In hosts with two or more strains, every strain present in a host that belonged to a different ST was also assumed to represent a different immigration/establishment event. When a host harboured two or more strains belonging to the same ST these strains could either represent independent immigration/establishment events or be the result of within-host evolution.

To distinguish between these two possibilities the average number of SNPs observed between strains belonging to the same ST isolated from different hosts was first assessed. Data was available for strains belonging to 16 STs (Table 3.2). On average, two strains belonging to the same ST but isolated from different hosts, differed by 4135 SNPs although this value varied considerably among STs.
Table 3.2: The average number of SNPs detected among strains belonging to the same sequence type (ST) but isolated from different patients.

<table>
<thead>
<tr>
<th>Sequence Type</th>
<th>Mean Number of SNPs</th>
<th>Number of strains</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>4370</td>
<td>2</td>
</tr>
<tr>
<td>349</td>
<td>3300</td>
<td>2</td>
</tr>
<tr>
<td>1727</td>
<td>9845</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>6676</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>61</td>
<td>2</td>
</tr>
<tr>
<td>91</td>
<td>1735</td>
<td>2</td>
</tr>
<tr>
<td>420</td>
<td>1458</td>
<td>2</td>
</tr>
<tr>
<td>132</td>
<td>849</td>
<td>3</td>
</tr>
<tr>
<td>569</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>58</td>
<td>4567</td>
<td>5</td>
</tr>
<tr>
<td>216</td>
<td>2448</td>
<td>3</td>
</tr>
<tr>
<td>362</td>
<td>3516</td>
<td>8</td>
</tr>
<tr>
<td>131</td>
<td>8730</td>
<td>7</td>
</tr>
<tr>
<td>73</td>
<td>4358</td>
<td>7</td>
</tr>
<tr>
<td>69</td>
<td>10771</td>
<td>5</td>
</tr>
<tr>
<td>95</td>
<td>3489</td>
<td>15</td>
</tr>
</tbody>
</table>

By contrast, strains belonging to the same ST from the same host differed by fewer than 12 SNPs (Figure 3.2). There was one exception, Host 23, harboured three ST73 strains. One of these strains differed from the other ST73 strains by an average of 6332 SNPs, while the other ST73 strains in the host differed by two SNPs. Thus, two quite different ST73 strains succeeded in establishing in this host and one has given rise to another variant as a result of within-host evolution.

Figure 3.2: Distribution of the number of SNP differences between strains isolate from the same host and belonging to the same ST.
The criteria used to distinguish between immigration/establishment events and within host evolution is also illustrated by depicting the phylogenetic relationships among all ST58, ST69, ST73, ST95, ST131 and ST 362 strains observed in this study (Fig. 3.3). The phylogenetic tree was constructed using MEGA to find the relatedness of major STs of this study.

**Figure 3.3.** Phylogenetic relationships among the sequence types 58, 69, 73, 65, 131 and 362 strains observed in this study. The first number in the strain name denotes the patient from which the isolates were taken. This phylogenetic tree is based on the strains of the same STs from same patients which is representing different immigration/establishment events within-host.
When a host harboured two or more strains belonging to the same phylogroup it could be due to immigration/establishment or within-host evolution. There were 33 cases of two of more strains belonging to the same phylogroup in a single host (Table 3.1). Within-host evolution accounted for the presence of another strain of the same phylogroup in 54 % of the cases, while in 45% of the cases this outcome was due to independent immigration/establishment events (Table 3.1).

The relative importance of within-host variation versus immigration/establishment in explaining the presence of two or more strains of the same phylogroup did not vary among phylogroups (Likelihood Ratio $X^2 = 3.6$, $p = 0.46$) (Table 3.3).

Table 3.3. Fraction of cases where the presence of two or more strains within a host was a consequence of within-host evolution versus migration/establishment as a function of the phylogroup membership of the strain.

<table>
<thead>
<tr>
<th>Phylogroup</th>
<th>N</th>
<th>Within Host Evolution</th>
<th>Migration</th>
</tr>
</thead>
<tbody>
<tr>
<td>A and F</td>
<td>2</td>
<td>100 %</td>
<td>0 %</td>
</tr>
<tr>
<td>B1</td>
<td>5</td>
<td>60 %</td>
<td>40 %</td>
</tr>
<tr>
<td>B2</td>
<td>18</td>
<td>61 %</td>
<td>39 %</td>
</tr>
<tr>
<td>D</td>
<td>8</td>
<td>37 %</td>
<td>63 %</td>
</tr>
</tbody>
</table>

3.6 Genetic changes and within-host evolution

Within-host evolution events accounted for a substantial portion of the within host strains diversity observed in this study. The nature of the changes that occurred in the strains that evolved within the host was investigated using the variant detection tool available at the PATRIC website (https://www.patricbrc.org/app/Variation).

Comparison of the strains inferred to have evolved within a host revealed that, on an average, 13% of the changes were in non-coding regions, 21% represented synonymous changes, and 66% non-synonymous changes (Table 3.4). However, the observed mutations occurred in genes with a variety of functions and there appeared to be little pattern in the functional nature of the genes in which mutations were observed (Table 3.4).
<table>
<thead>
<tr>
<th>Patient</th>
<th>Phylogroup ST</th>
<th>Strains compared</th>
<th>Strain 1</th>
<th>Strain 2</th>
<th>Mutation</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>B2 127</td>
<td>12-1-Ti12/12-2-Ti13</td>
<td>T</td>
<td>G</td>
<td>Synon</td>
<td>Oligopeptide transport permease protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Noncod</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Synon</td>
<td>MFS superfamily export protein YceL</td>
</tr>
<tr>
<td>13</td>
<td>B2 420</td>
<td>13-3-TC6/13-2-TCS</td>
<td>A</td>
<td>T</td>
<td>Nonsyn</td>
<td>Tyrosine-protein kinase Wzc (EC 2.7.10.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>G</td>
<td>Noncod</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>A 216</td>
<td>16-3-C13/16-2-Si18</td>
<td>T</td>
<td>C</td>
<td>Noncod</td>
<td>Aminobenzoyl-glutamate transport protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>G</td>
<td>Nonsyn</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>F 59</td>
<td>17-1-R20/17-2-DC14</td>
<td>T</td>
<td>C</td>
<td>Nonsyn</td>
<td>Lipoprotein NlpI</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>G</td>
<td>Nonsyn</td>
<td>LsrR, transcriptional repressor of lsr operon</td>
</tr>
<tr>
<td>19</td>
<td>B2 14</td>
<td>19-1-TG/19-1-Ti7</td>
<td>A</td>
<td>G</td>
<td>Nonsyn</td>
<td>PTS system, mannitol-specific cryptic IIB component (EC 2.7.1.69)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>A</td>
<td>Noncod</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>A</td>
<td>Synon</td>
<td>FIG00638808: hypothetical protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>G</td>
<td>Nonsyn</td>
<td>Co-activator of prophage gene expression lbrB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>G</td>
<td>Noncod</td>
<td>Uncharacterized substrate-binding protein YbaE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>C</td>
<td>Noncod</td>
<td>Probable diguanylate cyclase YeaP</td>
</tr>
<tr>
<td>19</td>
<td>B2 14</td>
<td>19-1-TG/19-1-Ti7</td>
<td>T</td>
<td>G</td>
<td>Synon</td>
<td>FIG00638808: hypothetical protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>A</td>
<td>Synon</td>
<td>FIG00638808: hypothetical protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>G</td>
<td>Synon</td>
<td>Co-activator of prophage gene expression lbrB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>G</td>
<td>Synon</td>
<td>Uncharacterized substrate-binding protein YbaE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>G</td>
<td>Synon</td>
<td>Co-activator of prophage gene expression lbrB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>T</td>
<td>Nonsyn</td>
<td>FIG00638808: hypothetical protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>A</td>
<td>Synon</td>
<td>FIG00638808: hypothetical protein</td>
</tr>
<tr>
<td>20</td>
<td>B1 1304</td>
<td>20-2-TG/20-3-TC12</td>
<td>T</td>
<td>C</td>
<td>Nonsyn</td>
<td>Outer membrane lipoprotein pcp precursor</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Nonsyn</td>
<td>Error-prone, lesion bypass DNA polymerase V (UmuC)</td>
</tr>
<tr>
<td>20</td>
<td>B1 1304</td>
<td>20-1-TG/20-2-TG13</td>
<td>A</td>
<td>G</td>
<td>Nonsyn</td>
<td>Outer membrane lipoprotein pcp precursor</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>T</td>
<td>Nonsyn</td>
<td>Error-prone, lesion bypass DNA polymerase V (UmuC)</td>
</tr>
<tr>
<td>20</td>
<td>B1 1304</td>
<td>20-1-TG/20-3-TG12</td>
<td>T</td>
<td>C</td>
<td>Nonsyn</td>
<td>Putative anti-Filc (2) FilD (4) factor YdiV</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Nonsyn</td>
<td>Putative anti-Filc (2) FilD (4) factor YdiV</td>
</tr>
<tr>
<td>21</td>
<td>B2 73</td>
<td>21-1-TG/21-2-C8</td>
<td>A</td>
<td>C</td>
<td>Nonsyn</td>
<td>Enterobactin synthetase component F, serine activating enzyme (EC 2.7.7.-)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>G</td>
<td>Nonsyn</td>
<td>Allantoinase (EC 3.5.2.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>G</td>
<td>Nonsyn</td>
<td>4-hydroxy-3-methylbut-2-enyl diphosphate reductase (EC 1.17.1.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Synon</td>
<td>Intramembrane protease RasP/YluC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>T</td>
<td>Nonsyn</td>
<td>Helicase PriA essential for oriC/DnaA-independent DNA replication</td>
</tr>
<tr>
<td>Patient</td>
<td>Phylogroup ST</td>
<td>Strains compared</td>
<td>Strain 1</td>
<td>Strain 2</td>
<td>Mutation</td>
<td>Function</td>
</tr>
<tr>
<td>---------</td>
<td>--------------</td>
<td>------------------</td>
<td>---------</td>
<td>---------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>23</td>
<td>B2 73</td>
<td>23-1-TC4/23-2-Ti8</td>
<td>C</td>
<td>A</td>
<td>Nonsyn</td>
<td>Phosphoenolpyruvate-protein phosphotransferase of PTS system (EC 2.7.3.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>G</td>
<td>Noncod</td>
<td>Phosphoenolpyruvate-protein phosphotransferase of PTS system (EC 2.7.3.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23-1-TC4/23-3-Ti9</td>
<td>C</td>
<td>A</td>
<td>Nonsyn</td>
<td>Phosphoenolpyruvate-protein phosphotransferase of PTS system (EC 2.7.3.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Noncod</td>
<td>Phosphoenolpyruvate-protein phosphotransferase of PTS system (EC 2.7.3.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23-1-TC4/23-2-AC12</td>
<td>C</td>
<td>A</td>
<td>Nonsyn</td>
<td>Phosphoenolpyruvate-protein phosphotransferase of PTS system (EC 2.7.3.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23-2-AC12/23-2-Ti8</td>
<td>A</td>
<td>G</td>
<td>Noncod</td>
<td>Phosphoenolpyruvate-protein phosphotransferase of PTS system (EC 2.7.3.9)</td>
</tr>
<tr>
<td>29</td>
<td>D 362</td>
<td>29-1-DC3/29-4-DC8</td>
<td>T</td>
<td>C</td>
<td>Nonsyn</td>
<td>Endonuclease IV (EC 3.1.21.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>A</td>
<td>Nonsyn</td>
<td>DNA sulfur modification protein DndB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>G</td>
<td>Nonsyn</td>
<td>Fig00642784: hypothetical protein</td>
</tr>
<tr>
<td>36</td>
<td>B2 429</td>
<td>36-1-TC9/36-1-Ti13</td>
<td>C</td>
<td>T</td>
<td>Synon</td>
<td>Peptidyl-prolyl cis-trans isomerase PpiD (EC 5.2.1.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Nonsyn</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>T</td>
<td>Noncod</td>
<td>Thiol: disulfide interchange protein DsbG precursor</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Synon</td>
<td>Quinolinate phosphoribosyltransferase [decarboxylating] (EC 2.4.2.19)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>A</td>
<td>Nonsyn</td>
<td>Peptidyl-prolyl cis-trans isomerase PpiD (EC 5.2.1.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>T</td>
<td>Noncod</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>T</td>
<td>Synon</td>
<td>LysR family transcriptional regulator YafC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Synon</td>
<td>Peptidyl-prolyl cis-trans isomerase PpiD (EC 5.2.1.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>C</td>
<td>Nonsyn</td>
<td>Formate hydrogenlyase subunit 3</td>
</tr>
<tr>
<td>40</td>
<td>B2 95</td>
<td>40-3-R7/40-1-R9</td>
<td>A</td>
<td>C</td>
<td>Nonsyn</td>
<td>Putative HTH-type transcriptional regulator ybaO</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Nonsyn</td>
<td>Formate hydrogenlyase subunit 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>T</td>
<td>Nonsyn</td>
<td>Formate hydrogenlyase subunit 3</td>
</tr>
<tr>
<td>44</td>
<td>B1 58</td>
<td>44-4-Ti15/44-1-Ti4</td>
<td>A</td>
<td>C</td>
<td>Nonsyn</td>
<td>HMP-PP hydrolase (pyridoxal phosphatase) Cof</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>A</td>
<td>Synon</td>
<td>Mobile element protein</td>
</tr>
<tr>
<td>53</td>
<td>B1 58</td>
<td>53-1-AC6/53-1-AUC1</td>
<td>C</td>
<td>G</td>
<td>Nonsyn</td>
<td>Membrane-bound lytic murein transglycosylase F (EC 4.2.2.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>A</td>
<td>Synon</td>
<td>Ribulokinase (EC 2.7.1.16)</td>
</tr>
<tr>
<td>54</td>
<td>B2 1919</td>
<td>54-1-TC4/54-1-Ti6</td>
<td>A</td>
<td>G</td>
<td>Nonsyn</td>
<td>Hybrid sensory histidine kinase</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>T</td>
<td>Noncod</td>
<td>Hmp-PP hydrolase (pyridoxal phosphatase) Cof</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>T</td>
<td>Nonsyn</td>
<td>Hmp-PP hydrolase (pyridoxal phosphatase) Cof</td>
</tr>
<tr>
<td>55</td>
<td>B2 131</td>
<td>55-1-TC9/55-1-Ti19</td>
<td>C</td>
<td>A</td>
<td>Nonsyn</td>
<td>Glutaminase (EC 3.5.1.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>T</td>
<td>Nonsyn</td>
<td>Phosphoenolpyruvate-protein phosphotransferase of PTS system (EC 2.7.3.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>T</td>
<td>Synon</td>
<td>ATP-dependent helicase HrpA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>A</td>
<td>Nonsyn</td>
<td>Probable diguanylate cyclase YedQ</td>
</tr>
<tr>
<td>Patient</td>
<td>Phylogroup ST</td>
<td>Strains compared</td>
<td>Strain 1</td>
<td>Strain 2</td>
<td>Mutation</td>
<td>Function</td>
</tr>
<tr>
<td>---------</td>
<td>---------------</td>
<td>------------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>56</td>
<td>D 6170</td>
<td>56-1-AC4/56-3-AC1</td>
<td>A</td>
<td>C</td>
<td>Nonsyn</td>
<td>[Protein-Pil] uridylyltransferase (EC 2.7.7.59)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>T</td>
<td>Nonsyn</td>
<td>Asparagine synthetase [glutamine-hydrolyzing] (EC 6.3.5.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>G</td>
<td>Synon</td>
<td>Putative structural protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td>56-1-AC4/56-4-R1</td>
<td>A</td>
<td>C</td>
<td>Nonsyn</td>
<td>[Protein-Pil] uridylyltransferase (EC 2.7.7.59)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>T</td>
<td>Synon</td>
<td>Lysine/cadaverine antiporter membrane protein CadB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Nonsyn</td>
<td>N-acetylglucosamine-6-phosphate deacetylase (EC 3.5.1.25)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>G</td>
<td>Synon</td>
<td>Putative structural protein</td>
</tr>
<tr>
<td>56</td>
<td>D 6170</td>
<td>56-3-AC1/56-4-R1</td>
<td>G</td>
<td>T</td>
<td>Synon</td>
<td>Lysine/cadaverine antiporter membrane protein CadB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>A</td>
<td>Synon</td>
<td>AI-2 transport protein TqsA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>A</td>
<td>Noncod</td>
<td>Acetyl-CoA acetyltransferase (EC 2.3.1.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>A</td>
<td>Nonsyn</td>
<td>Asparagine synthetase [glutamine-hydrolyzing] (EC 6.3.5.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Noncod</td>
<td>N-acetylglucosamine-6-phosphate deacetylase (EC 3.5.1.25)</td>
</tr>
<tr>
<td>65</td>
<td>D 1177</td>
<td>65-1-AC2/65-4-AC20</td>
<td>C</td>
<td>A</td>
<td>Nonsyn</td>
<td>Anaerobic selenate reductase</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Nonsyn</td>
<td>Protein involved in stability of MscS mechanosensitive channel</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Nonsyn</td>
<td>FIG00031715: Predicted metal-dependent phosphoesterases (PHP family)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>A</td>
<td>Nonsyn</td>
<td>FIG00640233: hypothetical protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>A</td>
<td>Noncod</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>G</td>
<td>Nonsyn</td>
<td>Electron transfer flavoprotein, beta subunit FxA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>G</td>
<td>Synon</td>
<td>Putative NAGC-like transcriptional regulator</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>A</td>
<td>Nonsyn</td>
<td>FhuE receptor precursor</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>T</td>
<td>Synon</td>
<td>Pseudouridine transporter PsuT (putative)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>C</td>
<td>Nonsyn</td>
<td>Sodium-Choline Symporter</td>
</tr>
<tr>
<td></td>
<td></td>
<td>65-1-AC2/65-3-TC1</td>
<td>T</td>
<td>A</td>
<td>Nonsyn</td>
<td>62kDa structural protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>T</td>
<td>Synon</td>
<td>Pseudouridine transporter PsuT (putative)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>G</td>
<td>Nonsyn</td>
<td>FIG006303: protein yraQ</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>C</td>
<td>Nonsyn</td>
<td>DEAD-box ATP-dependent RNA helicase CshA (EC 3.6.4.13)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Nonsyn</td>
<td>Lipopolysaccharide-assembly protein LptC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>A</td>
<td>Noncod</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>A</td>
<td>Nonsyn</td>
<td>FhuE receptor precursor</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Nonsyn</td>
<td>Putative transport protein YbjL</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>G</td>
<td>Nonsyn</td>
<td>Oligopeptide transport substrate-binding protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>T</td>
<td>Nonsyn</td>
<td>FIG00638098: hypothetical protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>T</td>
<td>Nonsyn</td>
<td>3-dehydro-L-gulonate 2-dehydrogenase (EC 1.1.1.130)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>Nonsyn</td>
<td>Transcriptional regulatory protein PhoP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>C</td>
<td>Nonsyn</td>
<td>Sodium-Choline Symporter</td>
</tr>
</tbody>
</table>
Chapter 4

DISCUSSION
Escherichia coli resides in the gastrointestinal tract of humans and the diversity, density and dominance of these strains depends on the numerous factors, such as, host age, sex, and diet, digesta retention times, gut morphology, as well as what strains the host is exposed to. This study asks the question, if a host harbours multiple strains belonging to the same phylogroup, then what fraction represent immigration/establishment events versus what fraction represent within host evolution events? In this study 83% of the strain richness observed in a host was a consequence of immigration/establishment events, rather than within host evolution. However, when only hosts harbouring strains of the same phylogroup are considered, then about half (54%) of the observed strain richness is a result of within host evolution events and the balance due to immigration/establishment events.

4.1 Caveats

Before discussing the significance of these results, I will examine the potential limitations of the study. Not all of the fingerprint types detected in every patient were available for sequencing. In order to avoid producing a freezer culture of every one of the almost 5000 isolates collected during the original study, it was decided to only make freezer cultures of one example of every fingerprint type detected in every gut region from each of the patients examined. Consequently, strains had to be fingerprinted and assigned a phylogroup prior to choosing the isolates for which freezer cultures would be made. While the isolates were being characterised they were stored on lysogeny broth plates at 5°C. During the time taken to characterise the isolates, some isolates died while being stored at 5°C. Those REP fingerprint types represented by a single isolate were more likely to be lost than those represented by multiple isolates. It also took significantly more time to characterise the isolates from the first set of patients examined compared to the isolates from subsequent sets of patients, and consequently more REP types were lost from this initial set of patients than subsequent sets. The loss of REP-PCR fingerprint types had the result that the number of strains observed per host based on whole genome sequencing (Table 3.1) was less than the average number of strains per host reported by Gordon et al. (2015).
The proportion of isolates for which WGS data was available that belonged to each phylogroup was not significantly different from the relative abundance of phylogroups reported in Gordon et al. (2015) (Table 4.1). This suggests that the isolates that were lost were likely a random subset of all isolates. In turn, this would suggest that the estimated ratio of immigration/establishment events versus within-host evolution events is also unbiased.

Table 4.1. Comparison of relative abundance of phylogroups distribution between Gordon et al. (2015) and this study.

<table>
<thead>
<tr>
<th>Phylogroup</th>
<th>Relative abundance of dominant phylogroups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gordon et al. (2015) study</td>
</tr>
<tr>
<td>A</td>
<td>7.2%</td>
</tr>
<tr>
<td>B1</td>
<td>14.5%</td>
</tr>
<tr>
<td>B2</td>
<td>44.9%</td>
</tr>
<tr>
<td>D</td>
<td>24.6%</td>
</tr>
<tr>
<td>E</td>
<td>2.9%</td>
</tr>
<tr>
<td>F</td>
<td>5.8%</td>
</tr>
</tbody>
</table>

A very conservative approach was taken in determining the number of SNP differences between two strains. Read mapping quality and base call quality scores were chosen to be greater than 20 (>99.9% confidence). Each potential SNP had to be represented by at least 10 reads. Less stringent requirements would result in greater absolute number of SNPs being called, but should not change the observation that the average number of SNPs observed between two strains belonging to the same ST and from the same patient should be greater than the average number of SNPs detected between the same strain sequenced twice and significantly smaller than the average number of SNPs detected between isolates of the same ST but from different patients. For strains representing within-host evolution events (belonging to the same ST and from the same host), SNP estimates were available using the CSIPhylogeny and variant detection tool in PATRIC (https://www.patricbrc.org/public/pdfs/Workshop-Variation-Service.pdf). Among the within host evolution comparisons CSIPhylogeny detected an average of 10 SNPs while PATRIC called an average of 78 SNPS and the correlation between the number of SNPs called by each method was (r=0.89).
4.2 The host as a determinant of \textit{E. coli} clonal diversity

There are now numerous studies that have shown that \textit{E. coli} genotypes are non-randomly distributed among hosts (Alm \textit{et al.}, 2011; Gordon \textit{et al.}, 2015). In particular, several studies have shown that when a host harbours two or more strains these strains are significantly more likely to be members of the same phylogroup. This study has shown that in about half of such cases, this is due to the immigration and establishment of strains and not within host evolution. Thus, the results of this study show that despite hosts being regularly exposed to a diversity of \textit{E. coli} through the food that they eat, factors related to the host, at least in part, determines what \textit{E. coli} strains succeed in establishing.

The precise nature of these factors is unknown. Studies examining the distribution of phylogroups in humans sampled in different parts of the world show systematic differences (Escobar-Páramo \textit{et al.}, 2004; Duries \textit{et al.}, 2001). B1 and A strains dominate in faecal samples collected from humans living in tropical/developing countries, while B2 strains are recovered from about 50\% of humans living in temperate/developed countries. There are substantial differences in the diets of people living in developed versus developing countries. In developing countries meat represents a smaller fraction of the diet relative to legumes and grains, while the opposite is true of the diets of people living in developed countries. People living in developed countries also have diets that are often based on highly processed foods. A recent study showed that over the past 30 years’ people in the Paris region of France have shown a switch from A and B1 strains to B2 strains. These differences may well be due to the increased use of highly processed foods in French diets. Similarly, another study examined the differences in the \textit{E. coli} populations of residents of metropolitan France compared to expatriates that relocated from France to French Guyana (Skurnik \textit{et al.}, 2008). The expatriates showed a lower prevalence of B2 strains compared to French residents. This observation provides additional evidence that differences in diet result in differences in phylogroup abundance. However, a limitation of all the studies examining the diversity of \textit{E. coli} in humans is that the effects of climate cannot be completely separated from the effects of diet. It may be that differences in survival in the external environment of strains belonging to the different phylogroups coupled with among-phylogroup differences in
transmission may explain the fact that B2 strains are less likely to be recovered from humans living in the tropics. However, there is no evidence to indicate that there are climate dependent differences in survival or transmission (Unno et al., 2009). Further, phylogenetic group B2 isolates were found be absent in human faecal samples collected in Jeonnam Province, South Korea, while phylogroup B2 strains were isolates from about 50% of faecal samples collected from residents of Tokyo, Japan (Obata-Yasuoka et al., 2002). As both localities have a temperate climate, this suggests that diet difference may be responsible for the differences in phylogroup abundance among different geographic localities.

Although differences in diet is the most likely explanation for the geographic variation in phylogroup abundance among humans, it is not known if diet exerts a direct effect, through the impact of diet, particularly fibre content and type, on gastro-intestinal transit times. Studies in rats indicate that the cell densities of phylogroup B2 strains to respond to changes in the fraction of non-fermentable fibre in the diet (O’Brien & Gordon, 2011). Another, non-mutually exclusive hypothesis, is that given it is well established that diet has a profound impact of the gut microbiota (Dore and Blottiere., 2015), it may be changes in the interactions of E. coli with the gut microbiota that explain the geographic differences in phylogroup abundance of E. coli isolated from humans. Further research is required before we can understand what drives the geographic variation in phylogroup abundance.

4.3 Within-host evolution

Most the E. coli diversity observed in a human is a consequence of immigration/establishment events. However, within host evolution did account for some of the observed diversity. Although samples sizes are small, there was no indication that the probability of observing a strain whose presence was inferred to be a result of within host evolution varied with the phylogenetic group membership of the strain (Table 3.4). The occurrence of within host evolution is not surprising given that there a numerous in vitro evolution studies of E. coli that show detectable variation can arise within a few generations (Kwon et al., 2015; Nisa et al., 2013), while within host variation has also been observed in vivo using a mouse model (Lescat et al., 2016; Batista-Barroso et al., 2014)
No patterns could be discerned in the nature of the differences between strains inferred to represent within host evolution events (Table 3.4). The lack of pattern is perhaps not surprising given that the patients sampled were of different sexes, ages and gut disease states.

In a study examining the extent of household sharing of ST131 *E. coli* strains (Johnson *et al.*, 2016), showed that ST131 strains isolated from different members of the same household were consistently more similar to one another, than were ST131 isolates recovered from different households. For 2 of the 6 households studied, within household variation was detected. In one household, the father and daughter shared a ST131 strain that was different to the ST131 strain present in the mother. In the other household, a mother and a child had indistinguishable isolates that differed from the ST131 isolate from the other child. Such an outcome indicates that variants can arise within a host and spread to other household members.

### 4.4 Measuring evolution of *E. coli* strains

Reeves *et al.* (2011) characterised 14 isolates of ST73 recovered from a single family over 3 years. Using this data, they inferred a mutation rate of about 1.1 per genome per year. These data can be used to infer how long the strains representing within host evolution events may have persisted in their hosts. Such an analysis would suggest that most strains had persisted in their hosts for 1.5-2.5 years, and one strain may have persisted for 11 years. Although there are very little data available, published estimates indicate that strains may persist for at least 6 years within a host (Johnson *et al.*, 2016). Given that no study has monitored hosts for more than 6 years, much longer resident times may well be possible.

Based on this study, it is clear when multiple strains belonging to the same phylogroup are present within a host, that independent immigration/establishment events are the explanation in about one half the observed cases. Further, research is required to understand that what are the host/gut microbiota factors that select for strains belonging to particular phylogroups.
REFERENCES


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


## Appendix A. Age, sex and disease condition of the patients/hosts

<table>
<thead>
<tr>
<th>Patient</th>
<th>Age</th>
<th>Sex</th>
<th>Disease condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56</td>
<td>F</td>
<td>Normal</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>F</td>
<td>Normal</td>
</tr>
<tr>
<td>4</td>
<td>49</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>5</td>
<td>18</td>
<td>F</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>6</td>
<td>44</td>
<td>F</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>7</td>
<td>47</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>8</td>
<td>62</td>
<td>M</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>9</td>
<td>73</td>
<td>M</td>
<td>Normal</td>
</tr>
<tr>
<td>10</td>
<td>54</td>
<td>M</td>
<td>Normal</td>
</tr>
<tr>
<td>11</td>
<td>69</td>
<td>M</td>
<td>Normal</td>
</tr>
<tr>
<td>12</td>
<td>61</td>
<td>M</td>
<td>Normal</td>
</tr>
<tr>
<td>13</td>
<td>43</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>14</td>
<td>18</td>
<td>F</td>
<td>Normal</td>
</tr>
<tr>
<td>15</td>
<td>35</td>
<td>M</td>
<td>Normal</td>
</tr>
<tr>
<td>16</td>
<td>53</td>
<td>F</td>
<td>Normal</td>
</tr>
<tr>
<td>17</td>
<td>31</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>18</td>
<td>21</td>
<td>M</td>
<td>Normal</td>
</tr>
<tr>
<td>19</td>
<td>26</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>20</td>
<td>40</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>21</td>
<td>71</td>
<td>M</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>22</td>
<td>26</td>
<td>M</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>23</td>
<td>50</td>
<td>F</td>
<td>Normal</td>
</tr>
<tr>
<td>24</td>
<td>43</td>
<td>F</td>
<td>Normal</td>
</tr>
<tr>
<td>25</td>
<td>24</td>
<td>M</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>26</td>
<td>26</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>27</td>
<td>23</td>
<td>M</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>28</td>
<td>40</td>
<td>M</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>29</td>
<td>23</td>
<td>M</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>30</td>
<td>29</td>
<td>M</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>32</td>
<td>24</td>
<td>M</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>33</td>
<td>53</td>
<td>M</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>35</td>
<td>51</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>36</td>
<td>49</td>
<td>M</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>37</td>
<td>47</td>
<td>F</td>
<td>Normal</td>
</tr>
<tr>
<td>39</td>
<td>28</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>40</td>
<td>52</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>41</td>
<td>65</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>42</td>
<td>22</td>
<td>M</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>43</td>
<td>24</td>
<td>M</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>44</td>
<td>61</td>
<td>M</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>Patient</td>
<td>Age</td>
<td>Sex</td>
<td>Disease condition</td>
</tr>
<tr>
<td>---------</td>
<td>-----</td>
<td>-----</td>
<td>-------------------</td>
</tr>
<tr>
<td>45</td>
<td>33</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>46</td>
<td>30</td>
<td>M</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>47</td>
<td>44</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>48</td>
<td>44</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>50</td>
<td>41</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>51</td>
<td>61</td>
<td>M</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>52</td>
<td>29</td>
<td>F</td>
<td>Normal</td>
</tr>
<tr>
<td>53</td>
<td>23</td>
<td>M</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>54</td>
<td>33</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>55</td>
<td>34</td>
<td>F</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>56</td>
<td>38</td>
<td>M</td>
<td>Normal</td>
</tr>
<tr>
<td>57</td>
<td>50</td>
<td>M</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>58</td>
<td>55</td>
<td>M</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>59</td>
<td>45</td>
<td>F</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>60</td>
<td>32</td>
<td>F</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>61</td>
<td>45</td>
<td>F</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>62</td>
<td>24</td>
<td>M</td>
<td>Normal</td>
</tr>
<tr>
<td>63</td>
<td>43</td>
<td>M</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>65</td>
<td>50</td>
<td>M</td>
<td>Normal</td>
</tr>
<tr>
<td>66</td>
<td>62</td>
<td>M</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>67</td>
<td>49</td>
<td>M</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>68</td>
<td>38</td>
<td>F</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>69</td>
<td>20</td>
<td>M</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>70</td>
<td>21</td>
<td>M</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>71</td>
<td>24</td>
<td>M</td>
<td>Normal</td>
</tr>
<tr>
<td>72</td>
<td>52</td>
<td>M</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>73</td>
<td>53</td>
<td>F</td>
<td>Ulcerative colitis</td>
</tr>
</tbody>
</table>

F=Female, M=Male
Appendix B. Whole genome sequence assembly data for the isolates examined for this study. Isolates depicted in red text were sequenced on the Illumina HiSeq2000 platform, whilst all other were sequenced on the Illumina MiSeq platform.

<table>
<thead>
<tr>
<th>No.</th>
<th>Patient</th>
<th>Isolate</th>
<th>Phylogroup</th>
<th>35-MiSeq assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N50</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1_3_Ti10</td>
<td>D</td>
<td>40,543</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1_2_C20</td>
<td>B2</td>
<td>102,381</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3_1_AU1</td>
<td>A</td>
<td>71,343</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4_1_C3</td>
<td>D</td>
<td>46,082</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4_2_Ti8</td>
<td>D</td>
<td>38,848</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>5_2_R8</td>
<td>B2</td>
<td>64,647</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>5_1_Ti12</td>
<td>B2</td>
<td>56,976</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>6_1_Ti16</td>
<td>B2</td>
<td>42,097</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>7_3_R4</td>
<td>D</td>
<td>37,681</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>8_1_AC8</td>
<td>D</td>
<td>53,688</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>8_3_DC15</td>
<td>B1</td>
<td>17,008</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>8_3_Ti3</td>
<td>B1</td>
<td>19,685</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
<td>8_2_Ti16</td>
<td>A</td>
<td>40,062</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>8_5_Ti14</td>
<td>A</td>
<td>37,591</td>
</tr>
<tr>
<td>15</td>
<td>9</td>
<td>9_5_R1</td>
<td>B2</td>
<td>41630</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
<td>10_1_R6</td>
<td>B2</td>
<td>37,213</td>
</tr>
<tr>
<td>17</td>
<td>11</td>
<td>11_1_Ti6</td>
<td>D</td>
<td>43,101</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>12_1_Ti12</td>
<td>B2</td>
<td>55,659</td>
</tr>
<tr>
<td>19</td>
<td>12</td>
<td>12_2_Ti13</td>
<td>B2</td>
<td>230,154</td>
</tr>
<tr>
<td>20</td>
<td>13</td>
<td>13_1_C11</td>
<td>B2</td>
<td>67,477</td>
</tr>
<tr>
<td>21</td>
<td>13</td>
<td>13_2_TC5</td>
<td>B2</td>
<td>58,720</td>
</tr>
<tr>
<td>22</td>
<td>13</td>
<td>13_3_TC6</td>
<td>B2</td>
<td>78,315</td>
</tr>
<tr>
<td>23</td>
<td>14</td>
<td>14_1_R13</td>
<td>D</td>
<td>52,243</td>
</tr>
<tr>
<td>24</td>
<td>14</td>
<td>14_2_R14</td>
<td>D</td>
<td>51,128</td>
</tr>
<tr>
<td>25</td>
<td>14</td>
<td>14_4_DC7</td>
<td>D</td>
<td>48,473</td>
</tr>
<tr>
<td>26</td>
<td>15</td>
<td>15_5_C3</td>
<td>B2</td>
<td>76,546</td>
</tr>
<tr>
<td>27</td>
<td>15</td>
<td>15_7_TC2</td>
<td>B2</td>
<td>48,434</td>
</tr>
<tr>
<td>28</td>
<td>16</td>
<td>16_1_Ti5</td>
<td>A</td>
<td>35,267</td>
</tr>
<tr>
<td>29</td>
<td>16</td>
<td>16_2_S18</td>
<td>A</td>
<td>56,783</td>
</tr>
<tr>
<td>30</td>
<td>16</td>
<td>16_3_C13</td>
<td>A</td>
<td>41,632</td>
</tr>
<tr>
<td>31</td>
<td>16</td>
<td>16_4_Ti11</td>
<td>A</td>
<td>38,233</td>
</tr>
<tr>
<td>32</td>
<td>16</td>
<td>16_5_Ti20</td>
<td>A</td>
<td>27,530</td>
</tr>
<tr>
<td>33</td>
<td>17</td>
<td>17_1_R20</td>
<td>F</td>
<td>71,132</td>
</tr>
<tr>
<td>34</td>
<td>17</td>
<td>17_2_DC14</td>
<td>F</td>
<td>35,374</td>
</tr>
<tr>
<td>35</td>
<td>18</td>
<td>18_1_DC10</td>
<td>B1</td>
<td>26,030</td>
</tr>
<tr>
<td>36</td>
<td>18</td>
<td>18_2_AC4</td>
<td>B1</td>
<td>71,927</td>
</tr>
<tr>
<td>37</td>
<td>18</td>
<td>18_3_Ti5</td>
<td>B2</td>
<td>243,933</td>
</tr>
<tr>
<td>38</td>
<td>18</td>
<td>18_4_Ti12</td>
<td>B2</td>
<td>296,258</td>
</tr>
<tr>
<td>39</td>
<td>18</td>
<td>18_3_AC5</td>
<td>B2</td>
<td>37,719</td>
</tr>
<tr>
<td>40</td>
<td>19</td>
<td>19_1_TC4</td>
<td>B2</td>
<td>195,371</td>
</tr>
<tr>
<td>41</td>
<td>19</td>
<td>19_1_Ti7</td>
<td>B2</td>
<td>284,439</td>
</tr>
<tr>
<td>42</td>
<td>19</td>
<td>19_2_TC6</td>
<td>B2</td>
<td>252,687</td>
</tr>
<tr>
<td>43</td>
<td>19</td>
<td>19_2_Ti6</td>
<td>B2</td>
<td>74,789</td>
</tr>
<tr>
<td>44</td>
<td>19</td>
<td>19_3_TC2</td>
<td>B2</td>
<td>33,537</td>
</tr>
<tr>
<td>No.</td>
<td>Patient</td>
<td>Strain</td>
<td>Phylogroup</td>
<td>A5-MiSeq assembly</td>
</tr>
<tr>
<td>-----</td>
<td>---------</td>
<td>--------</td>
<td>------------</td>
<td>------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N50</td>
</tr>
<tr>
<td>45</td>
<td>20</td>
<td>20_1_TC6</td>
<td>B1</td>
<td>67,633</td>
</tr>
<tr>
<td>46</td>
<td></td>
<td>20_2_TC13</td>
<td>B1</td>
<td>122,165</td>
</tr>
<tr>
<td>47</td>
<td></td>
<td>20_3_TC12</td>
<td>B1</td>
<td>114,619</td>
</tr>
<tr>
<td>48</td>
<td></td>
<td>20_5_R7</td>
<td>B2</td>
<td>123,680</td>
</tr>
<tr>
<td>49</td>
<td>21</td>
<td>21_1_TC7</td>
<td>B2</td>
<td>94,461</td>
</tr>
<tr>
<td>50</td>
<td></td>
<td>21_2_C8</td>
<td>B2</td>
<td>89,776</td>
</tr>
<tr>
<td>51</td>
<td>22</td>
<td>22_1_Ti17</td>
<td>D</td>
<td>95,206</td>
</tr>
<tr>
<td>52</td>
<td></td>
<td>22_1_Ti19</td>
<td>D</td>
<td>116,528</td>
</tr>
<tr>
<td>53</td>
<td></td>
<td>22_4_Ti18</td>
<td>D</td>
<td>153,839</td>
</tr>
<tr>
<td>54</td>
<td></td>
<td>22_3_TiNA12</td>
<td>D</td>
<td>61,909</td>
</tr>
<tr>
<td>55</td>
<td>23</td>
<td>23_1_TC4</td>
<td>B2</td>
<td>83,279</td>
</tr>
<tr>
<td>56</td>
<td></td>
<td>23_2_AC12</td>
<td>B2</td>
<td>60,124</td>
</tr>
<tr>
<td>57</td>
<td></td>
<td>23_2_Ti8</td>
<td>B2</td>
<td>48,193</td>
</tr>
<tr>
<td>58</td>
<td></td>
<td>23_3_Ti9</td>
<td>B2</td>
<td>86,000</td>
</tr>
<tr>
<td>59</td>
<td></td>
<td>23_4_R13</td>
<td>B2</td>
<td>84,833</td>
</tr>
<tr>
<td>60</td>
<td>24</td>
<td>24_1_Si1</td>
<td>B1</td>
<td>40,419</td>
</tr>
<tr>
<td>61</td>
<td></td>
<td>24_4_R6</td>
<td>B2</td>
<td>74,152</td>
</tr>
<tr>
<td>62</td>
<td>25</td>
<td>25_1_ADC_2</td>
<td>D</td>
<td>282,422</td>
</tr>
<tr>
<td>63</td>
<td></td>
<td>25_2_DC7</td>
<td>D</td>
<td>65,034</td>
</tr>
<tr>
<td>64</td>
<td></td>
<td>25_3_ADC_20</td>
<td>D</td>
<td>120,600</td>
</tr>
<tr>
<td>65</td>
<td></td>
<td>25_5_NDC20</td>
<td>D</td>
<td>60,092</td>
</tr>
<tr>
<td>66</td>
<td>26</td>
<td>26_1_C13</td>
<td>D</td>
<td>63,604</td>
</tr>
<tr>
<td>67</td>
<td></td>
<td>26_2_C12</td>
<td>D</td>
<td>55,007</td>
</tr>
<tr>
<td>68</td>
<td>27</td>
<td>27_1_R1</td>
<td>B2</td>
<td>123,984</td>
</tr>
<tr>
<td>69</td>
<td>29</td>
<td>29_1_DC3</td>
<td>D</td>
<td>85,983</td>
</tr>
<tr>
<td>70</td>
<td></td>
<td>29_4_DC8</td>
<td>D</td>
<td>65,610</td>
</tr>
<tr>
<td>71</td>
<td></td>
<td>29_2_DC4</td>
<td>B2</td>
<td>94,580</td>
</tr>
<tr>
<td>72</td>
<td></td>
<td>29_3_DC10</td>
<td>B1</td>
<td>64,809</td>
</tr>
<tr>
<td>73</td>
<td>30</td>
<td>30_1_R8</td>
<td>F</td>
<td>169,716</td>
</tr>
<tr>
<td>74</td>
<td></td>
<td>30_2_R7</td>
<td>F</td>
<td>64,655</td>
</tr>
<tr>
<td>75</td>
<td>32</td>
<td>32_1_DC18</td>
<td>B2</td>
<td>115,451</td>
</tr>
<tr>
<td>76</td>
<td></td>
<td>32_2_R12</td>
<td>B2</td>
<td>94,553</td>
</tr>
<tr>
<td>77</td>
<td>33</td>
<td>33_1_TC19</td>
<td>B2</td>
<td>200,666</td>
</tr>
<tr>
<td>78</td>
<td></td>
<td>33_3_Ti5</td>
<td>B2</td>
<td>88,880</td>
</tr>
<tr>
<td>79</td>
<td></td>
<td>33_2_AC11</td>
<td>B2</td>
<td>94,936</td>
</tr>
<tr>
<td>80</td>
<td></td>
<td>33_3_R3</td>
<td>B1</td>
<td>102,978</td>
</tr>
<tr>
<td>81</td>
<td>35</td>
<td>35_9_Ti14</td>
<td>B2</td>
<td>117,484</td>
</tr>
<tr>
<td>82</td>
<td>36</td>
<td>36_1_TC9</td>
<td>B2</td>
<td>114,927</td>
</tr>
<tr>
<td>83</td>
<td></td>
<td>36_1_Ti13</td>
<td>B2</td>
<td>135,729</td>
</tr>
<tr>
<td>84</td>
<td>37</td>
<td>37_2_AC20</td>
<td>B1</td>
<td>41,646</td>
</tr>
<tr>
<td>85</td>
<td>39</td>
<td>39_1_AC10</td>
<td>B2</td>
<td>88,788</td>
</tr>
<tr>
<td>86</td>
<td></td>
<td>39_2_Ti18</td>
<td>B2</td>
<td>234,240</td>
</tr>
<tr>
<td>87</td>
<td>40</td>
<td>40_1_R9</td>
<td>B2</td>
<td>95,952</td>
</tr>
<tr>
<td>88</td>
<td></td>
<td>40_3_R7</td>
<td>B2</td>
<td>63,667</td>
</tr>
<tr>
<td>89</td>
<td>41</td>
<td>41_1_Ti9</td>
<td>D</td>
<td>190,932</td>
</tr>
<tr>
<td>90</td>
<td></td>
<td>41_2_Ti13</td>
<td>D</td>
<td>195,608</td>
</tr>
<tr>
<td>91</td>
<td>42</td>
<td>42_1_R2</td>
<td>A</td>
<td>36,883</td>
</tr>
<tr>
<td>No.</td>
<td>Patient</td>
<td>Strain</td>
<td>Phylogroup</td>
<td>A5-MiSeq assembly</td>
</tr>
<tr>
<td>-----</td>
<td>---------</td>
<td>----------</td>
<td>------------</td>
<td>-------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N50</td>
</tr>
<tr>
<td>92</td>
<td>43</td>
<td>43_1_R3</td>
<td>B1</td>
<td>82,187</td>
</tr>
<tr>
<td>93</td>
<td>43</td>
<td>43_4_AC3</td>
<td>B1</td>
<td>158,385</td>
</tr>
<tr>
<td>94</td>
<td>43</td>
<td>43_2_R17</td>
<td>B1</td>
<td>49,515</td>
</tr>
<tr>
<td>95</td>
<td>44</td>
<td>44_1_Ti4</td>
<td>B1</td>
<td>87,815</td>
</tr>
<tr>
<td>96</td>
<td>44</td>
<td>44_2_Ti11</td>
<td>B1</td>
<td>56,705</td>
</tr>
<tr>
<td>97</td>
<td>44</td>
<td>44_4_Ti15</td>
<td>B1</td>
<td>114,847</td>
</tr>
<tr>
<td>98</td>
<td>45</td>
<td>45_1_Ti1</td>
<td>B2</td>
<td>146,037</td>
</tr>
<tr>
<td>99</td>
<td>45</td>
<td>45_1_HU2</td>
<td>B2</td>
<td>132,312</td>
</tr>
<tr>
<td>100</td>
<td>45</td>
<td>45_3_DC2</td>
<td>B2</td>
<td>114,683</td>
</tr>
<tr>
<td>101</td>
<td>45</td>
<td>45_2_HU10</td>
<td>A</td>
<td>153,160</td>
</tr>
<tr>
<td>102</td>
<td>45</td>
<td>45_7_HU9</td>
<td>A</td>
<td>92,985</td>
</tr>
<tr>
<td>103</td>
<td>46</td>
<td>46_1_Ti2</td>
<td>B2</td>
<td>358,431</td>
</tr>
<tr>
<td>104</td>
<td>46</td>
<td>46_2_Ti11</td>
<td>D</td>
<td>107,383</td>
</tr>
<tr>
<td>105</td>
<td>47</td>
<td>47_1_TC4</td>
<td>D</td>
<td>297,445</td>
</tr>
<tr>
<td>106</td>
<td>47</td>
<td>47_3_R4</td>
<td>B1</td>
<td>211,746</td>
</tr>
<tr>
<td>107</td>
<td>47</td>
<td>47_4_R1</td>
<td>D</td>
<td>311,185</td>
</tr>
<tr>
<td>108</td>
<td>48</td>
<td>48_1_R18</td>
<td>B2</td>
<td>339,742</td>
</tr>
<tr>
<td>109</td>
<td>48</td>
<td>48_5_R19</td>
<td>B2</td>
<td>160,035</td>
</tr>
<tr>
<td>110</td>
<td>50</td>
<td>50_4_ACUA1</td>
<td>B1</td>
<td>86,584</td>
</tr>
<tr>
<td>111</td>
<td>50</td>
<td>50_5_TCA1</td>
<td>B1</td>
<td>77,074</td>
</tr>
<tr>
<td>112</td>
<td>50</td>
<td>50_6_TCUA1</td>
<td>A</td>
<td>96,473</td>
</tr>
<tr>
<td>113</td>
<td>50</td>
<td>50_7_TCUA2</td>
<td>B1</td>
<td>27,450</td>
</tr>
<tr>
<td>114</td>
<td>51</td>
<td>51_1_Ti1</td>
<td>B1</td>
<td>197,462</td>
</tr>
<tr>
<td>115</td>
<td>51</td>
<td>51_2_R1</td>
<td>D</td>
<td>138,644</td>
</tr>
<tr>
<td>116</td>
<td>51</td>
<td>51_3_R19</td>
<td>D</td>
<td>180,933</td>
</tr>
<tr>
<td>117</td>
<td>51</td>
<td>51_6_R16</td>
<td>D</td>
<td>134,447</td>
</tr>
<tr>
<td>118</td>
<td>51</td>
<td>51_7_R18</td>
<td>D</td>
<td>104,075</td>
</tr>
<tr>
<td>119</td>
<td>52</td>
<td>52_1_Ti3</td>
<td>B2</td>
<td>242,632</td>
</tr>
<tr>
<td>120</td>
<td>52</td>
<td>52_2_Ti10</td>
<td>B2</td>
<td>219,999</td>
</tr>
<tr>
<td>121</td>
<td>53</td>
<td>53_1_AC6</td>
<td>B1</td>
<td>84,028</td>
</tr>
<tr>
<td>122</td>
<td>53</td>
<td>53_1_AUC1</td>
<td>B1</td>
<td>134,422</td>
</tr>
<tr>
<td>123</td>
<td>54</td>
<td>54_1_TC4</td>
<td>B2</td>
<td>553,572</td>
</tr>
<tr>
<td>124</td>
<td>54</td>
<td>54_1_Ti6</td>
<td>B2</td>
<td>653,360</td>
</tr>
<tr>
<td>125</td>
<td>55</td>
<td>55_1_TC9</td>
<td>B2</td>
<td>355,821</td>
</tr>
<tr>
<td>126</td>
<td>55</td>
<td>55_1_Ti19</td>
<td>B2</td>
<td>222,239</td>
</tr>
<tr>
<td>127</td>
<td>56</td>
<td>56_1_TC1</td>
<td>D</td>
<td>178,338</td>
</tr>
<tr>
<td>128</td>
<td>56</td>
<td>56_1_AC4</td>
<td>D</td>
<td>106,670</td>
</tr>
<tr>
<td>129</td>
<td>56</td>
<td>56_3_AC1</td>
<td>D</td>
<td>133,916</td>
</tr>
<tr>
<td>130</td>
<td>56</td>
<td>56_4_R1</td>
<td>D</td>
<td>139,738</td>
</tr>
<tr>
<td>131</td>
<td>56</td>
<td>56_2_AC5</td>
<td>B2</td>
<td>287,062</td>
</tr>
<tr>
<td>132</td>
<td>57</td>
<td>57_2_DC4</td>
<td>A</td>
<td>86,839</td>
</tr>
<tr>
<td>133</td>
<td>57</td>
<td>57_3_TC14</td>
<td>B2</td>
<td>168,969</td>
</tr>
<tr>
<td>134</td>
<td>57</td>
<td>57_3_Ti5</td>
<td>B2</td>
<td>407,054</td>
</tr>
<tr>
<td>135</td>
<td>57</td>
<td>57_5_R8</td>
<td>B2</td>
<td>207,989</td>
</tr>
<tr>
<td>136</td>
<td>58</td>
<td>58_2_AC1</td>
<td>B2</td>
<td>167,308</td>
</tr>
<tr>
<td>137</td>
<td>58</td>
<td>58_3_Ti7</td>
<td>A</td>
<td>73,351</td>
</tr>
<tr>
<td>138</td>
<td>59</td>
<td>59_1_AC3</td>
<td>A</td>
<td>70,024</td>
</tr>
<tr>
<td>No.</td>
<td>Patient</td>
<td>Strain</td>
<td>Phylogroup</td>
<td>A5-MiSeq assembly</td>
</tr>
<tr>
<td>-----</td>
<td>---------</td>
<td>------------</td>
<td>------------</td>
<td>-------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N50</td>
</tr>
<tr>
<td>139</td>
<td>60</td>
<td>60_1_Ti1</td>
<td>B2</td>
<td>429,061</td>
</tr>
<tr>
<td>140</td>
<td>61</td>
<td>61_1_Ti1</td>
<td>B2</td>
<td>245,678</td>
</tr>
<tr>
<td>141</td>
<td></td>
<td>61_2_AC9</td>
<td>B2</td>
<td>146,570</td>
</tr>
<tr>
<td>142</td>
<td>62</td>
<td>62_1_Ti3</td>
<td>B2</td>
<td>371,257</td>
</tr>
<tr>
<td>143</td>
<td></td>
<td>62_2_Ti6b</td>
<td>B2</td>
<td>102,671</td>
</tr>
<tr>
<td>144</td>
<td></td>
<td>62_4_Si7</td>
<td>B1</td>
<td>110,472</td>
</tr>
<tr>
<td>145</td>
<td>63</td>
<td>63_1_Ti1</td>
<td>D</td>
<td>183,569</td>
</tr>
<tr>
<td>146</td>
<td></td>
<td>63_2_TC7</td>
<td>B2</td>
<td>66,633</td>
</tr>
<tr>
<td>147</td>
<td>65</td>
<td>65_2_AC1</td>
<td>A</td>
<td>57,830</td>
</tr>
<tr>
<td>148</td>
<td></td>
<td>65_1_AC2</td>
<td>D</td>
<td>47,226</td>
</tr>
<tr>
<td>149</td>
<td></td>
<td>65_3_TC1</td>
<td>D</td>
<td>35,971</td>
</tr>
<tr>
<td>150</td>
<td></td>
<td>65_4_AC20</td>
<td>D</td>
<td>121,976</td>
</tr>
<tr>
<td>151</td>
<td>66</td>
<td>66_1_AC1</td>
<td>D</td>
<td>46,650</td>
</tr>
<tr>
<td>152</td>
<td>67</td>
<td>67_1_Ti13</td>
<td>D</td>
<td>49,694</td>
</tr>
<tr>
<td>153</td>
<td></td>
<td>67_3_Ti10</td>
<td>D</td>
<td>37,367</td>
</tr>
<tr>
<td>154</td>
<td></td>
<td>67_4_Ti7</td>
<td>A</td>
<td>57,737</td>
</tr>
<tr>
<td>155</td>
<td>68</td>
<td>68_2_Ti3</td>
<td>B1</td>
<td>41,496</td>
</tr>
<tr>
<td>156</td>
<td></td>
<td>68_4_TC14</td>
<td>B1</td>
<td>48,403</td>
</tr>
<tr>
<td>157</td>
<td></td>
<td>68_3_TC19</td>
<td>D</td>
<td>39,140</td>
</tr>
<tr>
<td>158</td>
<td>69</td>
<td>69_1_TC1</td>
<td>B2</td>
<td>58,220</td>
</tr>
<tr>
<td>159</td>
<td></td>
<td>69_1_Ti1</td>
<td>B2</td>
<td>662,554</td>
</tr>
<tr>
<td>160</td>
<td></td>
<td>69_2_R20</td>
<td>B2</td>
<td>148,853</td>
</tr>
<tr>
<td>161</td>
<td>70</td>
<td>70_1_AC3</td>
<td>D</td>
<td>48,361</td>
</tr>
<tr>
<td>162</td>
<td></td>
<td>70_2_TC2</td>
<td>D</td>
<td>152,551</td>
</tr>
<tr>
<td>163</td>
<td></td>
<td>70_4_DC20</td>
<td>B2</td>
<td>120,032</td>
</tr>
<tr>
<td>164</td>
<td></td>
<td>70_5_R4</td>
<td>B2</td>
<td>164,456</td>
</tr>
<tr>
<td>165</td>
<td></td>
<td>70_2_Ti12</td>
<td>B2</td>
<td>226,521</td>
</tr>
<tr>
<td>166</td>
<td></td>
<td>70_3_Ti8</td>
<td>D</td>
<td>83,335</td>
</tr>
<tr>
<td>167</td>
<td>71</td>
<td>71_1_AC1</td>
<td>F</td>
<td>73,839</td>
</tr>
<tr>
<td>168</td>
<td></td>
<td>71_2_Ti5</td>
<td>D</td>
<td>163,406</td>
</tr>
<tr>
<td>169</td>
<td>72</td>
<td>72_2_AC9</td>
<td>B2</td>
<td>84,746</td>
</tr>
<tr>
<td>170</td>
<td></td>
<td>72_3_Tc1</td>
<td>B2</td>
<td>34,270</td>
</tr>
<tr>
<td>171</td>
<td></td>
<td>72_4_T9</td>
<td>D</td>
<td>32,808</td>
</tr>
<tr>
<td>172</td>
<td></td>
<td>72_5_T14</td>
<td>D</td>
<td>54,421</td>
</tr>
<tr>
<td>173</td>
<td></td>
<td>72_6_T12</td>
<td>B2</td>
<td>654,244</td>
</tr>
<tr>
<td>174</td>
<td>73</td>
<td>73_1_Si1</td>
<td>B1</td>
<td>58,250</td>
</tr>
</tbody>
</table>
Appendix C: Two independent culture of the same strain of 19 samples were sequenced twice and used to find the SNPs between same strains.

<table>
<thead>
<tr>
<th>No.</th>
<th>Strain</th>
<th>Phylogroup</th>
<th>Source</th>
<th>A5-MiSeq assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N50</td>
</tr>
<tr>
<td>1</td>
<td>B271_G5</td>
<td>F</td>
<td>Bird faecal</td>
<td>22,665</td>
</tr>
<tr>
<td>2</td>
<td>B271_S26</td>
<td></td>
<td></td>
<td>153,908</td>
</tr>
<tr>
<td>3</td>
<td>B1549_00</td>
<td>F</td>
<td>Bird faecal</td>
<td>179,247</td>
</tr>
<tr>
<td>4</td>
<td>B1549_S36</td>
<td></td>
<td></td>
<td>153,815</td>
</tr>
<tr>
<td>5</td>
<td>W3_62_S12</td>
<td>A</td>
<td>Chicken</td>
<td>78,252</td>
</tr>
<tr>
<td>6</td>
<td>W3_62_S18</td>
<td></td>
<td></td>
<td>84,079</td>
</tr>
<tr>
<td>7</td>
<td>B313_S6</td>
<td>B1</td>
<td>Bird faecal</td>
<td>123,956</td>
</tr>
<tr>
<td>8</td>
<td>B313_S7</td>
<td></td>
<td></td>
<td>85,252</td>
</tr>
<tr>
<td>9</td>
<td>5_1_TC12</td>
<td>B2</td>
<td>Biopsy</td>
<td>56,976</td>
</tr>
<tr>
<td>10</td>
<td>5_1_TC12_redo</td>
<td></td>
<td></td>
<td>82,264</td>
</tr>
<tr>
<td>11</td>
<td>8_2_Ti16</td>
<td>A</td>
<td>Biopsy</td>
<td>40,062</td>
</tr>
<tr>
<td>12</td>
<td>8_2_Ti16_redo</td>
<td></td>
<td></td>
<td>54,647</td>
</tr>
<tr>
<td>13</td>
<td>8_3_Ti3</td>
<td>B1</td>
<td>Biopsy</td>
<td>19,685</td>
</tr>
<tr>
<td>14</td>
<td>8_3_Ti3_redo</td>
<td></td>
<td></td>
<td>19,703</td>
</tr>
<tr>
<td>15</td>
<td>B1932_GS</td>
<td>F</td>
<td>Bird faecal</td>
<td>185,361</td>
</tr>
<tr>
<td>16</td>
<td>B1932_S46</td>
<td></td>
<td></td>
<td>58,361</td>
</tr>
<tr>
<td>17</td>
<td>DMG_G1_S33</td>
<td>B2</td>
<td>Human faecal</td>
<td>44,169</td>
</tr>
<tr>
<td>18</td>
<td>DMG_G1_S38</td>
<td></td>
<td></td>
<td>84,641</td>
</tr>
<tr>
<td>19</td>
<td>DMG_G2_S34</td>
<td>B2</td>
<td>Human faecal</td>
<td>52,336</td>
</tr>
<tr>
<td>20</td>
<td>DMG_G2_S39</td>
<td></td>
<td></td>
<td>45,838</td>
</tr>
<tr>
<td>21</td>
<td>DMG_G3_S35</td>
<td>B2</td>
<td>Human faecal</td>
<td>71,636</td>
</tr>
<tr>
<td>22</td>
<td>DMG_G3_S40</td>
<td></td>
<td></td>
<td>57,055</td>
</tr>
<tr>
<td></td>
<td>Sample ID</td>
<td>Group</td>
<td>Sample Type</td>
<td>Reads</td>
</tr>
<tr>
<td>---</td>
<td>-----------</td>
<td>-------</td>
<td>-----------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>23</td>
<td>H173_GS</td>
<td>C</td>
<td>Human faecal</td>
<td>173,372</td>
</tr>
<tr>
<td></td>
<td>H173_S13</td>
<td></td>
<td></td>
<td>222,627</td>
</tr>
<tr>
<td>25</td>
<td>H440_GS</td>
<td>A</td>
<td>Human Blood</td>
<td>198,974</td>
</tr>
<tr>
<td></td>
<td>H440_S14</td>
<td></td>
<td></td>
<td>56,160</td>
</tr>
<tr>
<td>27</td>
<td>E2845_S33</td>
<td>B1</td>
<td>Water</td>
<td>155,561</td>
</tr>
<tr>
<td></td>
<td>E2845_S48</td>
<td></td>
<td></td>
<td>84,824</td>
</tr>
<tr>
<td>29</td>
<td>M1213_GS</td>
<td>E</td>
<td>Dasyurus maculatus faecal</td>
<td>121,977</td>
</tr>
<tr>
<td></td>
<td>M1213_S22</td>
<td></td>
<td></td>
<td>89,086</td>
</tr>
<tr>
<td>31</td>
<td>18_1_DC10</td>
<td>B1</td>
<td>Biopsy</td>
<td>26,030</td>
</tr>
<tr>
<td></td>
<td>18_1_DC10_redo</td>
<td></td>
<td></td>
<td>69,005</td>
</tr>
<tr>
<td>33</td>
<td>A1_66_S14</td>
<td>B2</td>
<td>Chicken meat</td>
<td>93,185</td>
</tr>
<tr>
<td></td>
<td>A1_66_S44</td>
<td></td>
<td></td>
<td>164,260</td>
</tr>
<tr>
<td>35</td>
<td>B1858_GS</td>
<td>F</td>
<td>Bird faecal</td>
<td>344,604</td>
</tr>
<tr>
<td></td>
<td>B1858_S45</td>
<td></td>
<td></td>
<td>89,136</td>
</tr>
<tr>
<td>37</td>
<td>W1_57_S13</td>
<td>B2</td>
<td>Chicken meat</td>
<td>156,811</td>
</tr>
<tr>
<td></td>
<td>W1_57_S27</td>
<td></td>
<td></td>
<td>133,412</td>
</tr>
</tbody>
</table>