The
University
Of
Sheffield.

Insights into the evolution of *Escherichia coli*

By:

Matthew Lawrence Howes MSc

A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Science
Department of Molecular Biology and Biotechnology

28th September 2018

Acknowledgements

I would like to thank my supervisor Dr Roy Chaudhuri for his support and for the useful discussions which helped me plan and conduct my investigations. I would also like to thank Dr Rachel Floyd and Professor Craig Winstanley for providing me with the *E. coli* urinary tract infection phenotype data which formed the basis for the work presented in Chapter 5. Finally I would like to thank the University of Sheffield and the Department of Molecular Biology and Biotechnology who provided me with the funding which enabled me to carry out this PhD.

Insights into the evolution of *Escherichia coli*

Contents

List of figures

List of tables

Abstract

E. coli is a highly diverse commensal and environmental organism also associated with a broad range of infectious diseases. The work presented in this thesis provides insights into the evolutionary history and genetic diversity of the species through using up-to-date genetically representative sets of genomes. Chapter 3 provides insights into the *E. coli* pan genome, rates of recombination, the clonal phylogeny, genes associated with major evolutionary diversification events, and identifies a potential new phylogenetic group, tentatively labelled 'G'. Chapter 4 provides evidence to support the use of a novel 256-gene *E. coli* core gene multi-locus sequence type (MLST) schema as reliable for assigning clonal and phylogenetic groups to strains in evolutionary studies, and as an alternative to using a core gene phylogeny, other core gene MLST schemas, or 7-15 locus MLST or multiplex schema methods. A novel 7-gene MLST schema and a 10-locus multiplex schema were also developed and presented which are inferred at their current stage of development to provide 100% correct phylogenetic group assignment to *E. coli* strains. Chapter 5 is an investigation to determine presence and evolutionary insights of genes strongly associated with decreases in ureter contractility observed during the early *E. coli* colonisation stages of mild to severe urinary tract infections. Contractility decrease phenotypes were found to be significantly linked to strains with genes from a specific set including those encoding two haemolysin operons, nitric oxide stress resistance proteins, and zinc and potassium uptake proteins. The final research chapter reported an analysis of the evolutionary history and genetic diversity of the *E. coli* type three secretion system 2 locus (ETT2), and the associated *eip* cluster. Previously unreported ETT2 cluster genes were identified in the genus *Citrobacter*, the species *E. fergusonii*, and the *E. coli* cryptic clades. Widespread truncations and deletions were found in specific lineages, together with multiple horizontal transfer events of ETT2 genes in group C-I strains.

Chapter 1: Introduction

1.1. *Escherichia coli*

Escherichia coli is a highly diverse facultatively anaerobic Gram negative bacterial species found in the environment and within animals (Savageau 1983). The bacterium is commonly present as an innocuous commensal organism in the mammalian gut, but there are also pathogenic forms that can cause mild to severe infection in hosts (Kaper *et al.* 2004). However, there is not a clear distinction between 'commensal' and 'pathogenic' *E. coli* as some commensals will cause disease if present in a particular site, such as the urinary tract in mammals but not in the intestinal region (Donnenberg 2002). Similarly, O157:H7 is a commensal 'strain' (defined as bacterial cells descended from a single genetically homogenous cell colony (Dijkshoorn et al. 2000)) in cows, but is pathogenic in humans (Jay *et al.* 2007).

Although uncommon, commensal forms of *E. coli* can be transformed by the acquisition of sequence regions from pathogenic forms by horizontal transfer (HT). These can be in the form of insertions containing one to ten genes or 'genomic island' mobile genetic elements (MGEs) up to 200 kb in length containing genes, and can encode a huge variety of proteins, some of which are termed virulence-associated factors (VAFs) as they can initiate and maintain disease symptoms in the host upon their expression (Ngeleka *et al.* 1996, Sussman 1997). Some types of VAFs include adhesins, toxins, secretion systems, membrane transporters, signalling structures, degradative enzymes, and capsules, which from the bacterial point of view, can be considered as "fitness factors", which enable virulent strains to infect hosts, overcome immune defences and exploit additional niches within the host, to promote survival and proliferation (Kaper *et al.* 2004).

Pathogenic *E. coli* are classified into different pathovars (Table 1.1), collectively capable of producing an extensive range of pathologies in human and animal hosts (Dobrindt *et al.* 2004). To humans, most *E. coli* pathovars represent a mild health risk in the western world, but some are potentially life-threatening (Croxen et al. 2013). However elsewhere they are more of an issue, such as with enterotoxigenic *E. coli* (ETEC)-mediated diarrhoea being the cause of significant mortality in under 5s in sub-Saharan Africa and south Asia (Kotloff et al. 2013). Pathogenic *E. coli* have been studied for over one hundred years because of their affinity with human disease (Kaper *et al.* 2004). *E. coli* deemed to cause the same pathologies are denoted with the same pathovar identity, with VAFs which can be highly diverse and present, absent, exclusive to the pathovar, or present in multiple pathovars. Pathovars active within the intestinal area are denoted diarrhoeagenic *E. coli* (DEC), while those active in host bodily regions outside the intestines are denoted extraintestinal pathogenic *E. coli* (ExPEC) (Dobrindt 2005, Kohler and Dobrindt 2011).

Table 1.1. Details of the twelve defined *E. coli* pathovars

	Pathovar	**Pathovar acronym**	**Disease(s) associated**	**Example strain**
Diarrhoeagenic *E. coli* (DEC)	Adherent invasive *E. coli*	AIEC	Crohn's disease	LF82
	Diffusely adherent *E. coli*	DAEC	Diarrhoeal disease	A22
	Enterotoxigenic *E. coli*	ETEC	Diarrhoeal disease	E24377A
	Enteropathogenic *E. coli* (typical and atypical)	tEPEC, aEPEC	Diarrhoeal disease	B171 O26: H11 C814-67
	Enteroinvasive *E. coli*	EIEC	Bacillary dysentery	53638
	Enteroaggregative *E. coli*	EAEC	Hemorrhagic colitis (HC), haemolytic uremic syndrome (HUS)	55989
	Enterohemorrhagic *E. coli*	EHEC	Hemorrhagic colitis, haemolytic uremic syndrome	O157:H7 EDL933
Extraintestinal pathogenic *E. coli* (ExPEC)	Avian pathogenic *E. coli*	APEC	Avian colibacillosis	APEC O1
	Neonatal Meningitis-associated *E. coli*	NMEC	Neonatal meningitis	RS218
	Septicaemia-associated *E. coli*	SEPEC	Septicaemia, bloodstream infection	SEPEC 06
	Uropathogenic *E. coli*	UPEC	Urinary tract infections (UTI), Kidney failure, bloodstream infection	CFT073
	Mammary pathogenic *E. coli*	MPEC	Mastitis	P4

1.2. Molecular evolution and phylogenetics

Genes are the physical and functional structures of heredity which are inherited directly from ancestors. Although replication mechanisms work to create identical copies of a DNA strand, mutational errors occur leaving substitutions (e.g. A instead of T), insertions (e.g. a new A between CG), or deletions (e.g. CGC to CC) in the copied sequence at a semi-predictable rate. When comparing related sequences, it is often ambiguous whether an insertion has occurred in one sequence, or a deletion has occurred in the other, so insertions and deletions are collectively referred to as 'indels'. When an indel occurs in a gene, it can have a significant impact on the structure of the protein it encodes. Within a gene, each set of 3 adjacent nucleotides from the start are each referred to as a 'codon' (Crick 1968, Streisinger et al. 1966, Wang et al. 2001, Koonin and Novozhilov 2009). Codons make up the genetic code of a gene and each codon directly contributes to determining the structure of the protein that the gene encodes as it is the order of the codons which give proteins their unique function (Crick 1968, Streisinger et al. 1966, Wang et al. 2001). Codons are assigned from the start of the gene sequence in sets of the 3 adjacent nucleotides so bases 1, 2, and 3 are assigned a specific codon, and bases 4, 5, and 6 are assigned the next codon. Start and stop codons are the first and last codons assigned for a gene sequence and mark the start and end of translation of a gene sequence into a protein sequence. When bases that are not in multiples of 3 are inserted or deleted, there is a shift in the reading frame, meaning that different codons are present along the rest of the coding sequence, and a different amino acid sequence is encoded (Streisinger et al. 1966, Koonin and Novozhilov 2009). For example, CCCTCT encodes Phenylalanine then Leucine, but a deletion of the second 'C' (CCTCT-) produces Proline and two left over bases 'CT'. Although indels can produce new functionality for a gene, they more commonly halt translation through causing the introduction of a premature stop codon (Streisinger et al. 1966). Proteins with halted translation are shorter and typically exhibit partial or no functionality (Streisinger et al. 1966).

When a base substitution occurs in a codon within a gene sequence, there is no truncation and the sequence length is retained (Crick 1968, Streisinger et al. 1966). However, if the substitution introduces or removes a start or stop codon, then this will lead to premature truncation or elongation of the gene sequence. All other base substitutions within a coding sequence can be classified as synonymous or non-synonymous depending on how that substitution affects the codon in which it lies (Crick 1968, Streisinger et al. 1966, Koonin and Novozhilov 2009, Copeland 2003). Synonymous substitutions produce the same amino acid despite the codon changing because of the redundancy in the code (for example a change from CAT to CAC will still encode Histidine) and the resulting protein is unchanged by the substitution (Crick 1968). Nonsynonymous substitutions produce a different amino acid with the new codon (AGC produces Serine, but AGA produces Arginine), and results in the substitution of a single amino acid in the encoded protein. This single change will either not affect the protein (neutral effect), impede its function (deleterious effect), or improve it (beneficial effect). Fitness is a concept which describes a bacterial cell's ability to replicate and survive in an environment (Subashchandrabose et al. 2015, Barrick et al. 2009). A quantification of the contribution of a gene to the fitness of the organism is often termed the fitness value or score (Barrick et al. 2009. Rubin et al. 2015). A deleterious mutation in a gene with a high contribution to the fitness of the organism would result in a reduced ability of the cell to replicate and survive in its current environment (Barrick et al. 2009, Rubin et al. 2015, Wiles et al. 2013). However, a beneficial mutation would improve it while a neutral mutation would not change its ability (Barrick et al. 2009, Wiles et al. 2013). Genes with a high fitness value may be subject to 'positive selection' whereby bacterial cells with the mutated gene preferentially replicate and survive over those without it, making the gene increasingly common in the population (Barrick et al. 2009, Wiles et al. 2013). This also occurs after gene duplication events through errors in DNA replication and repair (Yamanaka et al. 1998, Fyodor et al. 2002). Mutations caused by duplications, indels, and substitutions are therefore a significant driving force behind evolutionary development (Barrick et al. 2009, Fyodor et al. 2002). However specific *E. coli* genes encoding 'housekeeping' functions such as those relating to aerobic respiration (*arcA*), RNA polymerase production (*rpoS*), and cell membrane, nucleus, and transport system-associated functions essential to general cell survival typically exhibit stronger selection against nonsynonymous substitutions than other genes (Viscidi and Demma 2003, Reid et al. 2000). This is because they have high fitness values in their current states, meaning that almost all nonsynonymous substitutions will be deleterious (Viscidi and Demma 2003). Selection against deleterious substitutions is referred to as "purifying selection" (Jordan et al. 2002).

1.2.1. Horizontal transfer via mobile genetic elements

Horizontal transfer (HT) is an important factor in the molecular biology of bacteria and must be considered when making inferences about the evolutionary history of a species or simply the relatedness of a set of strains (Didelot et al. 2012). It is a process which allows genes from a donor genome or plasmid to be acquired by recipient cell genomes or cell plasmids (Hanahan 1983, Ochman et al. 2000). *E. coli* genomes include many "accessory genes" which are not conserved across all strains. These genes are important for survival in specific environments and are commonly found on MGEs such as insertion sequence (IS) elements, plasmids, and prophage (Dobrindt et al. 2004, Thomson et al. 2004). MGEs facilitate the phenomena of HT and enable the transfer of DNA segments ranging from a few bases to individual genes or operons (adjacent genes with products functioning as a single unit) and even genomic "islands" spanning several hundred kilobases (Ochman et al. 2000, Ren et al. 2004). Genes acquired via HT can be incorporated into the genome via a process called homologous recombination (referred to just as 'recombination' in this thesis) and may replace the native copy of that gene (Didelot and Maiden 2010) (Nehra et al. 2017).

1.2.2. Conjugation, transformation, and transduction

Conjugation, transformation, and transduction are the mechanisms by which genetic material is transferred between two bacterial cells in HT. Conjugation is the physical genetic transfer of DNA through a mating pore which spans between the membranes of both cells. (Guglielmini et al. 2011). Generally, MGE segments of up to 100 kb are transferred in conjugation. Plasmids (linear or circular double stranded extrachromosomal DNA) are often transferred through the pore and are common vectors of DNA transfer in conjugation (Burrus et al. 2002). Transferred DNA may then be inserted into the genome through the activity of MGEs such as transposons, insertion sequences, and phage genes, or replace an existing sequence via recombination.

Transformation is where naked DNA strands from the surrounding intercellular environment (typically left after cell lysis) are taken up by the cell and incorporated into the genome via insertion or recombination (Hanahan et al. 1983, Ochman et al. 2000). Only a minority percentage of bacterial cells in a uniform or mixed population can uptake this DNA. Such "competent cells" express factors including cell wall modifications that allow DNA to be bound and uptaken into the cell (Hanahan, 1983; Nielsen & Van Elsas, 2001). The acquired DNA may then be inserted into the genome through MGE activity or replace an existing sequence via recombination (Hanahan et al. 1983, Ochman et al. 2000).

Transduction is the transfer of DNA between bacteria mediated by bacterial viruses called bacteriophages (Schicklmaier and Schmieger 1995, Ochman et al. 2000). With generalised transduction, a random section of bacterial DNA sequence becomes enclosed in a viral capsid protein (also referred to as a viral envelope) during viral packaging in an infected donor cell (Ochman et al. 2000). This can occur via 'headful packaging' whereby the bacteriophage tends to incorporate non-viral DNA into its genome if there is size capacity (≤110 kb) in its capsid (Coren et al. 1995). Upon replication of the bacteriophage, lysis of the donor, and subsequent infection of another bacterium, recombination occurs between the donor DNA sequence transferred during infection and the recipient genome's homologous DNA sequence after the virus takes control of the cell in order to replicate its own DNA (Goh et al. 2016, Ochman et al. 2000). In specialised transduction, the bacteriophage inserts its DNA into the host chromosome (called a 'prophage' once integrated) through involvement of the bacterial RecA and RecBCD enzymes, and the prophage is then replicated with the surrounding bacterial genome (Thomason et al. 2007, Chen et al. 2018). Then during induction, viral DNA is enzymatically excised and is packaged in a protein capsid to form a bacteriophage particle (Goh et al. 2016, Ochman et al. 2000). When the phage genes excise from the host chromosome, host chromosomal DNA can be erroneously incorporated into the bacteriophage genome by imprecise excision followed by packaging where a protein capsule surrounds the DNA to become part of the bacteriophage (Chen et al. 218, Ochman et al. 2000). The bacteriophage then replicates, lyses the cell, matures, and reinfects a different recipient cell, inserting its prophage-host DNA sequence mix into the genome which becomes integrated with other chromosomal genes (Chen et al. 2018, Ochman et al. 2000). A recipient cell with the integrated DNA is referred to as lysogenic cell (Goh et al. 2016). The quantity of DNA transferred in transduction events is limited by the phage capsule size but can reach 100 kb (Thomason et al. 2007, Ochman et al. 2000).

1.2.3. Insertion sequences

Insertion sequence elements are MGEs which facilitate self-transposition, typically shorter than 2.5 kb and commonly found on plasmids or inserted into bacterial chromosomes that encode genes (Mahillon and Chandler 1998, Mahillon et al. 1999). Their structure usually includes a central transposase gene (Tpase) and a promotor sequence which initiates transcription of the Tpase gene, surrounded by inverted repeated base (IR) sequences that define the IS element borders (Mahillon and Chandler, 1998). In *E. coli* IRs range from 10 to 40 bp in length (Mahillon et al. 1999). IS elements typically only encode functions which mediate their own translocation; movement of the whole sequence which can hop between locations within the same genome or between genomes through insertion at the point of a target sequence, with the central Tpase genes carrying out the transposition (Mahillon and Chandler, 1998; Mahillon et al. 1999). The donor sequence can then be transcribed and translated with surrounding genes (Mahillon et al. 1999). IS elements additionally have a role in gene deletion in bacterial genomes. When two copies of the same IS element exist in the same genome, recombination can occur between the repeat sequences in the IS which can lead to a deletion of genes situated between the repeat sequences. IS elements thus can play an important role in initiating genome HT, deletions, rearrangement, and diversification (Siguier et al. 2006).

In a long-term evolution study (Lee et al. 2016), 41 copies of ISs from 14 families were found between 520 *E. coli* strains descended from a founding *E. coli* strain labelled PFM2. The study monitored genome rearrangements with lengths of 1-40 kb in all isolates over a total of 2.2 million generations and found 758 novel insertions and 98 recombination events to have occurred as a result of the ISs (Lee et al. 2016). A similar study identified 110 rearrangements with lengths of ≥5 kb over 40,000 generations including 82 deletions and 19 inversions in strains descended from *E. coli* strains REL606 or REL607 (Raeside et al. 2014). In nature, such rearrangements can disturb the set genome order of ancestral genes and cause significant levels of gene loss and the creation of pseudogenes (gene inactivation). HT of genes is thought to be balanced by gene loss mediated by recombination between IS elements, so bacterial genome size will not increase continually (Parkhill et al. 2001). An example of this can be seen in *Bordetella* and *Yersinia pestis* where strains with smaller genomes were found to have more ISs than those with larger genomes (Parkhill et al. 2001).

1.2.4. Genomic Islands

Genomic islands (GIs) are distinct chromosomal elements 9 to 200 kb in length, which are often flanked by short repeats, IS elements or transfer RNA (tRNA) genes. GIs are often found inserted alongside tRNA genes, and typically have a GC content that differs from the core genomic regions (Kaper et al. 2004, Ochman et al. 2000). The higher integration activity is due to the presence of genes that encode transposases and integrases, enzymes which catalyse the insertion of mobile genetic elements into the recipient strain chromosome (Dobrindt et al. 2004). GIs were originally identified as clusters of genes encoding VAFs, referred to as “pathogenicity islands” (PIs; Hacker et al. 1997, Buchrieser et al. 1998, Kaper and Hacker 1999), but the availability of complete genome sequences indicated the presence of chromosomal insertions which had no obvious link to pathogenicity (Perna et al. 2001), hence the adoption of the more general term GIs.

A PI may confer upon the host strain a specific virulence phenotype due to the nature of the encoded proteins. One medically important and well-characterised *E. coli* PI is the Locus for Enterocyte Effacement (LEE). The LEE is a principal virulence system present in the genomes of enteropathogenic *E. coli* (EPEC) and enterohaemorrhagic *E. coli* (EHEC) (Nataro and Kaper, 1998). It encodes a Type III secretion system which acts as a “molecular syringe”, to secrete virulence-associated factors from the bacterial cell to the host across both the bacterial and host cell membranes. The LEE can be found on plasmids or within the chromosome, inserted alongside the *pheU*, *selC*, or *pheV* tRNA genes (Deng et al. 2001, Jores et al. 2004). It is 35.6 kb in length with a G+C content of 38.4%, far lower than the *E. coli* genome average of 50.8% (Frankel et al. 1998).

1.2.5. Phylogenetics and the evolutionary tree

Substitutions and indels of chromosomal genes and MGEs are an important source of information used for deducing the genetic and evolutionary relationships between a set of bacterial strains (Duchêne et al. 2016, Woese 2000). Since Charles Darwin drew the first diagrammatic tree of evolutionary relationships, branching depictions of genetic kinship have been described in the same way in the form of phylogenetic trees (phylogenies) (Woese 2000). Today molecular phylogenies provide a highly informative estimation of molecular genetic relatedness between bacterial isolates or genomes which can be used to address research questions relating to pathogenicity and evolution (Woese 2000). Modern molecular phylogenetics uses the distribution of substitutions between aligned nucleotide sequences to determine their respective evolutionary distances to one another, and to infer the evolutionary relationships between them (Chatzou et al. 2016). Phylogenetic trees include external (or leaf) nodes representing the sequenced organism, internal nodes representing inferred ancestral sequences, branch lines representing evolutionary distance that connects nodes to their ancestors, and branch support values which are a measure of the statistical support for the grouping of the descendent leaf nodes based on the input DNA sequences (Ahrenfeldt et al. 2017) (Figure 1.1). The branching pattern of the tree is referred to as its "topology".

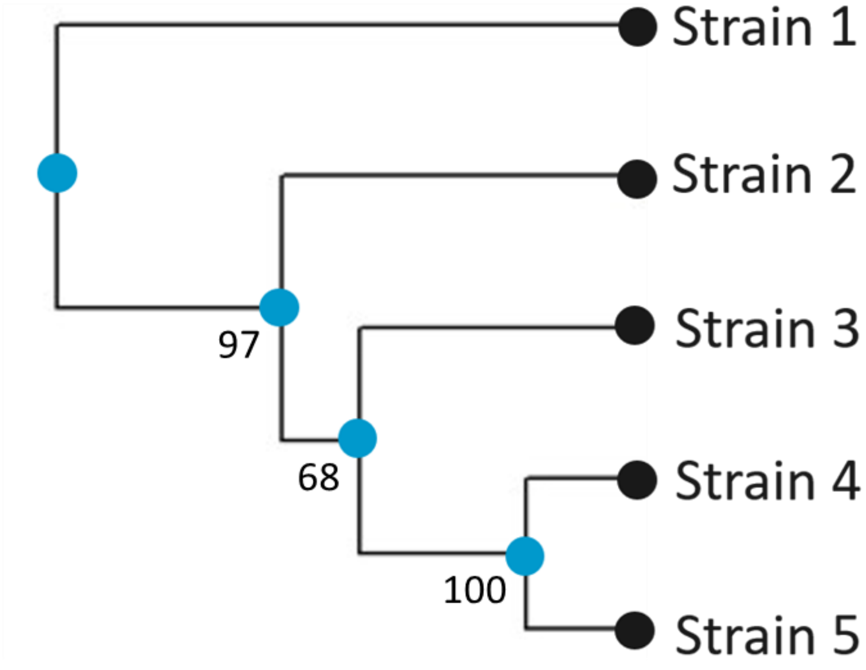


Figure 1.1. Diagram of a phylogenetic tree for five bacterial strains. Leaf nodes are shown as black circles and inferred ancestral (internal) nodes are shown as blue circles. The lines connecting nodes are branches. The numbers adjacent to the internal branches are example branch support values, which are measures of statistical support for the grouping of the descendent nodes under each internal node, expressed as a percentage.

1.2.6. Phylogeny and recombination

The possibility of recombination via HT should be considered when evolutionary relationships of bacterial strains are being inferred using a phylogenetic tree (Escobar-paramo et al. 2003, Didelot et al. 2010). A phylogenetic tree constructed using DNA sequences from genes which have undergone recombination may exhibit a different topology to that constructed using DNA from genes which have not undergone recombination (Didelot et al. 2010). This is because a recombination event between two isolates means that the recipient appears to cluster more closely with the donor strain in the phylogeny obtained from the recombinant gene sequences. However, the two strains in fact have a different phylogenetic relationship if inferred using DNA sequences not involved in recombination, which reflects the “true” evolutionary relationships between the isolates (Figure 1.2) (Didelot and Maiden 2010).

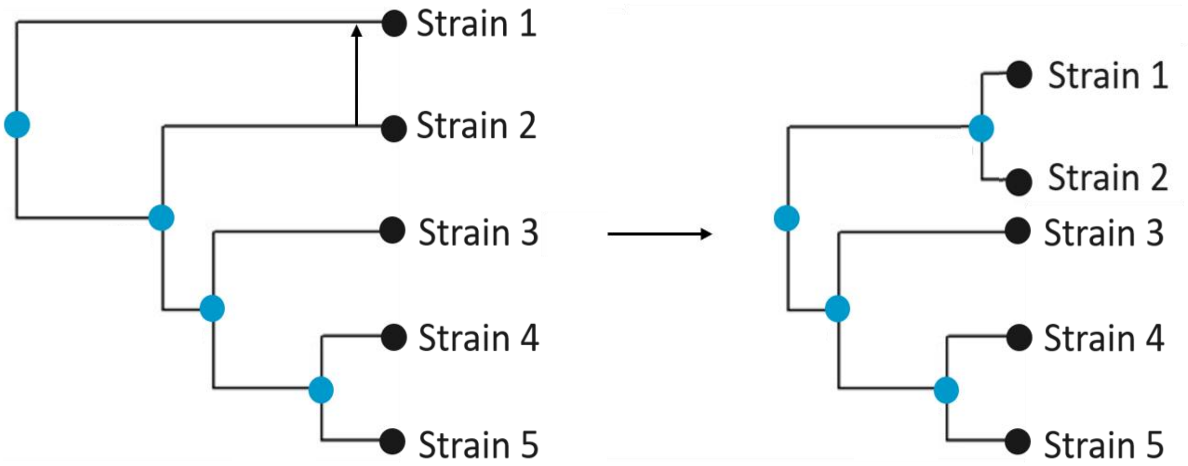


Figure 1.2. Diagram illustrating the phylogenetic effects of horizontal transfer (HT) via recombination. Left: the phylogeny created using gene sequence without a history of recombination with an arrow showing the point at which a previously present gene is acquired by strain 1 from strain 2 in a homologous recombination event. Right: A phylogeny constructed from the gene affected by the recombination event, showing that strains 1 and 2 are clustered more closely than in the first phylogeny but have diverged independently since the recombination event.

Recombination has implications for the determination of a bacterial species phylogeny (Didelot and Maiden 2010). Bacteria clonally reproduce through binary fission so a bacterial species phylogeny depicts inferred binary fission events as internal nodes in the phylogeny (Didelot and Maiden 2010). The species phylogeny is referred to as the 'clonal phylogeny' (Wieler et al. 1997), which is the term used throughout this thesis. However, to construct a clonal phylogeny recombinant core genome sequence must be removed prior to phylogeny construction (Didelot et al. 2010).

1.2.7. Whole genome sequencing of bacteria

Whole genome sequencing of *E. coli* is an important way to study the genetic diversity, evolutionary history, and virulence of a set of isolates. The first complete bacterial genome (*Haemophilus influenzae*) was sequenced in 1995 (Fleischmann et al. 1995) using Sanger sequencing (Sanger et al. 1977). At that time genome sequencing was not scalable due to the costs associated with the Sanger method and the coverage produced in sequenced reads was also limited and not consistently detailed (Woolley and Mathies 1995, Ruan et al. 1995, Loman et al. 2012). In 2005 the first high-throughput sequencing technologies (HTS; also referred to as 'next-generation' sequencing, NGS) were developed which produced much longer read lengths, removing the requirement for a reference genome. Initially the most widely used HTS platform for bacterial sequencing was the Roche 454 (Margulies et al. 2005) (read length 700-800 bp). However, the technology was found to have a high rate of indel errors (Loman et al. 2012) meaning it was superseded by Illumina sequencing (read lengths 200-300 bp). Illumina sequencing is the currently most used technology and involves fragmenting DNA, attachment of adaptor DNA, cluster generation in a unique cluster growth and bridge amplification step, and a single-base at a time imaging method of sequencing and included analysis of basecalls and read data as part of each run. The technology overall provides greater accuracy at reporting indel regions than the Roche 454 sequencer (Loman et al. 2012, Quainoo et al. 2017). Additionally, two alternative technologies have most recently been developed for use in high-throughput sequencing; the PacBio platform (Rhoads and Au 2015) and the pocket-sized MinION sequencer (Jain et al. 2016). The PacBio platform provides a single-molecule, real-time de-novo sequencing approach and was designed to produce longer read lengths than existing technologies (>10 kb) which allows it to produce single-contig bacterial genomes (Rhoads and Au 2015). However, the PacBIO exhibits a comparatively high error rate in repeat sequence regions so is currently most suitable for metagenomic studies where high-throughput sequencing is required and where errors in repeat sequences can still allow identification of a sequence to the species level (Loman et al. 2012). The MinION is a portable sequencer weighing less than 100 g which has allowed new remote research possibilities and can produce reads of lengths 100-300 kb (Mikheyev et al. 2014). However, it currently exhibits a significant error rate in sequence deletion regions (Mikheyev et al. 2014) so it and the PacBIO technologies require development. However, the speed at which sequencing technologies are being developed (Quainoo et al. 2017) indicate that the pace of bacterial whole genome sequencing is likely to continue and increase in the future as costs decrease and higher quality sequences are produced more efficiently than previously (Loman et al. 2012).

1.2. *E. coli* genetic diversity and evolutionary history

This section provides a discussion of previous research efforts to determine the genetic diversity present across the *E. coli* species and the major evolutionary events which have characterised its evolutionary history.

1.3.1. *E. coli* genome structure and diversity

Studies of *E. coli* strain genomes indicate that the *E. coli* species genome varies considerably in size, ranging from 4.6 Mb to 5.5 Mb (Bergthorsson et al. 1995, Bergthorsson and Ochman 1998). The first *E. coli* genome sequence to be published was of strain K-12 MG1655 (Blattner et al. 1997), followed by strains O157: H7 Sakai (Hayashi et al. 2001) and O157:H7 EDL933 (Perna et al. 2001) around the same time and then strain CFT073 (Welch et al. 2002). Comparisons of these genomes to one another revealed that the *E. coli* genome exhibits a shared co-linear backbone which is 'punctuated' by hundreds of genomic islands many of which are shared between strains K-12 MG1655 and O157:H7 EDL933 (Perna et al. 2001) (Figure 1.3) and CFT073 (Welch et al. 2002).

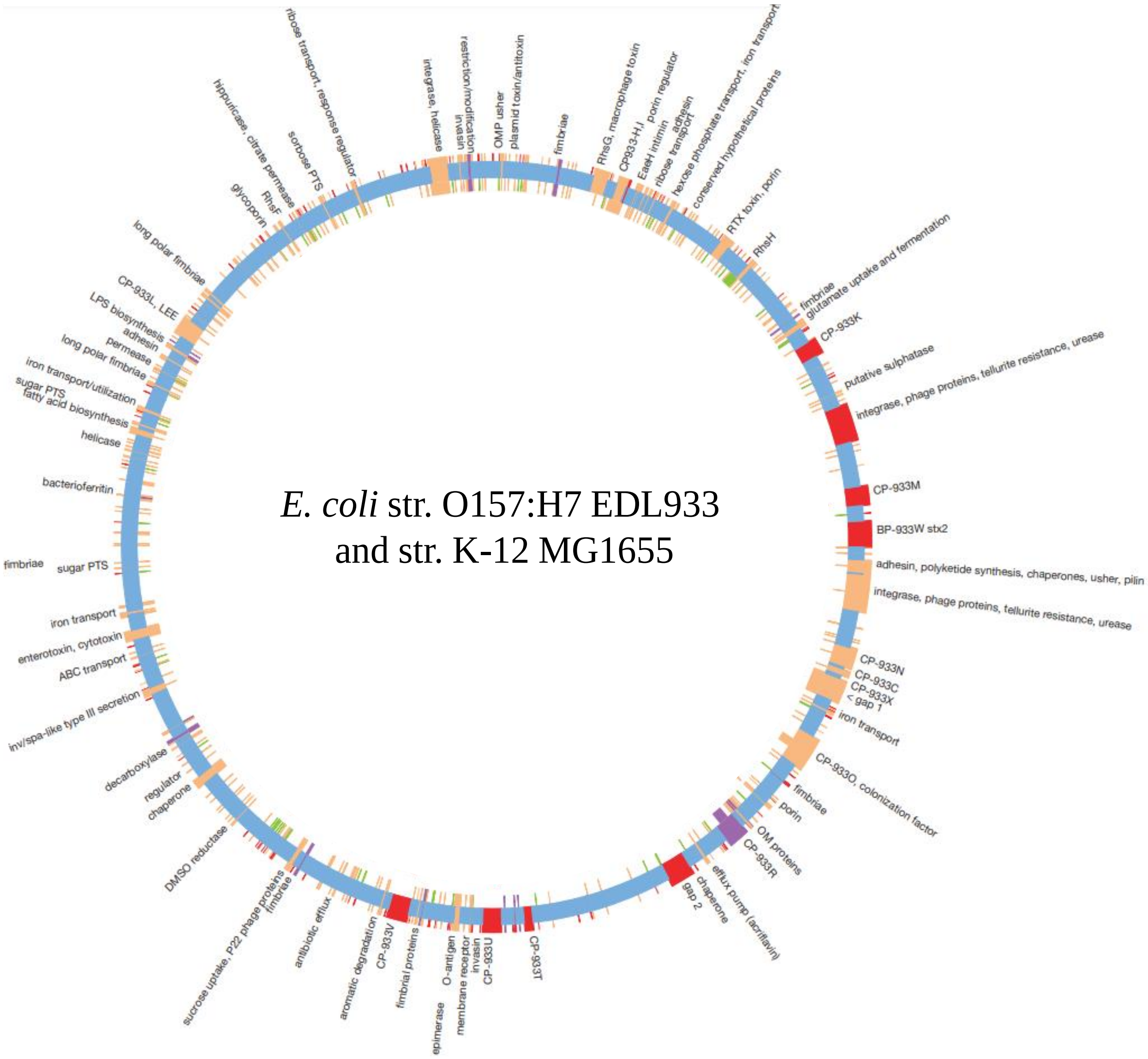


Figure 1.3. Circular diagram of the O157:H7 EDL933 genome (outer circle) compared to that of the str. K-12 MG1655 genome (inner circle) with their shared co-linear core genome backbone (blue middle circle). Positions of sequences comprising individual genes and genomic islands specific to EDL933 are in red and those specific to K-12 MG1655 are in green. Genes and genomic islands which are shared by both genomes are in tan and purple and those in the same position in both as in EDL933 are detailed in the core genome backbone. Functional annotations for proteins encoded by example genes and islands are detailed next to the outer ring. Edited from Figure 1 of Perna et al. (2001) (Reprinted and adapted by permission from Springer Nature [COPYRIGHT] 2001).

The term pan genome was coined by Tettelin et al. (2005) to describe all genes that are present in the chromosomes of a set of strains. The term can be used for a given set of strains or for all strains of a bacterial species (the species pan genome). The pan genome can be divided into the dispensable genome (Medini et al. 2005), also referred to as the variable (Lukjancenko et al. 2010), or accessory (Touchon et al. 2009) genome, and the core genome. The dispensable genome comprises of genes which are present in the species, but which are absent from the genomes of one or more strains. The core genome comprises of genes which are present as orthologues (typically defined as genes which have ≥ 80%, ≥ 85%, ≥ 90%, or ≥ 95% amino acid identity to one another, depending on the analysis), present in 100% of genomes (strict definition) or ≥ 99% of genomes (permissive definition to account for core genes which have been deleted in 1% of strains, the definition used in this thesis) of a genome set (core genes to the set) or species (core genes to the species) (Touchon et al. 2009, Lukjancenko et al. 2010). A pan and core genome plot created by Lukjancenko et al. (2010) showed that the change in pan genome, illustrated by the cumulative number of gene families identified across genomes, rises with the addition of each *E. coli* genome (Figure 1.4). In the plot the core genome is illustrated as the conserved number of gene families identified across genomes which decrease with each genome.

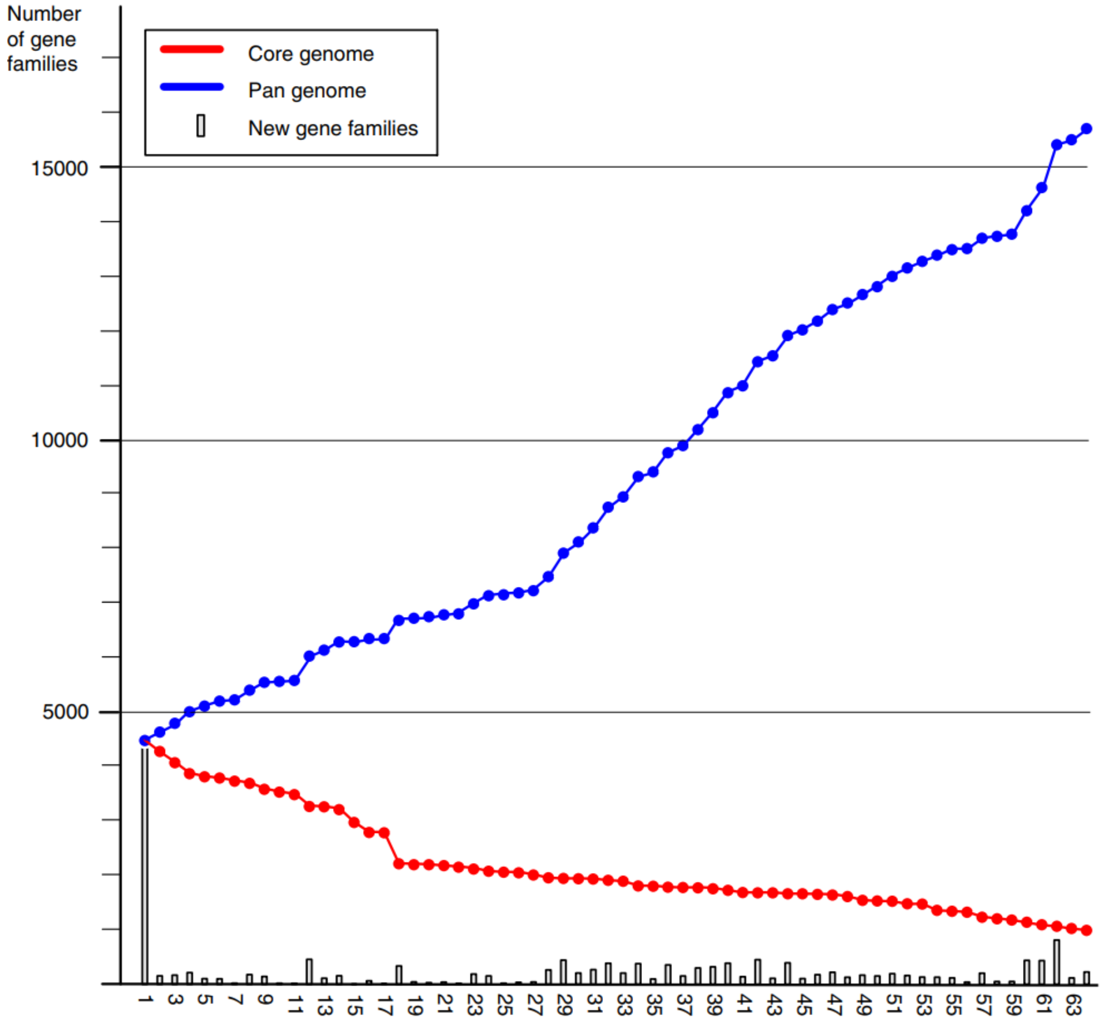


Figure 1.4. Pan and core genome plot showing that the change in pan genome, illustrated by the cumulative number of gene families identified, rises with the addition of each genome. The core genome is illustrated as the conserved number of gene families identified which decrease with each additional genome. The number of gene families identified with each new genome which are previously unseen are also shown as bars. Edited from Figure 4 of Lukjancenko and Wassenaar (2010), Creative Commons use: http://creativecommons .org/licenses/by /2.0/.

Using complete genome sequences, previous studies have reported the *E. coli* pan genome and accessory genome to effectively be infinite and encode a broad range of functions (Touchon et al. 2009). Conversely, core genes are likely to be responsible for functions relating to basic metabolism, replication, translation, and transcription (Touchon et al.

2009, Rasko et al. 2008, Chaudhuri et al. 2010). Genome content differences between strains occur primarily through gene duplication and subsequent divergence of one copy (Teichman and Babu 2004) or through acquisition via HT from other lineages (Dobrindt et al. 2004). Genome content differences provide evidence for the accessory genome, which allows for specific adaptation to the range of changing habitats where *E. coli* are found (Monk and Bosi 2018, Tenaillon et al. 2010, Dobrindt et al. 2004).

1.3.2. The *E. coli* species phylogeny

Efforts to determine the *E. coli* species phylogeny (or clonal phylogeny) date back over 40 years (Milkman 1973). A species phylogeny is useful as a point of reference for genetic diversity and population genetics studies. *E. coli* phylogenetics started with the measurement of electrophoretic mobility of the same set of enzymes from different strains, and quantifying diversity based on the observed variation in electrophoretic profiles (Milkman 1973). A given number of selected cellular soluble metabolic proteins are first extracted from isolate cells and then placed into starch gel which has a current passed through it (Selander et al. 1986). Gels are prepared by using 48g starch to 420 ml of gel buffer (no. 2901-02; Connaught Laboratories as used by Selander et al. (1986)) and a constant voltage of 100-350 V is passed through the gel for the duration of the experiment (Selander et al. 1986). Optimum electrophoretic conditions for each enzyme of a given bacterial species can be determined prior to analysis through trialling a range of buffers with differing PH values (Selander et al. 1973, Selander et al. 1986). Gels are then incubated at 37 degrees centigrade in the dark for a period to 10 minutes to four hours until the point where electrophoresed and stained enzymes appear on the gel as defined narrow bands (Tenover et al. 1994, Maslow et al. 1994). Variations in electrophoretic mobility of the enzymes are then interpreted as reflections of amino acid substitutions in the enzymes which alter protein charge and motility through the gel under influence of a current

(Tenover et al. 1994, Maslow et al. 1994). Phylogenetic relationships among strains are inferred based on measured distances of enzymes in the gel from the point of origin and from one another and are assigned numbers to aid with this process (Selander et al. 1973, Selander et al. 1986).

The most noted weaknesses of MLEE was that carrying it out was regarded as technically demanding in terms of time and was not always accessible due to the expense of the resources required to carry it out (Maslow et al. 1993, Tenover et al. 1994). It was also reported to not achieve adequate levels of discrimination between loosely related isolates in epidemiological studies (Maslow et al. 1993, Tenover et al. 1994). For example, when used to analyse pyelonephritis patients infected with UPEC, virulent strains were represented in MLEE by a limited number of closely related lineages which were difficult to distinguish (Maslow et al. 1993). However, the approach did provide sufficient utility at the time of its development for distinguishing bacterial strains and isolates (Ochman and Selaner 1984). It became the standard approach for quantifying *E. coli* diversity, and a series of similar *E. coli* MLEE studies over the following years increasingly supported the hypothesis that the species evolved clonally with little recombination (Ochman and Selaner 1984, Whittam *et al.* 1983). The MLEE findings led Ochman and Selander (1984) to create the 72-strain *E. coli* reference (ECOR) collection, which consists of strains isolated from 17 mammals including humans and chosen to represent the electrophoretic enzyme diversity seen across all *E. coli*. Midpoint rooting (selection of the longest branch in the phylogeny to be the out group root) of a phylogeny constructed using 38 enzyme loci from these strains by Herzer et al. (1990) defined the four groups A, B1, B2, and D with several unclassifiable strains as group E (Figure 1.5). A group named C had also previously been defined in earlier analyses, though this study revealed it is not a well-defined phylogenetic group and so it is no longer recognised (Chaudhuri and Henderson 2012).

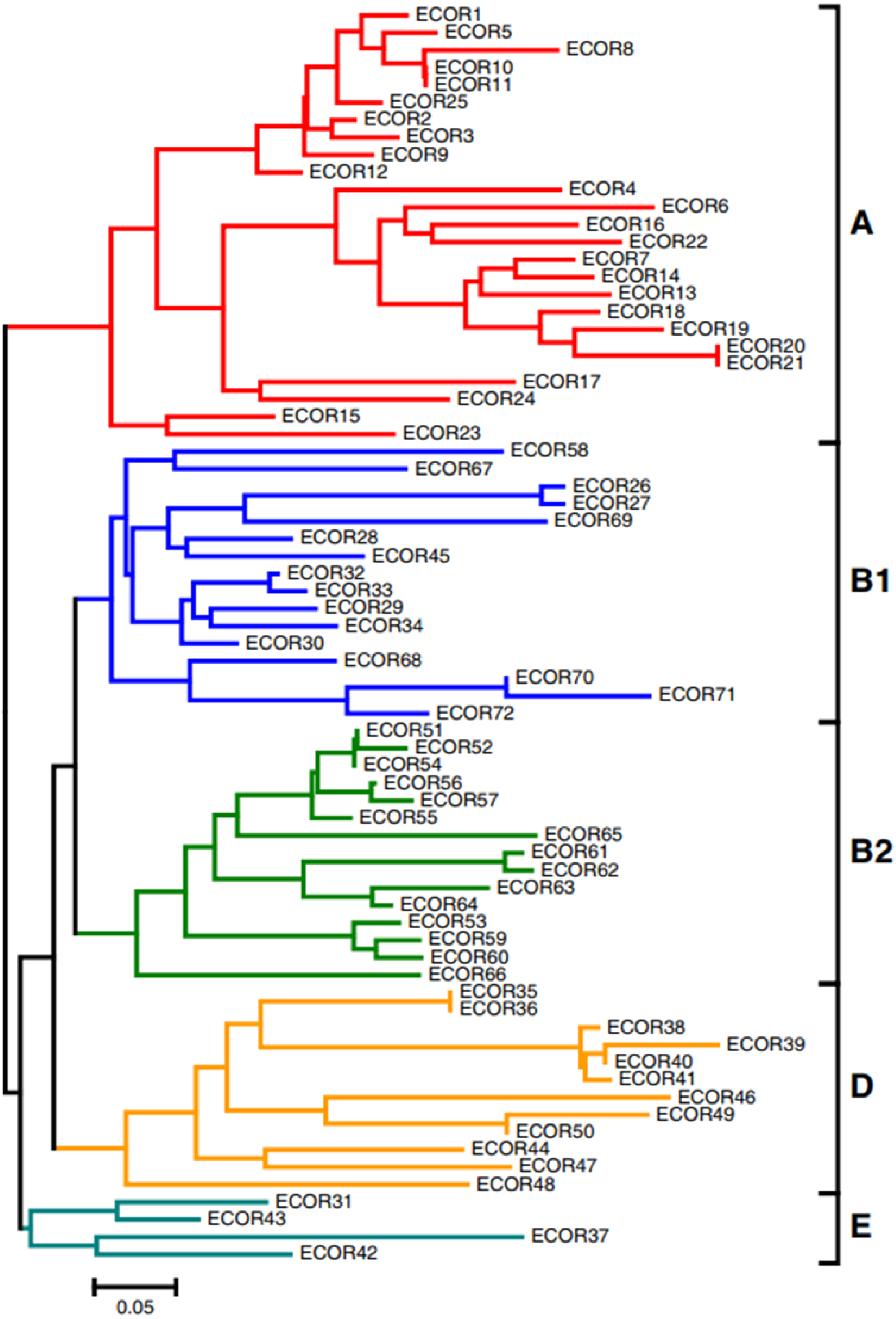


Figure 1.5. MLEE phylogeny of 38 enzyme loci using 72 strains of the ECOR collection constructed by Chaudhuri and Henderson (2012) using the neighbor-joining method (Saitou and Nei, 1987), equivalent to Figure 1 in Herzer *et al.* (1990) (used with permission from Elsevier). The scale bar on the bottom left indicates the number of substitutions per site represented by the branch length shown.

The major MLEE phylo-groups identified by Herzer et al. (1990) have since been supported by variation across nucleotide sequences of a 29.9 kb gene cluster in a genomic location separate from the genes encoding the enzymes used (Ren *et al*. 2004). Moreover, Clermont *et al.* (2000) developed a diagnostic method, a triplex PCR, for efficiently characterizing strains into one of the four groups with an 85-90% accuracy rate using variation in the sequenced genes *chuA*, *yjA*, or sequence TSPE4.C to differentiate the 4 groups. This was updated in 2013, with an updated quadruplex PCR method to include previously unincluded groups denoted E and D2/F and divergent clades C-I, C-III, C-IV and C-V using the gene *arpA* additionally in PCR (Clermont *et al.* 2013).

MLEE data was surpassed in accuracy by the availability of nucleotide and amino acid sequence data, which have a markedly reduced probability of convergence, where distantly related strains by chance show a similar phenotype and cluster together, which falsely indicates that the strains are more closely related than they are (Bisercic *et al.* 1991). Phylogenetic analysis of sequence data facilitated studies to investigate genetic diversity. Initially this was done on individual genes, but this led to the development of multi-locus sequence typing (MLST), which involves phylogenetic analysis using multiple genes to phylogenetically place strains usually with a number called a ‘sequence type’. The approach is carried out by first designing oligodeoxyribonucleotide (oligo) primer sequences which are constructed to be complementary to the desired genes which will be used in the MLST analysis (Escobar-Paramo et al. 2004). Oligos are then subjected to the polymerase chain reaction and are used to generate DNA fragments which have overlapping ends, which then combine in a fusion reaction where the overlapping ends anneal. This causes the 3’overlap of each strand which then serves as a primer for the 3’ component of the complementary strand (Ho et al. 1989). Further amplification of the fusion product then occurs to produce a final amplicon product which is then DNA sequenced (Ho et al. 1989, Maiden et al. 2013). Computational phylogenetic analysis of these sequences then follows to determine the genetic relationships of the sequences and assign the sequence type based on the isolate or strain position in the phylogeny topology (Maiden et al. 2013). The development of MLST meant that isolates could be phylogenetically typed using a greater quantity of information compared to MLEE by using base polymorphisms compared to electrophoretic profiles, which yielded greater typing accuracy (Maiden et al. 2013). The comparatively superior phylogenetic typing accuracy that the approach provided meant it became widely adopted (Maiden et al. 2013). Four different *E. coli* MLST schemas currently exist, all of which use 'housekeeping' genes that are inferred to be involved in metabolic functions (Maiden et al. 2013) (Table 1.2): Achtman (7 genes), which is the most commonly used (Clermont et al. 2015, Wirth et al. 2006), Pasteur (8 genes) (http://www.pasteur.fr/recherche/enopole/PF8/mlst/EColi.html), and EcMLST, which consists of two schemas, one of 7 genes, and one of 15 (Qi et al. 2004).

Table 1.2. Four widely used MLST schemas used to phylogenetically place *E. coli*.

MLST schema	**Genes**	**Origin**	**Website**
Achtman	*adk, fumC, gyrB, icd, mdh, purA, recA*	Warwick Medical School	http://mlst.warwick.ac.uk/dbs/Ecoli
Pasteur	*dinB, icdA, pabB, polB, putP, trpA, trpB, uidA*	Pasteur Institute	http://www.pasteur.fr/recherche/enopole/PF8/mlst/EColi.html
EcMLST: 7 genes	*aspC, clpX, fadD, icdA, lysP, mdh, uidA*	Michigan State University	http://www.shigatox.net/ecmlst/cgi-bin/index
EcMLST: 15 genes	*aspC, clpX, fadD, icdA, lysP, mdh, uidA, mtlD, mutS, rpoS, grpE, dnaG, cyaA, arcA, aroE*	Michigan State University	http://www.shigatox.net/ecmlst/cgi-bin/index

In an MLST study, Escobar-Paramo *et al.* (2004) used the sequences of 6 genes from the Pasteur schema (*trpA*, *trpB*, *pabB*, *putP*, *icd*, and *polB*) from 98 isolates and produced a tree topology similar to the MLEE phylogenies, but with group B2 as the outgroup rather than group A (Figure 1.6). This topology is largely supported by an unrooted phylogeny generated by Turrientes et al (2014) when the Achtman schema was used with a set of 80 *E. coli* strains (Figure 1.7). However, in the Achtman-based phylogeny group A was split across two clades rather than clustering as a monophyletic group. Group B2 is connected to the other groups by the longest branch.

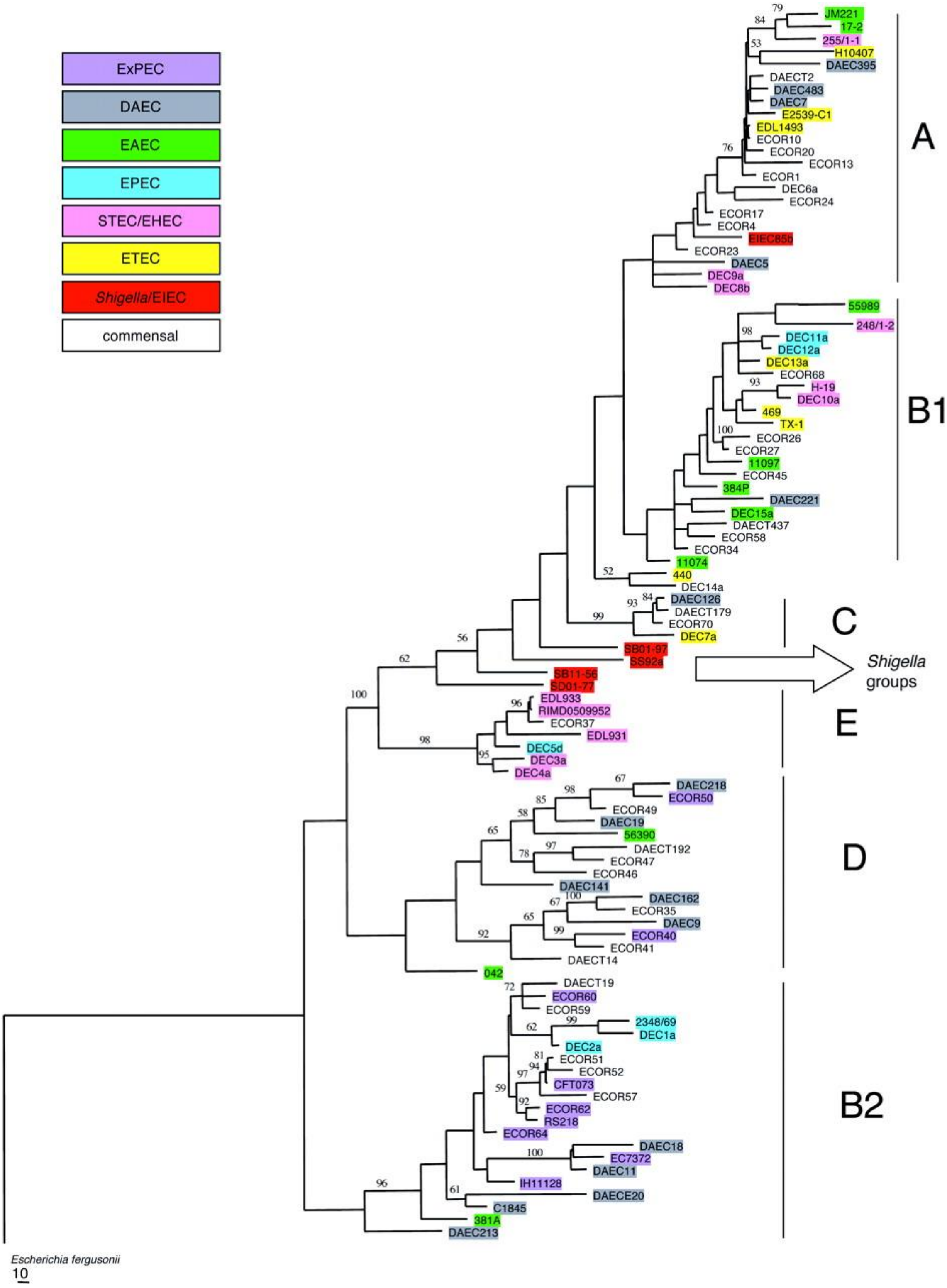


Figure 1.6. MLST phylogeny based on 6 chromosomal loci using 98 *E. coli* strains which includes strains from the ECOR collection. *E. coli* phylogenetic groups are labelled A-E (Figure 1 of Escobar-Paramo et al. 2004, used with permission from Oxford University Press). Percentage bootstrap support values are shown on internal branches. The scale bar on the bottom left indicates the number of substitutions per site represented by the branch length shown.

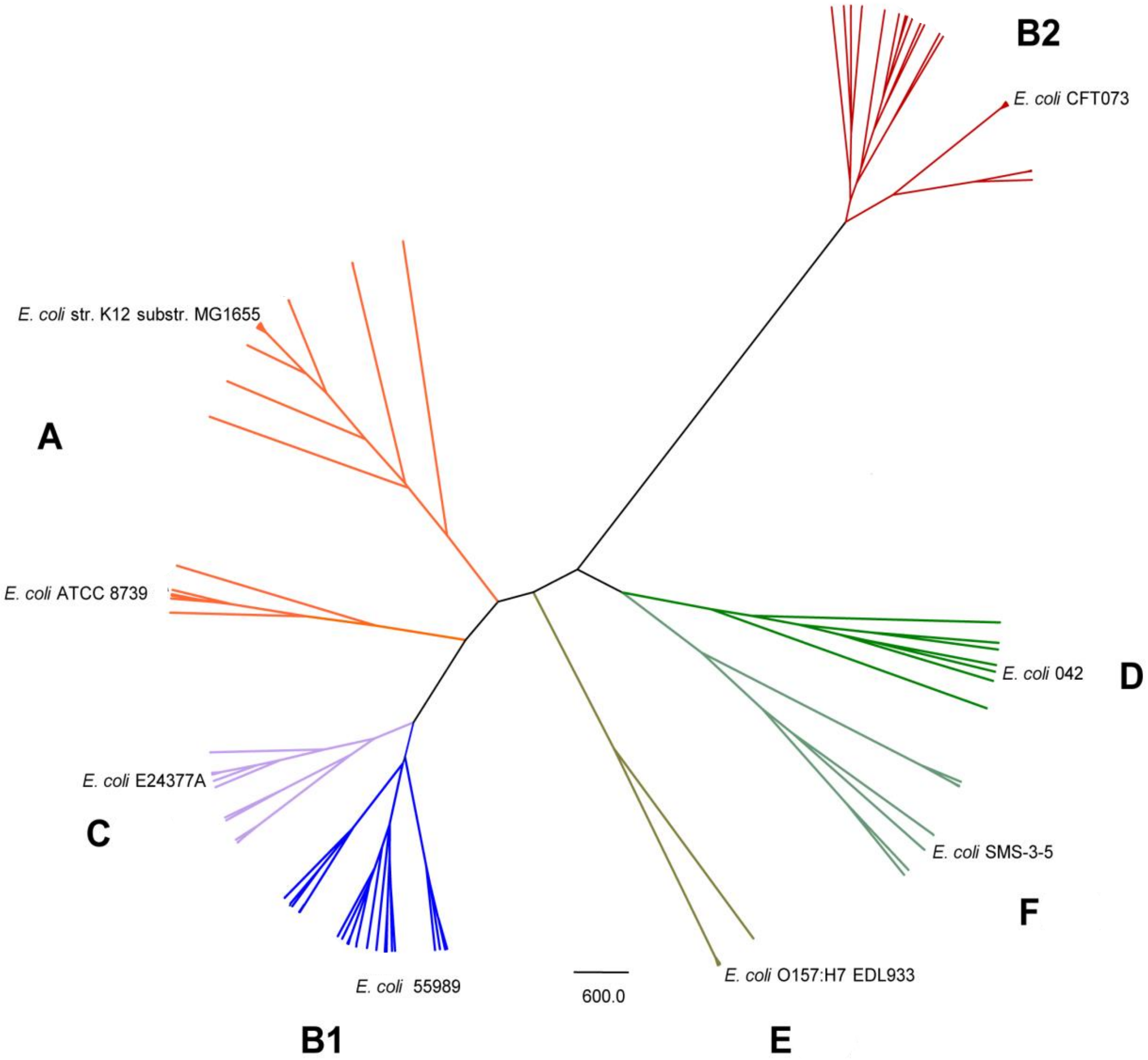


Figure 1.7. Unrooted phylogenetic tree constructed using genes from the Achtman schema from 80 *E. coli* strains. Phylogenetic groups are labelled A-E. Edited from Figure 2 of Turrientes et al. (2014), Creative Commons use: http://creative commons.org /licenses/by/4.0/. The scale bar indicates the number of substitutions per site represented by the branch length shown.

A weakness of MLST is that it is inherently limited for phylogenetic reconstruction (Been et al. 2015). This is because using a restricted selection of core or housekeeping genes to construct a clonal phylogeny can mean that informative sequences are excluded. Such informative sequence may support genetic divergence or close relationship between groups of isolates and excluding it can introduce inaccuracy through falsely indicating strains are more closely or distantly related than they are (Chaudhuri and Henderson 2012). There is also a chance that some of the included genes have previously been subject to recombination via HT, and since only a limited number of 7-15 genes are included in the analysis, any recombinant sequences may disproportionately affect the inferred phylogeny. This is because recombinant sequence will show a false indication of the isolates position in the phylogeny where its gene sequence will cluster more closely to gene sequence used in the phylogeny that resembles the recombinant sequence. For an isolate, the greater the ratio of recombinant to non-recombinant sequence, the greater the likelihood that the isolate exhibits a phylogenetic position which more closely resembles that of the donor of the recombinant sequence (Chaudhuri and Henderson 2012).

One way to circumvent these limitations is to create a core genome phylogeny (also referred to as a whole genome sequence (WGS) phylogeny) using all shared genes, which is more likely to be representative of the clonal species phylogeny than a phylogeny inferred using a subset of the core genes. This is because the quantity of recombinant sequence compared to non-recombinant sequence is likely to be lower when an increased number of genes are included. In a phylogeny constructed using these genes, the recombinant sequence therefore contributes significantly less to phylogenetic signal and topology than the recombination-free sequence (Touchon et al. 2009, Chaudhuri and Henderson 2012). A strength of the approach is the comparatively larger quantity of informative sequence included. It makes core genome phylogenetic typing a method that provides a superior accuracy than both MLST and MLEE approaches (Quainoo et al. 2017, Chaudhuri and Henderson 2012). This has resulted in a tendency demand for its increased use in genetic studies of *E. coli* and other bacteria (Nyolm et al. 2015, Ferdous et al. 2016, Quainoo et al. 2017). However, a weakness of the approach is that it requires access to sequencing technology capable of sequencing whole genomes and a computer with the processing power required to conduct phylogenetic analysis using whole core genome sequences

(Tenaillon et al. 2010, Quainoo et al. 2017). The time required to conduct phylogenetic analysis of core genomes can be prohibitively lengthy without a capable computer (Quainoo et al. 2017). However, sequencing technologies and capable computers are becoming increasingly available with the decreasing costs of equipment responsible for increased processing capabilities (Quainoo et al. 2017)

The first *E. coli* core genome phylogeny was published by Touchon et al. (2009) (Figure 1.8) using an alignment of 1,878 core genes. In this phylogeny the root strain was selected via midpoint rooting whereby the longest branch in the phylogeny is selected to be the root. The root split group D into two, with one half clustering with group B2 as an out group, and the other half clustering with groups A, B1, and E.

However, Kaas et al. (2012) constructed a phylogeny using 1,278 core genes (equivalent to ~1.28Mb) from 186 *E. coli* strains (Figure 1.9). Like the phylogeny by Touchon et al. (2009) the phylogeny defined a similar topology to the MLEE and MLST trees, although a division of group D (labelled as F and referred in later papers as D2/F) was reported as found by Touchon et al (2009) and with D2/F and B2 representing out groups to other A-E branches. The phylogeny agreed with the definition of seven major phylogenetic groups: A, B1, B2, D1, D2/F, E, and C-I (Walk *et al.* 2009, Clermont *et al.* 2000, 2013). It included group C-I, which clustered as an out group to the A-E strains (Figure 1.9). The group is from a larger clade comprising of phylogenetic groups C-I, C-III, C-IV, and C-V (Luo et al. 2011) (Figure 1.10). This clade of highly diverse ‘environmental *E. coli*’ are newly characterized and are currently not well studied but appear to exhibit a degree of genetic isolation from one another. Genetic isolation was inferred based on the relative greater number of synonymous substitutions per site observed for pairs cryptic clade strains compared to pairs of strains from groups A-E in an alignment of 1,910 core genes. The authors inferred the divergence had occurred as result of ecological separation functioning as a barrier that prevented the exchange of genetic material between strains of different groups (Luo et al. 2011).

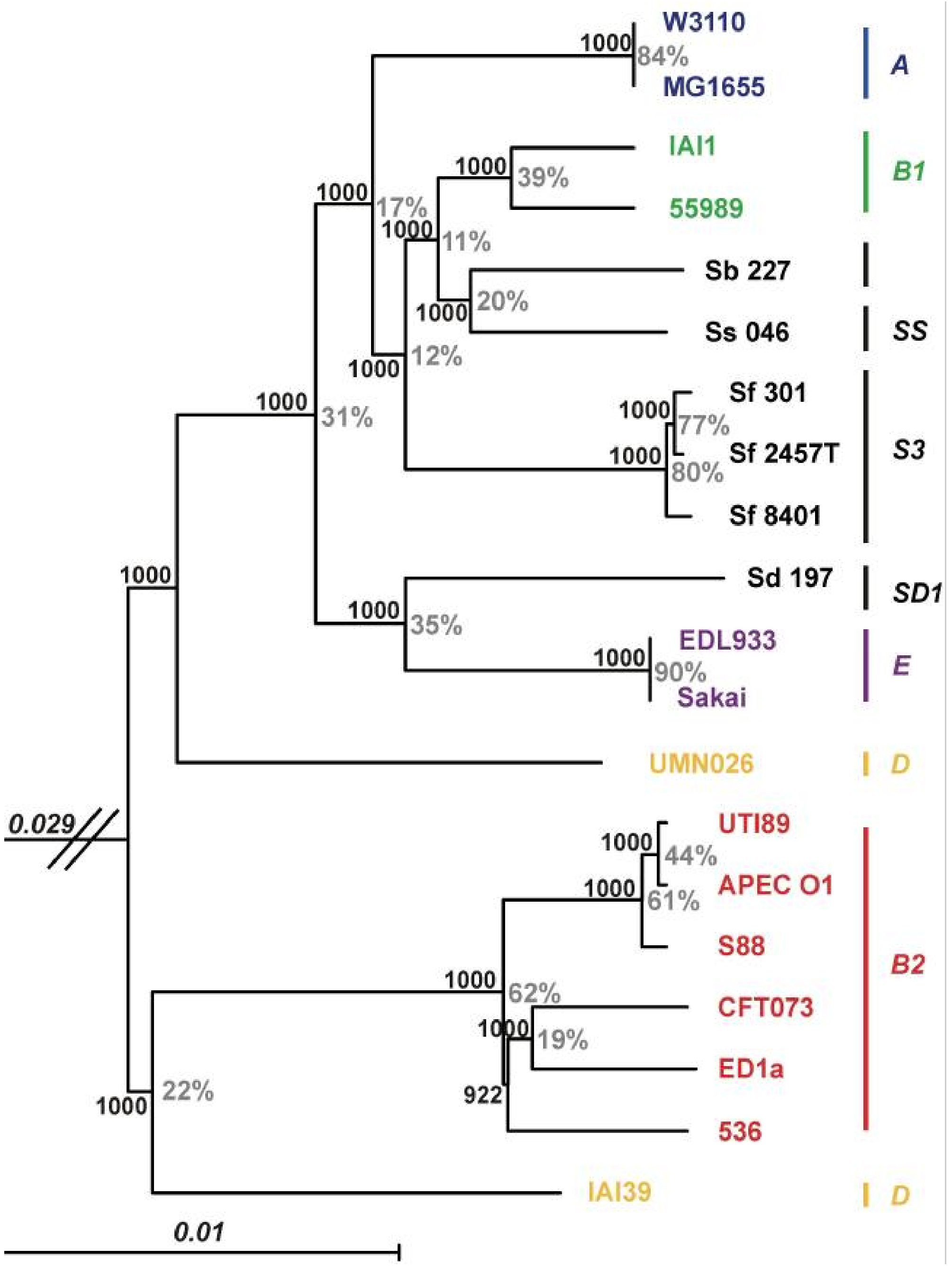


Figure 1.8. Rooted core genome phylogenetic tree of 14 *E. coli* and 6 *Shigella* strains as reconstructed from the sequences of 1,878 core genes. Percentage bootstrap support values are shown on internal branches. The scale bar on the bottom left indicates the number of substitutions per site represented by the branch length shown genes (Figure 4 from Touchon et al. 2009, used with permission from PLOS Genetics).

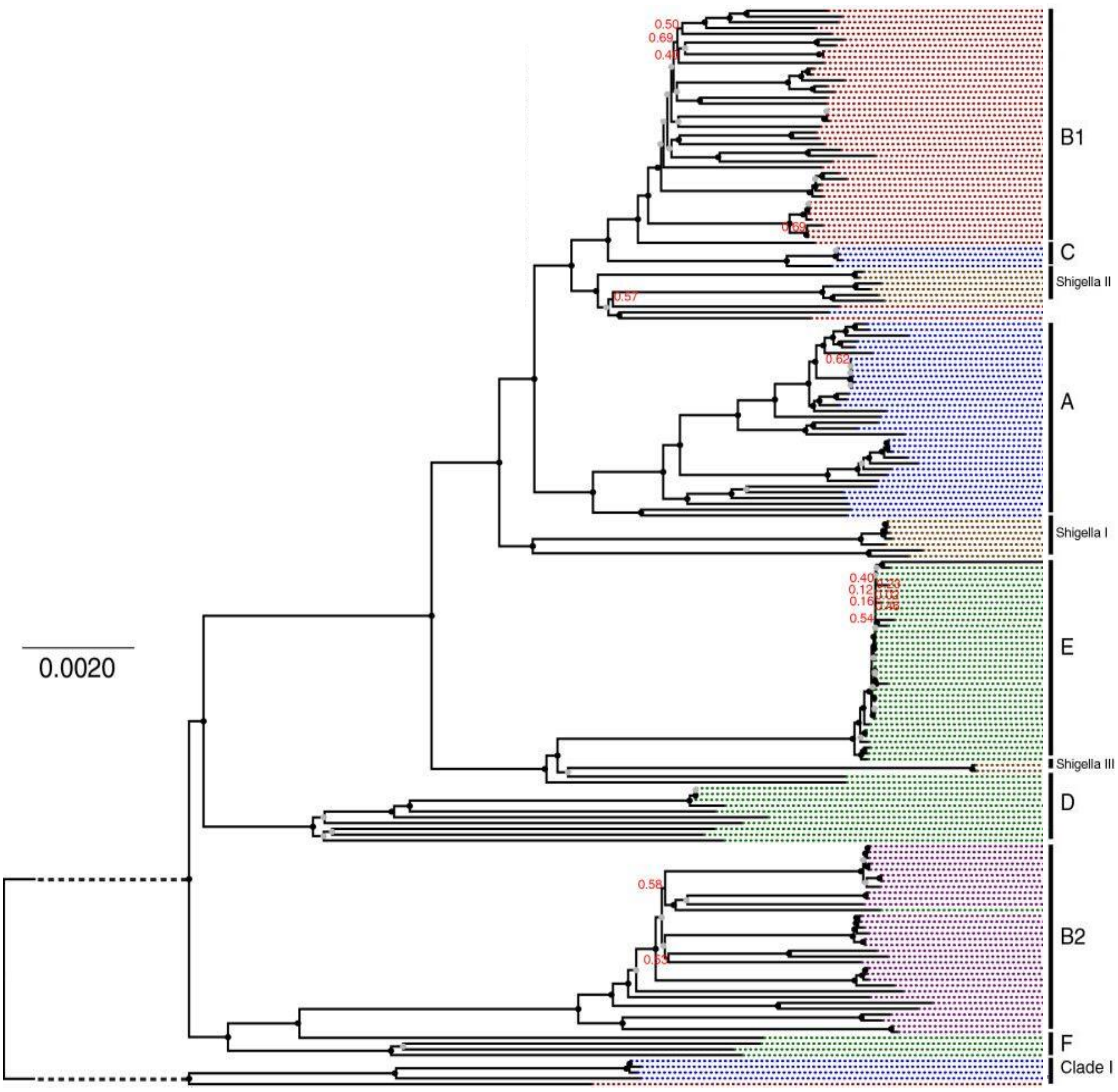


Figure 1.9. A midpoint rooted *E. coli* phylogeny constructed using 1,278 core genes (equivalent to ~1.28Mb) from 186 *E. coli* strains. Major phylogenetic groups are defined to the right. Percentage bootstrap support values are shown on internal branches. The scale bar indicates the number of substitutions per site represented by the branch length shown (Figure 6 of Kaas et al. 2012, Creative Commons use: http://creativecommons. org/licenses/by /2.0/).

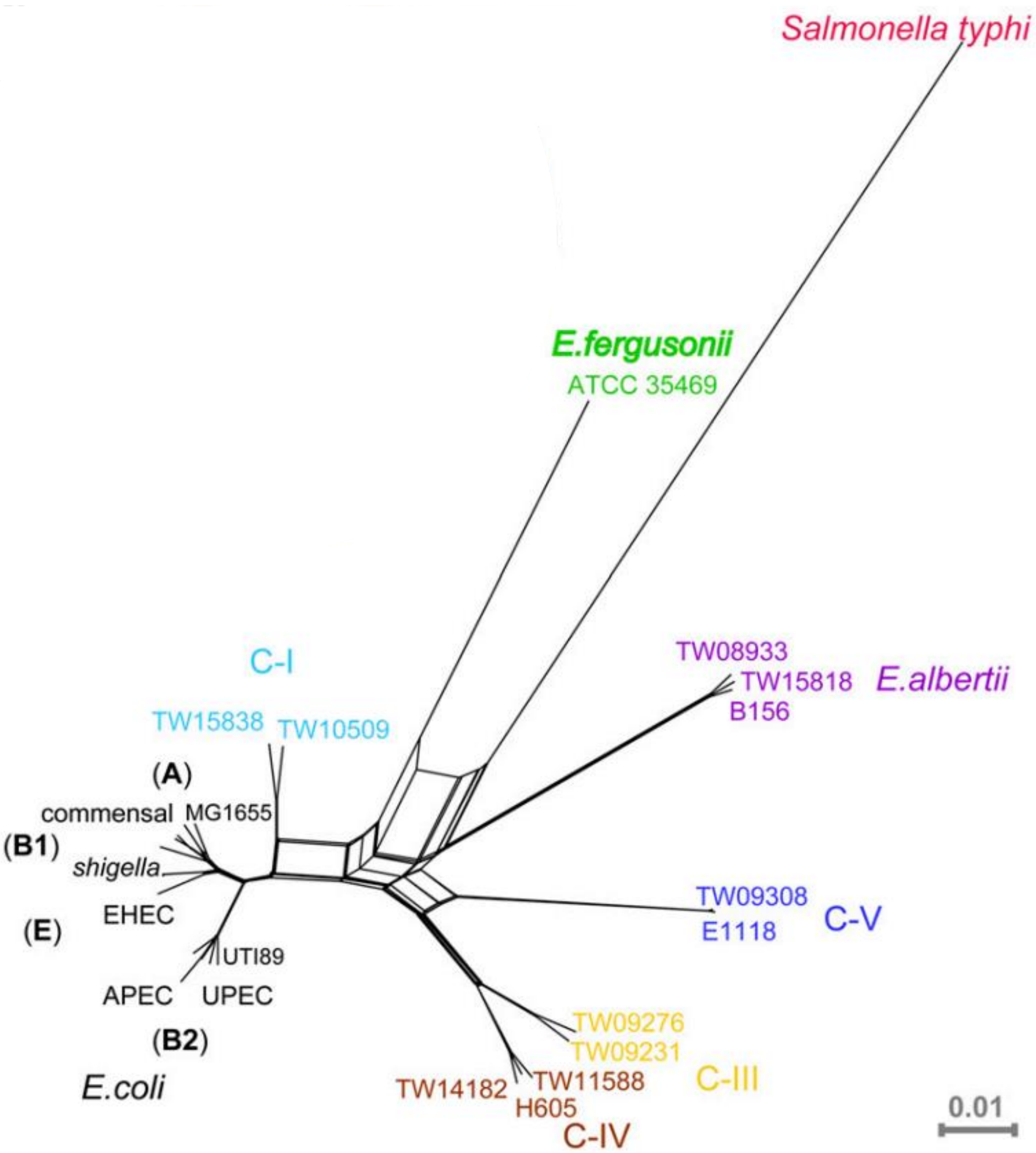


Figure 1.10. Whole-genome unrooted phylogenetic tree of 24 *Escherichia* genomes and *Salmonella typhi* using 1,910 core gene nucleotide sequences. It shows the divergent position of groups C I-V relative to A-E group *E. coli* strains. The scale bar indicates the number of substitutions per site represented by the branch length shown (Figure 2 of Luo *et al.* 2011, used with permission from PNAS).

Chaudhuri and Henderson (2012) also constructed a core genome phylogeny using a 2.78 Mb alignment (equivalent to ~2,780 genes) of 20 *E. coli* and 4 *Shigella* genomes (Figure

1.11). The phylogeny defined an identical topology of A-E groups and group C-I as found by Kaas et al. (2012) (Figure 1.9).

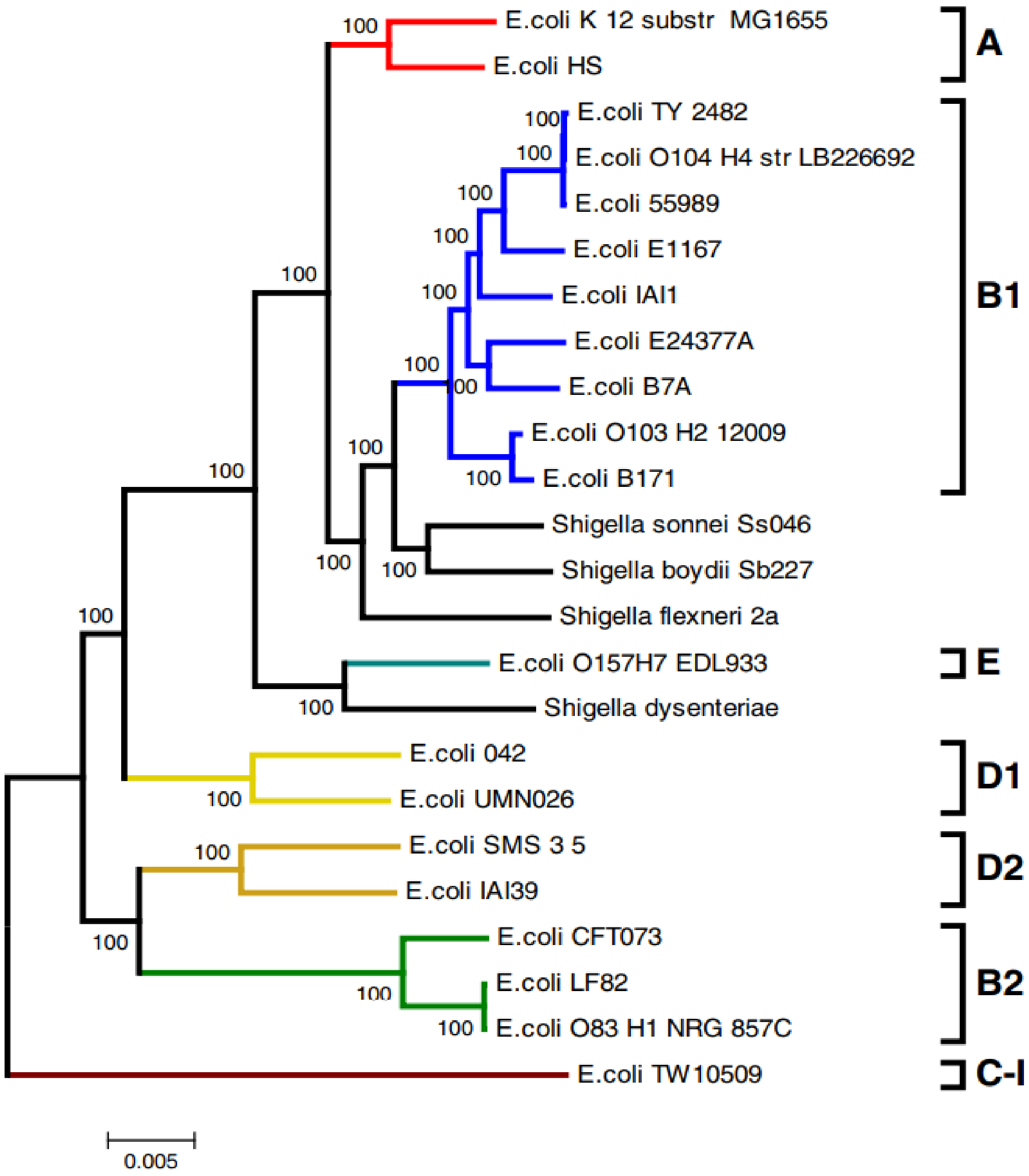


Figure 1.11. Whole-genome unrooted phylogenetic tree of 24 *Escherichia* genomes and *Salmonella typhi* of complete and draft whole-genome sequences. The major phylogenetic groups are defined A-E. Percentage bootstrap support values are shown on internal branches. The scale bar indicates the number of substitutions per site represented by the branch length shown (Figure 4 from Chaudhuri and Henderson 2012, used with permission from Elsevier).

This same topology was also reported in a recent core genome phylogeny by Dunne et al. (2017) of 29 *E. coli* and 4 *Shigella* genomes constructed using 2,173 core genes (Figure 1.12), and in a phylogeny by McNally et al. (2013) (Figure 1.13) based on a 2.3 Mb alignment of 62 *E. coli* genomes.

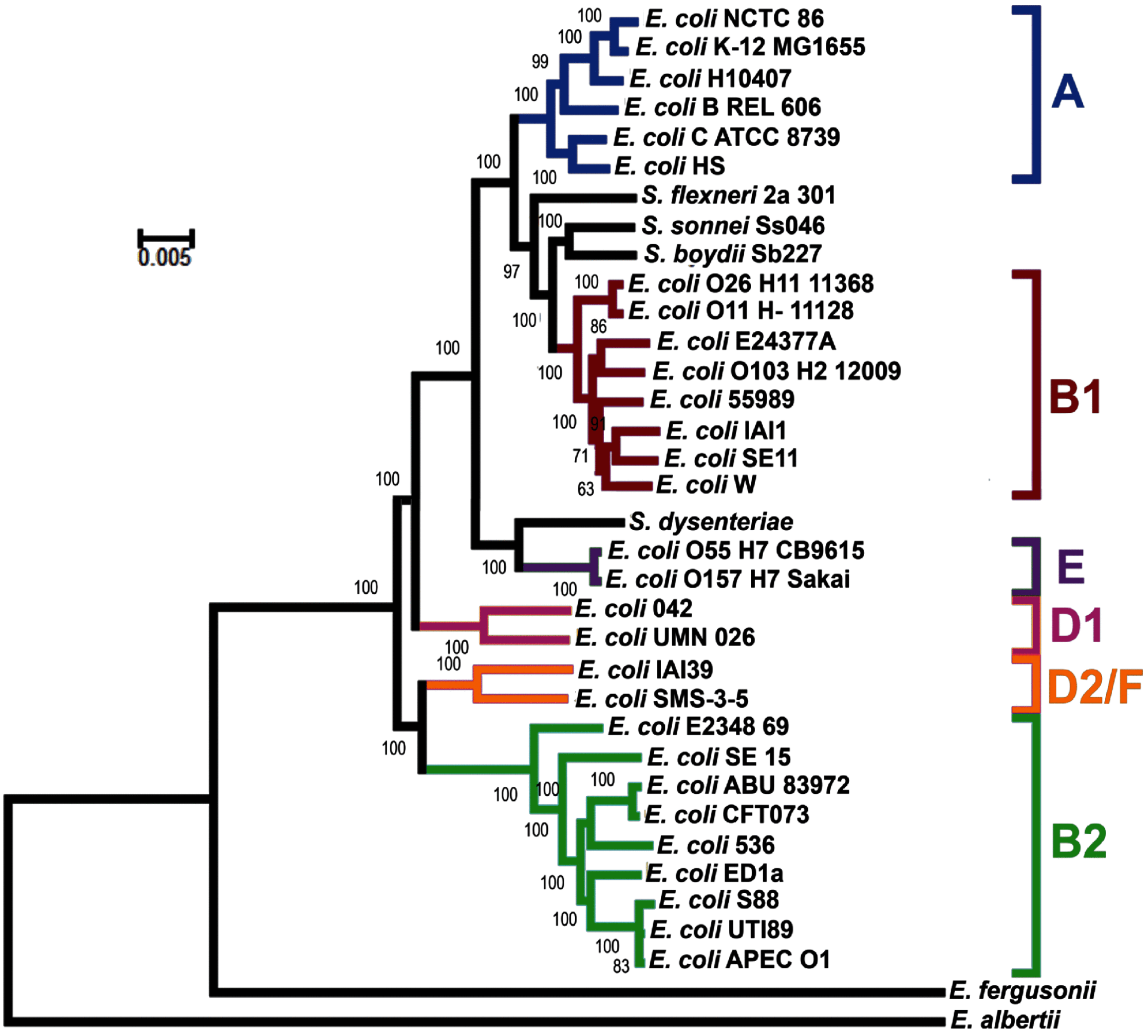


Figure 1.12. A rooted phylogeny of 29 *E. coli* and 4 *Shigella* genomes constructed using 2,173 core genes (equivalent to ~2.17Mb). Major phylogenetic groups are labelled to the right. Percentage bootstrap support values are shown on internal branches. The scale bar indicates the number of substitutions per site represented by the branch length shown (Figure 3 from Dunne et al. 2017, used with permission from Microbial Genomics).

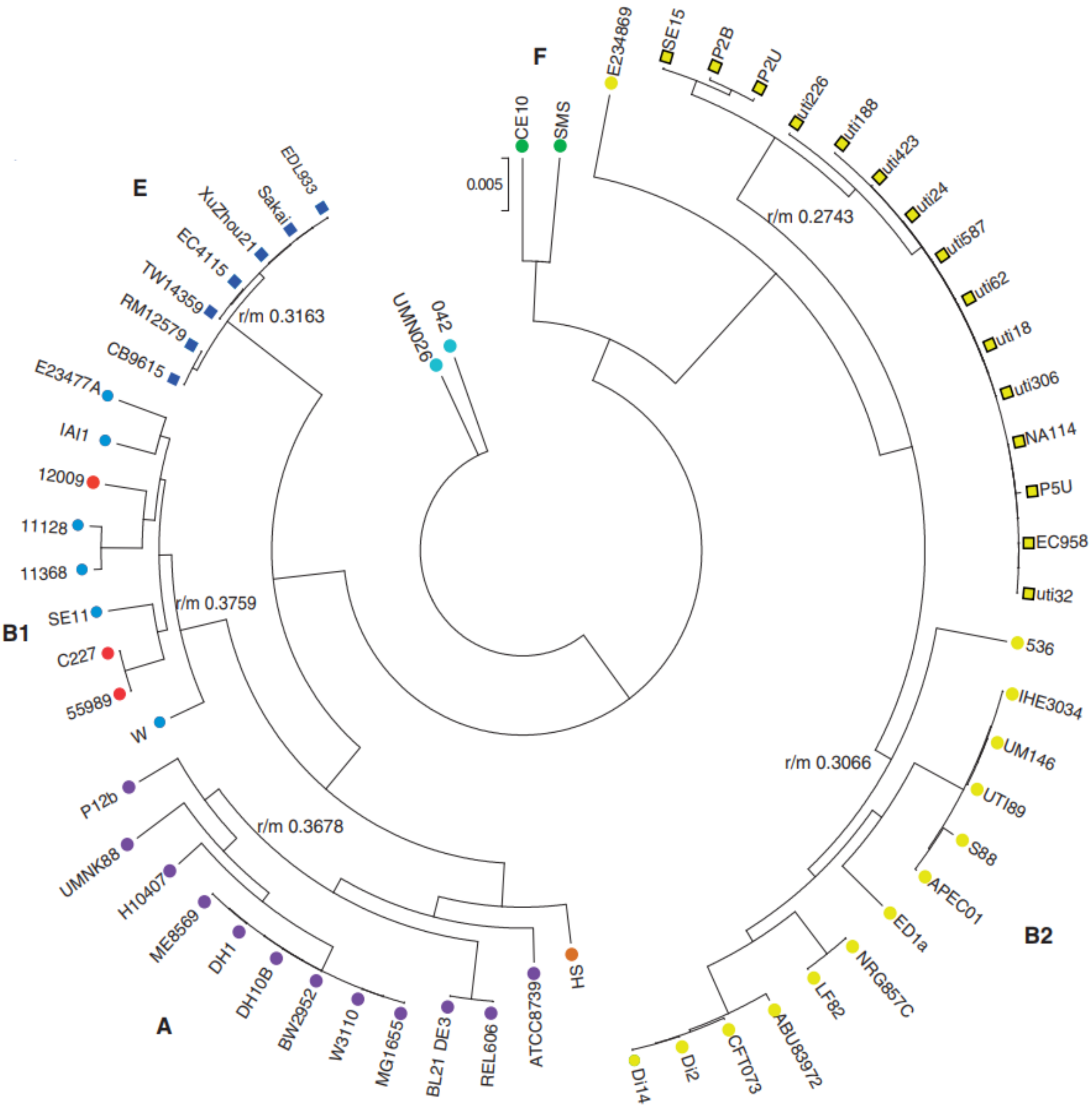


Figure 1.13. A circular core genome phylogeny of 62 *E. coli* strains constructed using a 2.3 Mb alignment. Major phylogenetic groups are labelled. The scale bar indicates the number of substitutions per site represented by the branch length shown (Figure 1 of McNally et al. 2013, Creative Commons use: https://creativecommons.org/licenses/by-nc/4.0/).

A core genome phylogeny is not likely to be fully consistent with the clonal species phylogeny (Tenallion et al. 2010). This is because it includes both recombinant and recombination-free sequence so the former may distort the underlying phylogenetic signal of clonal relationships during phylogeny construction (Tenallion et al. 2010).

Recombination can include events which occurred between ancestral lineages and or more recent inter/intra-phylogenetic group recombination events between strains (Tenallion et al. 2010). However, the inclusion of all clonal sequence in a core genome phylogeny does reduce the relative proportion of recombinant signal present in the data relative to the underlying clonal signal in the data when a phylogeny is being constructed making it a more reliable depiction of the clonal phylogeny than using a limited gene set as with MLST (Tenallion et al. 2010). Nonetheless to truly determine the *E. coli* clonal phylogeny, recombinant sequence must be removed from the core genome alignment prior to phylogenetic construction. Didelot and Falush (2006) developed a program "ClonalFrame" to remove recombinant sequences from MLST datasets. It uses substitution and indel information in a core genome alignment to construct an inferred phylogenetic tree and ancestral sequences, outputting a phylogeny which is inferred to be most supported by the underlying signal of clonal relationships in the data (Didelot and Falush. 2006). It was used by Didelot et al. (2012) to construct the *E. coli* clonal species phylogeny of A, B1, E, and B2 groups using core genes with a concatenated length of 3.3 Mb from 27 genomes (Figure 1.14). The phylogeny of the four groups described a topology which was essentially the same to that published by Kaas et al. (2012), Chaudhuri and Henderson et al. (2013), and Dunne et al. (2017). This was also the case for a ClonalFrame phylogeny constructed using 8 housekeeping genes by Tenaillon et al. (2010) with a 4,095 bp alignment from 72 *E. coli* genomes (Figure 1.15). Interestingly all 6 A-E groups had similar topologies as the core gene phylogenies seen in the three mentioned studies indicating with the Didelot et al. (2012) phylogeny (Figure 1.12), that the best estimations of the *E. coli* core genome phylogeny are essentially the same to best estimations of the clonal species phylogeny.

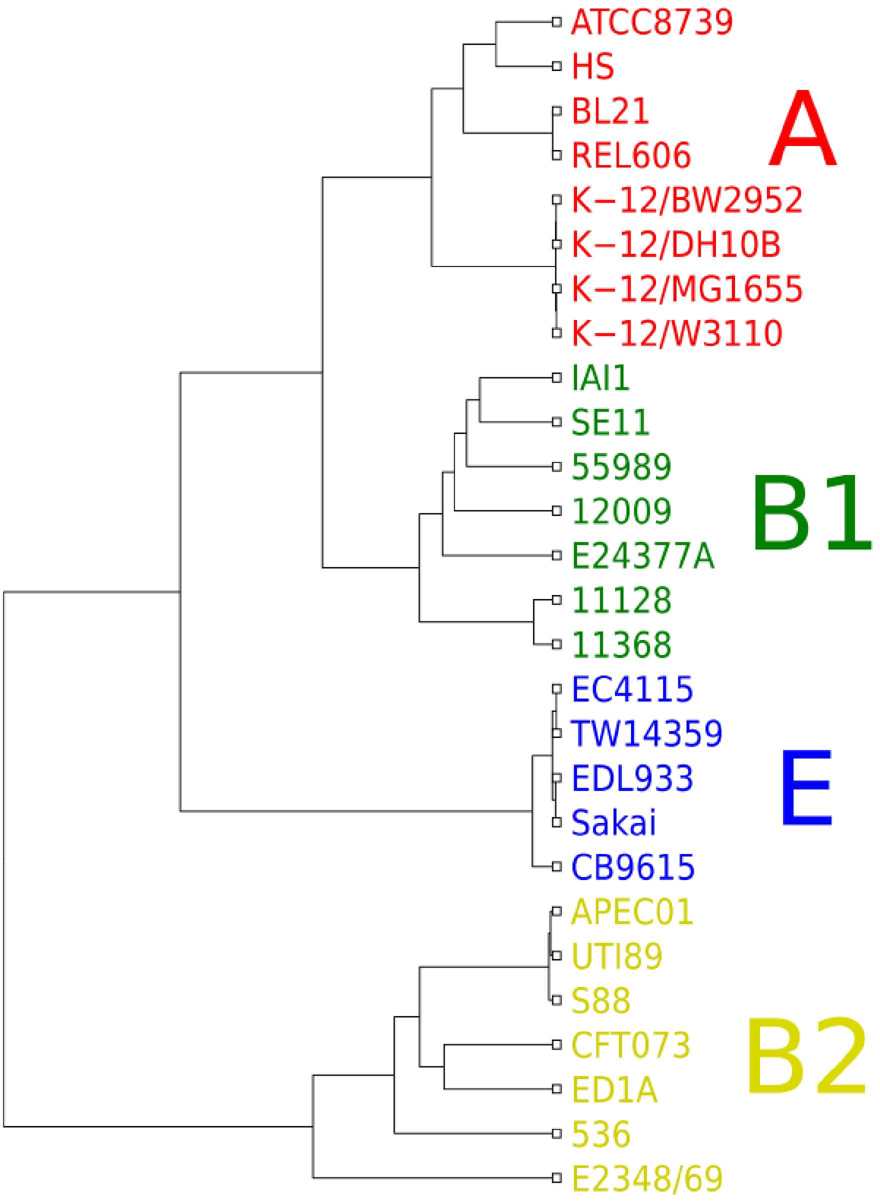


Figure 1.14. Rooted *E. coli* clonal species phylogeny of core genes with a total alignment length of 3.3 Mb using ClonalFrame (see main text) (Figure 4a from Didelot et al. 2012, Creative Commons use: http://creativecommons.org/licenses/by/4.0/).

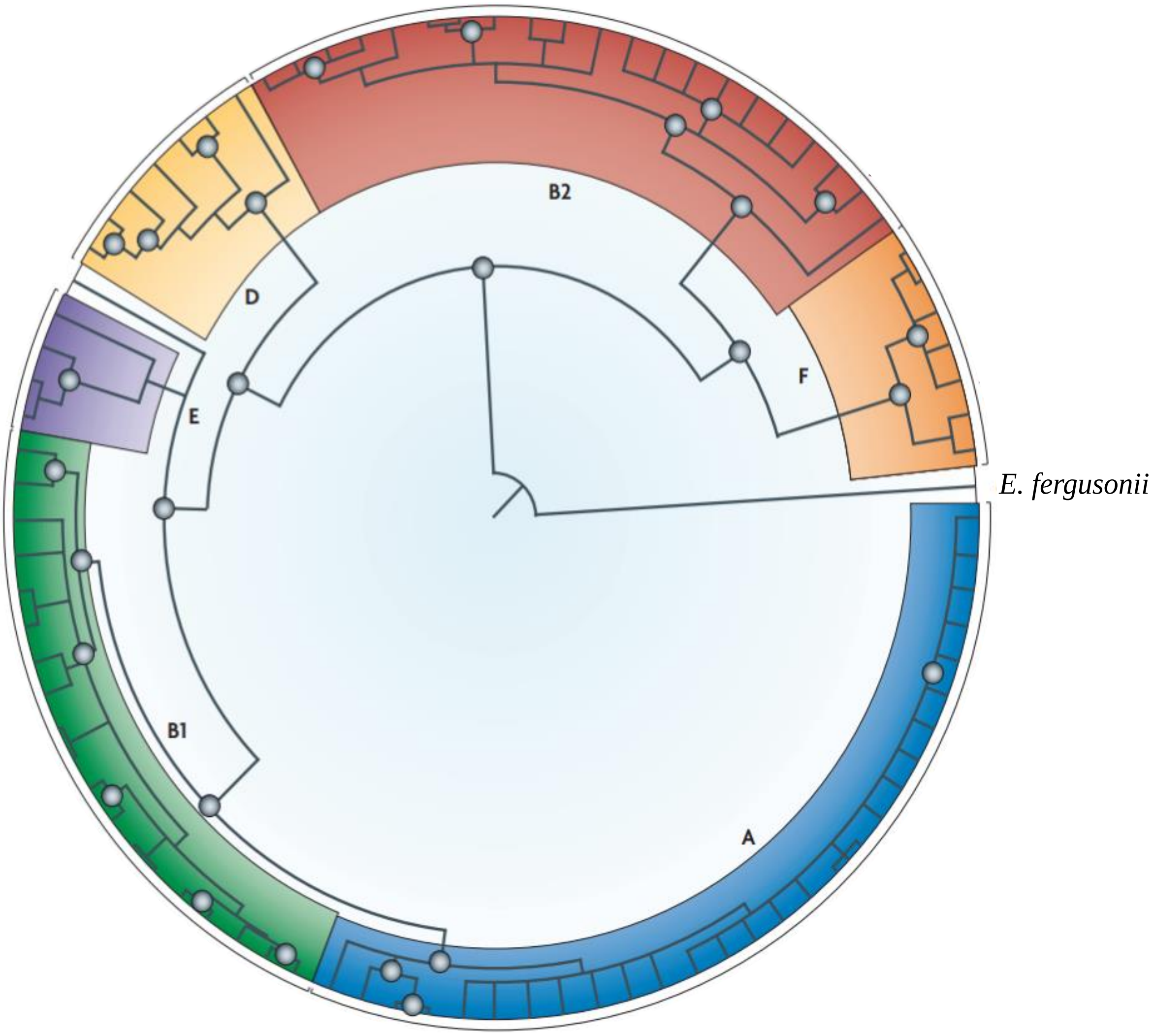


Figure 1.15. Rooted circular *E. coli* clonal species phylogeny constructed with a 4,095 bp alignment of 8 housekeeping genes in 72 *E. coli* genomes obtained using ClonalFrame (see main text), Group D is referred to as D1, and F as D2/F in the main text (Figure 3 from Tenaillon et al. 2010, used with permission from Springer Nature [COPYRIGHT] 2012).

1.3.3. Variability of the accessory genome across the A-E *E. coli* phylogenetic groups

Studies by Didelot et al. (2012) and Kaas et al. (2012) illustrated gene content variations between the phylogenetic groups. Trees which clustered strains based on shared accessory gene contents between A, B1, E, and B2 (Figure 1.16), and all 6 A-E groups (Figure 1.17) were constructed to achieve this respectively. Didelot et al. (2012) reported group E to exhibit the most unique group pan genome consistent with the finding by Touchon et al. (2009) that group E exhibited the largest number of accessory genes. Groups A was clustered with a split group B1 as a single group indicating the two groups share a high number of genes. This finding was also reported in the dengrogram published by Kaas et al. (2012). However, group B2 and not E was reported as having the most unique group pan genome.

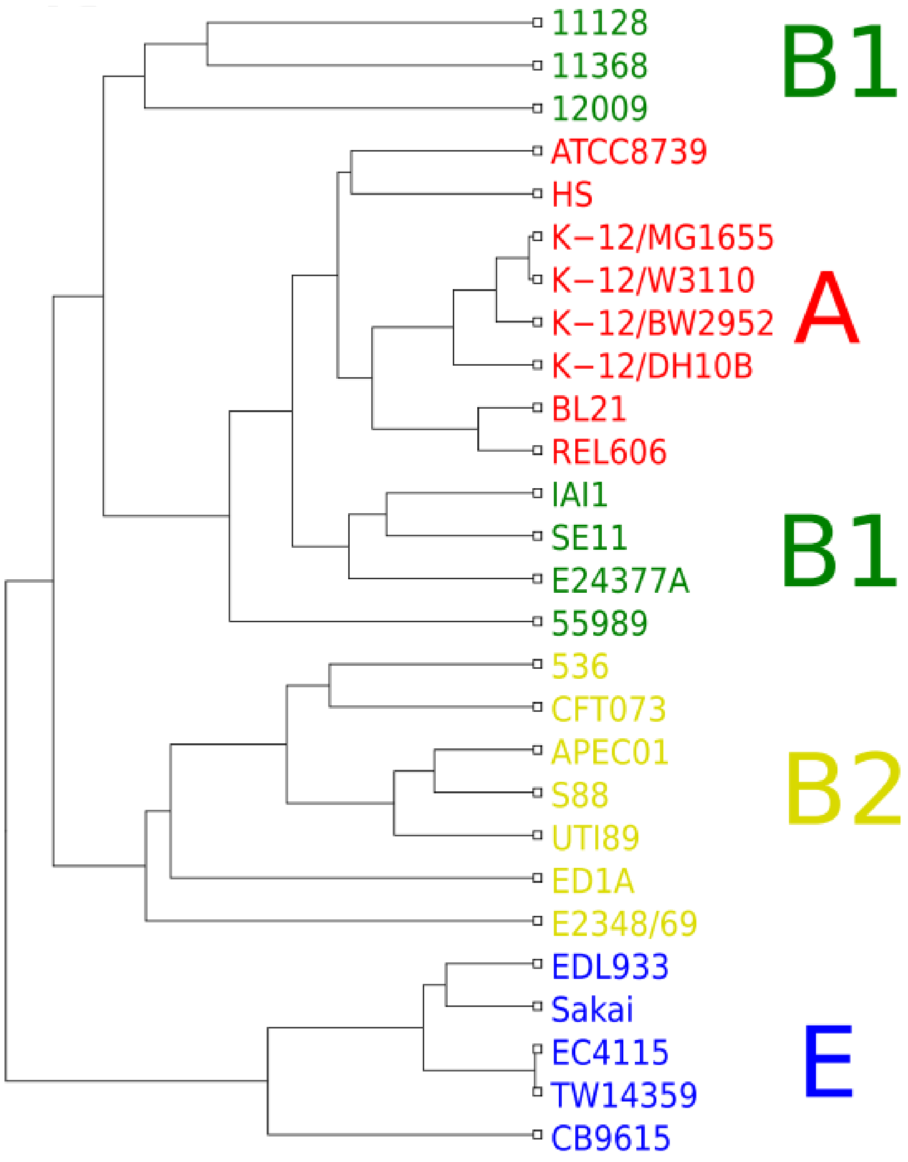


Figure 1.16. Shared accessory gene presence between 27 *E. coli* strains of the phylogenetic groups A, B1, E, and B2 illustrated as a cluster dendrogram tree which clusters strains based on shared gene contents (Figure 4b from Didelot et al. 2012, Creative Commons use: http://creativecommons.org/licenses/by/4.0/).

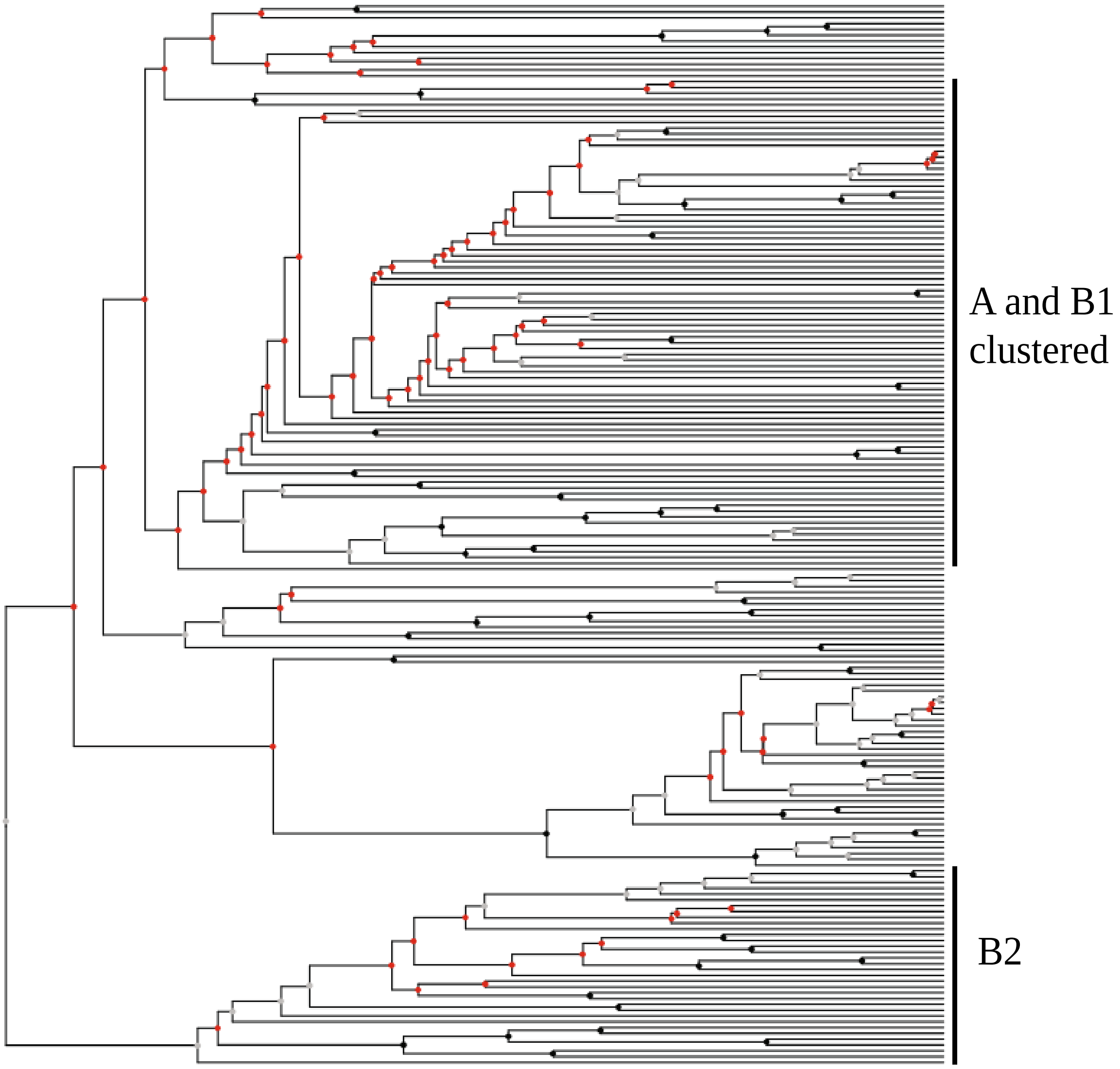
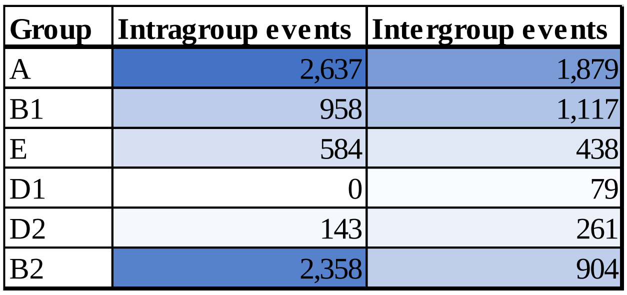


Figure 1.17. Shared accessory gene presence of 186 *E. coli* from phylogenetic groups A, B1, E, D1, D2, and B2 illustrated as a cluster dendrogram tree which clusters strains based on shared gene contents. Clustering of the phylogenetic groups A and B1 together, and B2 as identified in a core genome phylogeny are labelled (Figure 7 in Kaas et al. 2012, used with permission from Springer Nature [COPYRIGHT] 2012).

1.2.4. Recombination between and within A-E phylogenetic groups, and speciation

E. coli were originally thought to have evolved with little to no recombination (Ochman and Selander 1984, Whittam *et al.* 1983). However, more recent studies have found recombination to have significantly contributed to the diversity of the species seen today in phylogenetic groups A-E (Gonzalez-Gonzalez et al. 2013). In a study of recombination in a 2.3 Mb core genome (where core here was defined as 100% presence across strains of a given gene) of 62 *E. coli* genomes McNally et al. (2013) reported 6,680 intragroup and 4,678 intergroup recombination events (where each event involved genomic segments containing at least 1 gene) across the six phylogenetic groups defined with a core genome phylogeny (Figure 1.13) (Table 1.3).

Table 1.3. The number of intragroup and intergroup recombination events for each of the *E. coli* phylogenetic groups A, B1, E, and B2 as inferred by McNally et al. (2013).



Group	Intragroup events	Intergroup events
A	2,637	1,879
B1	958	1,117
E	584	438
D1	0	79
D2	143	261
B2	2,358	904

Didelot et al. (2012) carried out a similar analysis within the 3.3 Mb core genome of 27 complete genomes from the four groups A, B1, E, and B2 defined using a core genome phylogeny (where core here was defined as 100% presence across strains of a given gene, Figure 1.14). 18,590 intragroup (A: 2,151, B1: 6,443, E: 46, B2: 9,950) and 13,666 intergroup recombination events were reported (Table 1.4).

Table 1.4. The number of intragroup (18,590 total) and intergroup (13,666 total) recombination events each of the *E. coli* phylogenetic groups A, B1, E, and B2 are associated with, and the number of intergroup recombination events inferred to have occurred between each pair of the four groups (last four columns) inferred by Didelot et al. (2012).

Group	Intragroup	Intergroup	A	B1	E	B2
A	2,151	9,005	2,151			
B1	6,443	10,556	6,058	6,443		
E	46	548	155	230	46	
B2	9,950	7,223	2,792	4,268	163	9,950

Both studies' results are not fully consistent on whether there is a bias of intragroup or intergroup events for each of the phylogenetic groups A-E. In the McNally et al. (2013) study groups A, E, and B2 exhibited more intragroup recombination but in the latter study, it was only group B2. Where a group was found to exhibit increased intragroup compared to intergroup recombination it was suggested by authors that they are evolutionarily diverging away from the other respective A-E groups (McNally et al. 2013, Didelot et al. 2012). However, neither study provided details of gene functions encoded by recombinant regions (only quantities of contiguous core genome regions) so it is difficult to draw conclusions about the evolutionary mechanisms underlying each of the biases.

Didelot et al. (2012) speculated that the intragroup recombination bias seen in their results of the groups A+B1, E, and B2 may be explained by an adaptation of each group to separate ecological niches or reproductive cycles. This is also consistent with Touchon et al's (2009) finding that groups A+B1 and B2 are metabolically diverging. It was further suggested that disruption of normal genetic flow which usually reduces intragroup diversification could result in speciation of the three lineages in the future (Didelot et al. 2012). The gene presence tree (Figure 1.16) supports this suggestion as groups A and B1 are clustered together as a single group, whereas groups E and B2 form separate monophyletic groups indicating the three groups exhibit distinct gene contents (Didelot et al. 2012). Figure 1.17 similarly shows A+B1 and B2 clustering as separate groups (Kaas et al. 2012).

1.2.5. Recombination between ancestral lineages

The McNally et al. (2013) study uniquely includes an analysis of *E. coli* recombination including that between ancestral lineages. The authors constructed a phylogeny using core genome sequences identified as both recent and ancestral recombination using the program BratNextGen (Martinnen et al. 2011). BratNextGen determines recombination through clustering 5 kb sections of a core genome (where core here was defined as 100% presence across strains of a given gene) based on nucleotide similarity before construction of a 'shared ancestry tree' which represented the inferred pattern of clonal inheritance of strains using the clustered sequences (Martinnen et al. 2011). To infer both recent and ancestral recombination, nucleotide variations are then analysed between strains descended from major groups of this ancestry tree (Martinnen et al. 2011). The phylogeny constructed using all recombinant core genome sequence (Figure 1.18) was consistent with a core genome phylogeny of the same strains (Figure 1.13) other than group D1's position. The author suggested that this indicated there was no significant recombination between phylogenetic groups at the core genome level (McNally et al. 2013).

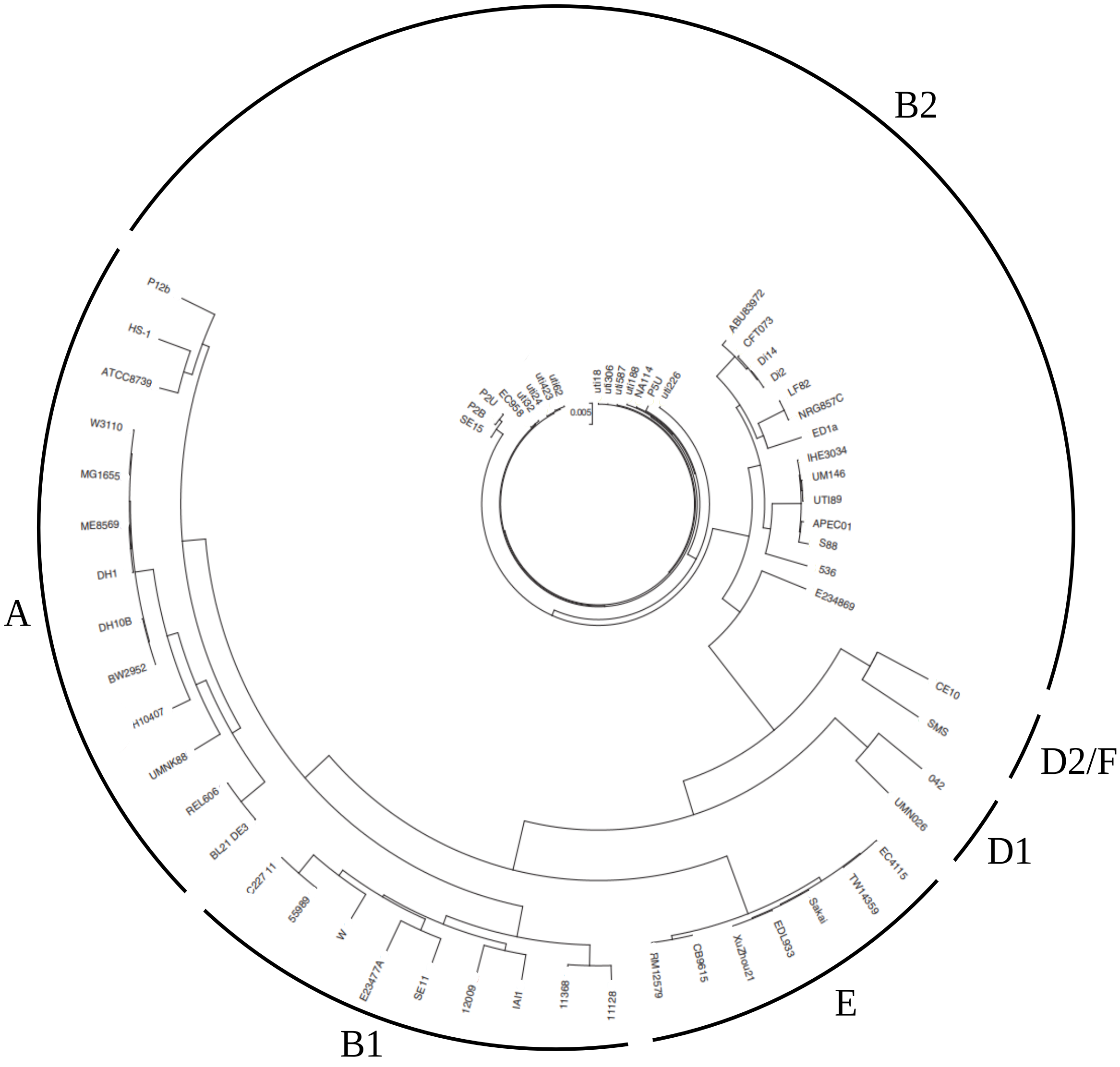


Figure 1.18. Circular maximum likelihood phylogeny of recombinant core genome regions identified by McNally et al. (2013). Major *E. coli* phylogenetic groups are labelled A-E. The scale bar indicates the number of substitutions per site represented by the branch length shown. Creative Commons use: https://creativecommons.org/licenses/by-nc/4.0/

1.4. This thesis

The overall aim of this thesis was to explore the evolutionary history and genetic diversity of *E. coli* using all publicly available *E. coli* genome sequences to contribute to the understanding of the species and impact the way it is viewed and studied. This aim was approached in the work presented across Chapters 3, 4, 5, and 6 in different ways. Of the work presented in Chapter 3, the overall aim was to create a reference set of *E. coli* genome sequences representative of the full genetic and sample diversity available across all publicly available sequences. It was then to use the reference set to determine an up-to-date *E. coli* clonal frame phylogeny and to test the objectives designed to produce specific information about *E. coli* evolution. This was followed by Chapter 4, for which the overall aim of the work presented was to explore evidence to support the preferential use of a proposed core gene MLST schema with specific benefits in increased accuracy and reduced analysis time compared to existing schemas. The work presented in Chapter 5 focused on pathogenicity within a specific lineage of 20 group B2 UPEC strains, analysed through computational analysis of genomes and phenotype data. The overall aim of the work was to reveal the genetic basis to UPEC ureter contractility inhibition phenotypes and provide insights about the evolution of such phenotypes. The work presented in the final research chapter focused on the evolution of the *E. coli* type three secretion system II (ETT2) and *eip* gene clusters which have been previously described as having genotypes subject to mutational attrition and implicated in virulence. The overall aim of the work was to present an up-to-date account of the evolutionary history and genetic diversity of these clusters and to test objectives designed to confirm the results of previous research, provide a deeper understanding of how their various genotypes evolved, and infer any likely ETT2 or *eip* cluster phenotypes based on genotype.

Chapter 2: General Methods

2.1. General nucleotide and protein sequence manipulation

Custom programs for sequence manipulation were written in the programming language Perl (version 5, Wall 1994) using BioPerl modules (Stajich et al. 2002) or Python (version 3, Sanner et al. 1999) using Biopython modules (Cock et al. 2009) depending on program requirements. Structure of input and output files and parameters were defined and the optimal way to parse nucleotide and protein sequences to produce the desired manipulated output was planned. Programs were written using a text editor within a Linux BASH computing environment and tested on a reduced version of the file which would be analysed by the program. Parallelisation of programs was implemented using GNU Parallel (Tange 2011), a command line tool which allowed processes to be distributed across multiple cores. Graphical plots were displayed using R (R core team 2013).

2.2. Obtaining bacterial strain genome sequences

Complete and draft bacterial genome sequences were downloaded in GenBank or plain FASTA format from GenBank at the National Centre for Biotechnology Information (NCBI) accessible via https://www.ncbi.nlm.nih.gov.

2.3. Filtering genome sequences by assembly quality

Genomes were filtered based on genome assembly quality so had to have an ‘N50’, i.e. size-ordered median contig length, of 100,000 bases to pass filtering. Filtered out genomes were replaced with a genome with higher assembly quality.

2.4. Filtering genome sequences by sample information

Genomes were filtered firstly based on author, isolate, and strain name information to remove closely-related strains, indicated by slight differences in the strain name. To do this, strain names were programmatically checked for identical strings of letters and identification numbers and subsequently manually inspected. Genomes published by the same author and which originated from identical isolation sources were also filtered to prevent the inclusion of groups of genome sequences with limited genetic diversity. This was facilitated by a program which reported genomes with similar author and sampling information based on identical strings of letters and numbers in the annotated genome sequence files, followed by manual inspection.

2.5. Annotation of gene sequences using Prokka

To standardise coding sequence annotations for downloaded strain genomes, Prokka (Seemann 2014) was used. Prokka is a prokaryotic genome annotation pipeline that predicted and functionally annotated bacterial genomic features to the strain genome sequences. Prokka used BLAST+ (Altschul et al. 1990) for homology searches, Prodigal (Hyatt 2010) for coding sequence prediction, RNAmmer (Lagesen et al. 2007) for annotation of rRNA genes, Aragorn (Laslett and Canback 2004) for annotation of tRNA genes, SignalP (Petersen et al. 2011) to identify signal peptides, and Infernal (Kolbe and Eddy 2011) to predict potential non-coding RNA genes. Genes with no predicted function were annotated as “hypothetical protein”.

2.6. Annotation of gene sequences: Non-Prokka methods

Annotations for gene sequences were obtained by carrying out a protein BLAST to identify orthologues in *E. coli* genomes annotated by the Wellcome Trust Sanger Institute, which have high quality manually inspected annotations. These were the strains E2348/69 (Iguchi et al. 2009), 042 (Chaudhuri et al. 2010), and H10407 (Crossman et al. 2010) with GenBank accessions FM180568, FN554766, and FN649414 respectively. Two other genomes with high-quality annotation were also used for this purpose, strains K-12 MG1655 (Blattner et al. 1997) and O157 H7 Sakai (Hayashi et al. 2001) with GenBank accessions U00096 and BA000007 respectively. Information about metabolic profiling was obtained for genes present in *E. coli* genome sequences 042 and K-12 MG1655 as these data have been made available by Chaudhuri et al. (2010) and Feist et al. (2007) respectively. Next, putative bacterial cell functional information relevant to molecular, pathogenic, and ecosystem related molecular pathways was obtained for each gene (where available) using the online service: Kyoto Encyclopedia of Genes and Genomes (KEGG, Kanehisa and Goto 2000).

2.7. Pan genome analysis using Roary

Roary (Page et al. 2015) was used to carry out pan-genome analysis to compile a pan-genome and identify core genes (present in ≥99% of strains) and genes shared across all strain genome combinations across the strain genome set. Prior to running Roary analysis, a set of genome sequences with standardised annotations obtained using Prokka (Seemann 2014) was collated. These were used as input for Roary, and a predetermined appropriate user-specified percentage identity cut off was specified as a parameter to be used for identifying pairs of orthologues during analysis.

In analysis using the Prokka-annotated genome set, Roary first extracted coding sequence coordinates, then collated protein sequences. It then carried out filtering to remove incomplete sequences and pre-clustering using CD-HIT (Fu et al. 2012). The protein sequences were then subjected to an all-against-all protein BLAST and matches meeting the user specified percentage identity criterion were recorded**.** Match sequences were then subjected to clustering with MCL (Enright et al. 2002) and merged with the earlier pre-clustering results generated with CD-HIT. Next, conserved gene neighbourhood information was used to group orthologous sets of sequences and split paralogues into groups of true orthologues (Page et al. 2015). Roary produced individual gene and concatenated core gene sequence alignments and gene presence and absence data for all strain combinations. These alignments and information were then kept for later use in relevant analyses.

2.8 Using BLAST to obtain a set of orthologous gene sequences

Reference protein sequences were collated using a Python program which employed the module Biopython (Cock et al. 2009) which extracted specified protein sequences from an annotated genome sequence using a list of gene names as input. Genome sequences that were to be searched for the reference sequences were then prepared. This involved extracting all protein sequences for each strain genome from an annotated GenBank or Prokka file. Protein BLAST (Altschul et al. 1997) analysis was then used to detect localised sequence alignments which were optimal between the genome sequences and the reference sequences. For BLAST analysis using a given reference sequence, computation began with separation of each triplet of reference sequence amino acids and detection of their frequency within each genome sequence, location of occurrence, and details of cases where one of the three residues in the triplet differs. Next matches to each triplet were identified and alignment scores were computed with length and identity statistics for matches of the whole reference sequence to a location within each of the genome sequences (Altschul et al. 1997). Using an in-house program named 'Mutualbest', the reciprocal BLAST analysis was then performed where the protein sequences of a given genome took the place of the reference gene (Tatusov et al. 1997). The results of both analyses were then compared. If the best hit matches for the original reference gene to a protein sequence in the first BLAST analysis were reciprocated in the second BLAST analysis (i.e. the reference gene was the best hit for the same protein), the program reported the original reference gene and the protein to be orthologue matches. If the BLAST amino acid identity between the orthologues was greater than the predetermined cut-off, the genes were recorded as true orthologues. The orthologue matches for the given reference gene were then extracted from the BLAST result files in either protein or nucleotide sequence format.

2.9. Nucleotide and amino acid alignment using Muscle

To carry out an alignment of nucleotide or amino acid orthologue or homologue gene sequences to a reference sequence, all gene sequences were first collated into a single file. Using the input file, Muscle (Edgar et al. 2004) was run with default parameters for a “100 iterations: slow accurate” analysis. In Muscle analysis, the similarity of each pair of sequences was first determined using k-mer (short adjacent sets of residues with specific states) counting and through global alignment of each pair of sequences to calculate a fractional identity for the pair. Distance estimates were then obtained by computing a triangular distance matrix from sequence similarities between sequence pairs. A distance rooted tree was then constructed from the matrix and progressive alignment followed using the order of branching in the tree as a guide, which resulted in a multiple alignment. Similarity between sequence pairs was then determined using the fractional identity information computed from each pairs’ mutual alignment in the current alignment. Using this information, construction of a second tree then occurred using a Kimura distance algorithm-based method. The new and first tree were then compared, and ancestral tree nodes in the new tree in which branching order differs were simultaneously identified. This was repeated until the number of branches differing in the new tree compared to the first tree did not decrease any further before progressive alignment followed using the new tree as a guide. Next, iterative refinement of the new multiple sequence alignment occurred. In each iteration internal branches were removed to divide sequences into two subsets of putatively clustered strains before the multiple alignment profile of each subset was extracted and columns containing no residues were discarded. The two profiles were then re-aligned to one another using profile to profile alignment and a summed alignment score was calculated. If the score for those two exact profiles was higher than that of a previous iteration, the new alignment was kept but discarded if not. Once all internal branches had been used for dividing sequences into subsets and no observed score change was seen or if a user defined iteration number was reached this last phase terminated, otherwise it was repeated, before the new multiple sequence alignment was outputted (Edgar et al. 2004). Finished Muscle alignments were manually inspected, and minor corrections were made if necessary using Seaview (Galtier et al. 1996), an alignment editing tool that differentially colours residues for ease of viewing.

2.10. Maximum Likelihood phylogeny construction using RAxML

RAxML (version 8, Stamatakis et al 2014) was used for maximum likelihood phylogeny construction. As a parameter, the general time reversible model of nucleotide substitution was used with nucleotide alignments, which has six parameters that allow a different rate of substitution for each pair of nucleotides (Zwickl and Holder 2004). For amino acid alignments the WAG model was used, a model based on empirical observation of amino acid exchangeability in families of closely related amino acids (Whelan and Goldman 2001). 100 bootstrap replicates were selected as a parameter prior to analysis, meaning analysis would be repeated 100 times to provide a measure of support for each constructed branch in the phylogeny. In analysis, RAxML first employed the principle of likelihood, which is the probability of observing the patterns in the sequence data given the assumptions used in either the GTR or WAG substitution models and a particular topology (Graur et al. 2000). The likelihood was calculated for each alignment site by consideration of each unknown ancestral state and calculation of their associated probabilities. Several hypothetical trees were trialled, and the likelihood of each was calculated as the product of the likelihood values for all alignment sites. This was calculated as the sum of the logarithms of the likelihoods for each site, or log likelihood (*lnL*). The analysis was repeated 100 times (specified with the 100 bootstrap replicates parameter) and the phylogenetic branches which were supported most frequently across all 100 analysis replicates were included in the final outputted phylogeny.

Chapter 3: Insights into the evolution of *Escherichia coli*

3.1. Introduction

5,623 *E. coli* whole genome sequences from diverse sources have been published and deposited in GenBank, with an increased proportion of these published in the last 10 years (Jiang et al. 2014, Tang 2016, Hur and Young 2015, Lawsin et al. 2017, Messerer 2016). This is most likely due to progressive developments in and decreasing costs of bacterial whole genome sequencing (Quainoo et al. 2017). These *E. coli* genome sequences collectively represent an unprecedented degree of sample diversity. They were sampled from a wide range of hosts and environments worldwide and likely represent previously unreported genetic diversity. Previous genetic evolutionary research was limited by lack of diverse sequences and additionally relatively recently, computational resources capable of genomic analysis of hundreds of genomes over hours or days (Quainoo et al. 2017). The purpose of the work in this chapter was to make use of the available genome sequence data and produce an up-to-date narrative describing deduced major *E. coli* evolutionary events through addressing a hypothesis, an overall aim, and objectives using a computer with up-to-date processing capabilities for genomic analysis. I predicted that the full dataset of 5,623 *E.coli* genomes is too large for practical computational analysis within reasonable times, but the sample and genetic diversity present within the full set can instead be represented using a carefully selected subset of 100 genomes. Producing a species phylogeny using a genome set representative of the species such as this is arguably the most important component of any study of a species' evolutionary history. Species phylogenies are also reference frameworks for which other evolutionary investigations are based upon. In the case of *E. coli*, determining the species phylogeny involves determining the clonal frame phylogeny, which is constructed by excluding recombinant core gene sequence and only using the remaining sequence in phylogeny construction. This is sequence which has a history of vertical inheritance which depicts the clonal frame.

3.1.1. Hypothesis

The hypothesis for the work presented this chapter was designed to determine if the deduced clonal frame phylogeny created using an up-to date set of *E. coli* strain genomes is consistent with previous accounts of the *E. coli* clonal frame phylogeny published by Didelot et al. 2012 and Tenaillon et al. 2010 (Figure 1.14, Figure 1.15).

The *E. coli* clonal frame phylogeny constructed using an up-to-date representative set of *E. coli* genome sequences depicts consistency with the established clonal frame phylogeny for groups A-E and the core gene phylogeny for cryptic clade groups, by exhibiting a divergence pattern of six major *E. coli* phylogeny groups and four cryptic clade groups which have an established clustering pattern (Figure 3.1).

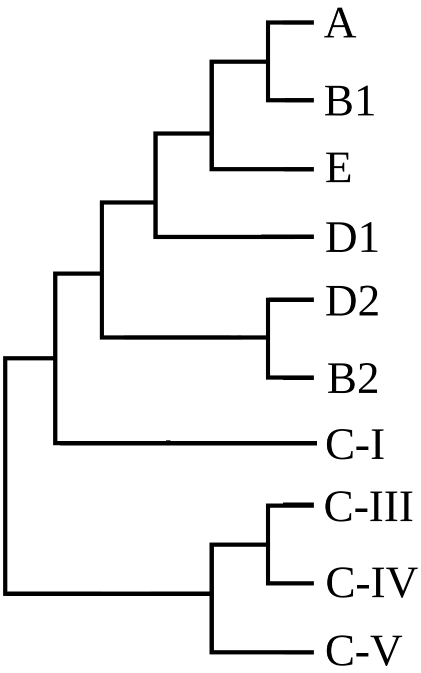


Figure 3.1. Established phylogenetic clustering patterns observed in the clonal frame phylogeny for *E. coli* phylogenetic groups A-E and the core gene phylogeny for cryptic clade groups.

3.1.2. Aims and objectives

The overall aim was to create a reference set of *E. coli* genome sequences representative of the full sample and genetic diversity available across all publicly available sequences as of January 2018. To use the reference set to determine the *E. coli* clonal frame phylogeny and to test the following objectives designed to produce specific information about *E. coli* evolution:

1. To determine the prevalence and type of recombination events which occurred between the major phylogenetic group ancestors (pre-divergence; before major groups diverged), compared to more recently between strains or lineages of separate groups (post-divergence; after major groups diverged).
2. Identify what proportion of each *E. coli* major phylogenetic group's pan genome is shared with that of each other group.
3. Identify which specific genes can be deduced to have most contributed to the divergence of each *E. coli* major phylogenetic group, and to what functional categories do they belong.

3.2. Methods

3.2.1 Filtering strain genomes based on phylogenetic diversity

To filter strain genomes by the relative amount of phylogenetic diversity they contribute, a given strain genome was selected manually from the total number of strain genomes on the basis of its position in a phylogeny. Strains represented in the phylogeny which were separated by a relative longer branch than those separating other strains or clusters of strains were manually picked from the phylogeny to create a set of phylogenetically diverse strain genomes.

3.2.2 Assigning *E. coli* phylogenetic groups to strains

A given *E. coli* strain genome was assigned a phylogenetic group manually on the basis of its relative, close phylogenetic clustering proximity to a published phylogenetic group representative, all of which have a complete genome sequence, other than those for groups C-III, C-IV, and C-V (Table 3.1).

Table 3.1. Details of published strain genomes which represent the *E. coli* phylogenetic groups A-E and cryptic clade groups C-I to C-V.

Phylogenetic group	Representative	GenBank Accession
A	str. K-12 substr. MG1655	U00096
B1	O104:H4 str. 2009EL-2050	CP003297
B2	O127:H6 str. E2348/69	FM180568
D1	042	BA000007
D2	SMS 3 5	FN554766
E	O157:H7 str. Sakai substr RIMD 0509952	CP000970
C-I	TW10509	GL872204
C-III	RCE03	JUDX00000000
C-IV	TW11588	AEMF00000000
C-V	TW09308	AEME00000000

3.2.3. Quartet analysis

Individual core gene alignments which comprised the core gene alignment outputted by Roary were extracted. A new reordered core gene alignment was constructed by concatenating gene alignments using the order that each gene was observed in the strain genome K-12 substr. MG1655 as a guide. The reordered core gene alignment was used as input into quartet analysis, which was a method based on quartet phylogenetic inference (Strimmer and Von Haeseler 1997). Prior to analysis the alignment was divided into sliding window sections of size 10 kb. The analysis worked by taking four individual sequences (one from each phylogenetic group) from each 10 kb section and reporting which of the three possible 4-taxa unrooted phylogenetic trees were supported by different 10 kb sections (Figure 3.2). The analysis was repeated for different combinations of sequences.

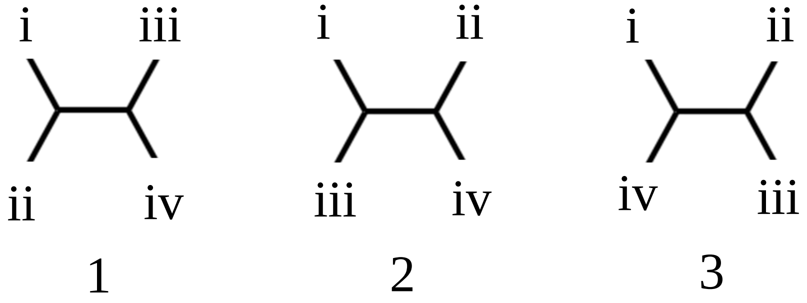


Figure 3.2. The three possible unrooted phylogenies for four phylogenetic groups represented by numerals i, ii, iii, and iv.

The method was implemented in Python version 3 (Sanner et al. 1999), using the Python module ETE Toolkit (Huerta-Cepas et al. 2016). For each window, the method cycled through strains of a specific "test" phylogenetic group and each strain was compared in turn with a consistent set of 3 reference strains chosen from 3 other phylogenetic groups to create a quartet phylogeny of the four strains. The best supported topology and the bootstrap support value for the middle branch was recorded for each quartet. If bootstrap values were ≥ 50% the quartet phylogeny was considered for further analysis. The pre-divergence phylogenetic clustering pattern for the four groups for each window was taken to be the pattern exhibited by ≥ 50% (the majority) of quartet phylogenies computed for that window. The alternative clustering patterns for each section were then recorded as post-divergence recombination events for each analysis between the given four phylogenetic groups. For example, if phylogenetic group A clusters with group E and D1 with B2 ((A, E) (D1, B2)) in ≥ 50% of quartets within a particular window it is considered the pre-divergence clustering pattern for that window section and clustering of A with D1 or B2 relative to E would be considered evidence of post-divergence recombination clustering pattern occurring in that window region. If no single pattern made up ≥50% of the quartet phylogenies the window was deemed phylogenetically unresolved and excluded from further analysis.

For a single window, the pre-divergence clustering pattern was deduced as that present in ≥ 50% of quartets. For a whole quartet analysis, the pre-divergence clustering pattern for the four phylogenetic groups was determined to be the most prevalent pre-divergence pattern across windows in the analysis. For the whole *E. coli* species, the pre-divergence clustering pattern (the *E. coli* clonal frame clustering pattern) was defined as the most prevalent pre-divergence clustering pattern observed across all 35 possible quartet analyses of different combinations of phylogenetic groups. The other clustering patterns were therefore considered as pre-divergence recombinant clustering patterns and taken as evidence of pre-divergence recombination events. These pre-divergence clustering patterns were observable in single quartet analyses between sets of four phylogenetic groups and in individual windows within a given quartet analysis.

3.2.4. Obtaining the pan genome for each phylogenetic group and identifying unique and shared genes

A program was written to process the gene presence/absence data file produced by Roary and extract information for a set of strains from a specific phylogenetic group (or groups). The program calculated the group pan genome size, the number of core and accessory genes present, and the number of core and accessory genes which were unique or highly enriched within the group, as well as the number shared with each of the other groups.

3.2.5. Determining genes unique and highly enriched in phylogenetic groups and clades with an in-house program

An in-house program was written that took the gene presence absence data file outputted by Roary as input with a list of strain names to analyse for the given phylogenetic group or groups. The program cycled through genes present for each strain and identified genes present in and unique to ≥ 90% of strains of each of the phylogenetic groups or group of phylogenetic groups (a clade).

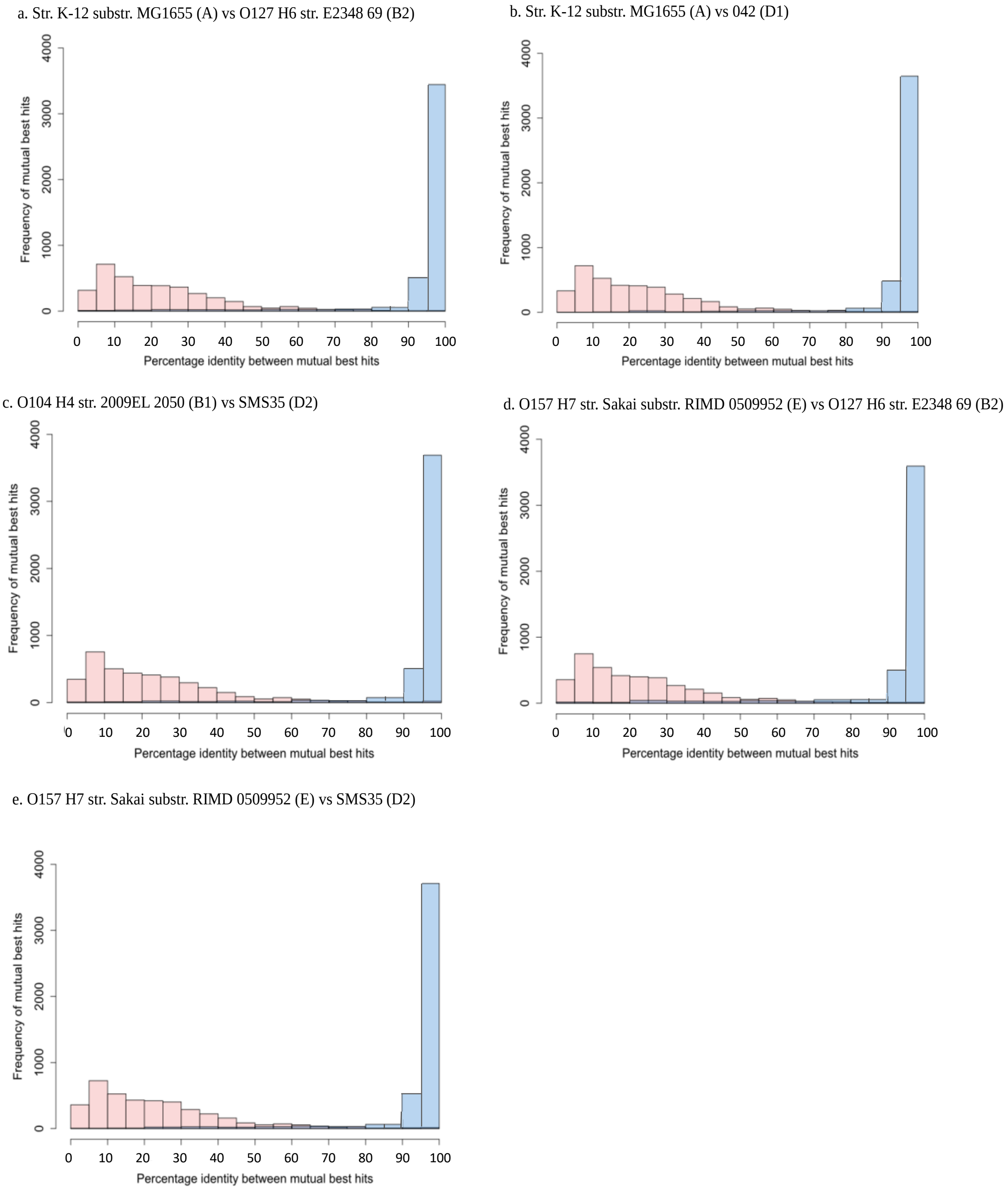
3.3. Results

3.3.1. Creating a reference set of *E. coli* genome sequences representative of available sample and genetic diversity

A literature review was conducted, after which 5,623 complete and draft *E. coli* genome sequences were downloaded. After filtering for genome assembly quality, this was reduced to 4,923, and after filtering to remove genomes with identical sample information to other genomes, the number was reduced to 4,170. To determine how phylogenetically diverse the 4,170 strain genomes were in relation to each other, it was necessary to observe phylogenetic branch lengths separating strain genomes in a core gene phylogeny. Construction of a complete core genome phylogeny from all 4,170 genomes was inferred to have not been possible with the available computational resources, so it was decided an alignment of 500 core *E. coli* genes from each of the 4,170 genomes would be obtained. Using 500 genes was a compromise which included sufficient polymorphic nucleotide sites to produce a phylogeny representative of the core genome phylogeny, whilst remaining potentially computationally tractable.

To obtain orthologues for 500 *E. coli* individual core genes from each of the 4,170 genomes, reference core genes were first compiled from *E. coli* K-12 substr. MG1655

(GenBank accession U00096). To identify their orthologues in the other 4,169 genomes, it was first necessary to determine an appropriate amino acid identity cut-off value to apply in a BLAST analysis. To do this, a mutual best hit and second best hit analysis was carried out (Tatusov et al. 1997) between all gene sequences (protein format) encoded by the reference genomes (Table 3.1) for each phylogenetic group. Based on the mutual best hit analysis, the appropriate *E. coli* identity cut-off value for identifying an orthologue for a given gene sequence in BLAST was determined to be 95% identity (Figure 3.3). The reason for this was because across comparisons in the mutual best plots, 95% was observed to mark the sequence identity value between where the greatest dip from best to second best hit matches occurred.


a. Str. K-12 substr. MG1655 (A) vs O127 H6 str. E2348 69 (B2)

b. Str. K-12 substr. MG1655 (A) vs 042 (D1)

c. O104 H4 str. 2009EL 2050 (B1) vs SMS35 (D2)

d. O157 H7 str. Sakai substr. RIMD 0509952 (E) vs O127 H6 str. E2348 69 (B2)

e. O157 H7 str. Sakai substr. RIMD 0509952 (E) vs SMS35 (D2)

Figure 3.3. Histogram of percentage amino acid identity for mutual best and second-best BLAST hits between representatives of the phylogenetic groups A-E. For each comparison in a-e, the best hits (blue) represent matches between potentially orthologous genes and second-best hits (red) represent matches between non-orthologous genes. Both plots are transparent, so the overlap can be seen. The percentage identity value at which orthologues can be identified between strains of these groups was determined to be the crossover point between the distributions, estimated at 95% for all group comparisons.

A BLAST analysis of all gene sequences in protein format was performed using the 500 *E. coli* reference core genes against the 4,170 genome sequences and the 95% identity cut-off deduced as appropriate for identifying *E. coli* orthologue sequences. The output of the BLAST analysis was inspected and orthologues for *E. coli* reference core genes present in 100% of the 4,170 strains were retained. A gene from *E. coli* strain genome K-12 substr. MG1655 was extracted and used for identifying orthologues followed by BLAST report inspection and the process continued until orthologues for 500 *E. coli* reference core genes were collated from each of the 4,170 genomes. The 500 orthologue groups were concatenated, aligned using Muscle and construction of a maximum likelihood phylogeny was attempted using RAxML. However, this proved to be computationally too complex. Smaller numbers of genes were also investigated: 50, 80, 120, 150, 300. It was determined that using orthologues from 120 *E. coli* reference core was an optimal compromise of maximising phylogenetic information, whilst maintaining a reasonable time to compute the phylogeny.

On inspecting the phylogeny of the 120 concatenated and aligned core genes (Figure 3.4), of particular interest was a clade of 29 strains that did not cluster closely with groups A, B1, E, D1, D2, or B2. These were labelled as the putative group G. Two strains did not cluster with any phylogenetic group (labelled in blue in Figure 3.4). The available genomes from the cryptic clade phylogenetic groups C-I to C-V (27 genomes) were considered too few to fully represent the diversity of those groups, so were excluded from the 120-gene phylogeny and further analysis.

After genomes from phylogenetic groups A-G were sampled for phylogenetic diversity using the 120-core gene phylogeny as a guide, 723 diverse genomes were chosen. After further sampling for pathovar and sample diversity, a preliminary set of genomes from 50 commensal, environmental, and laboratory strains, and 50 pathogenic strains were selected.

This set of 100 strains (highlighted with red dots in Figure 3.4 and listed in full in Table 3.2) represented the full phylogenetic and sample diversity observed in the original 4,170 strains and are the focus of the subsequent work in this chapter.

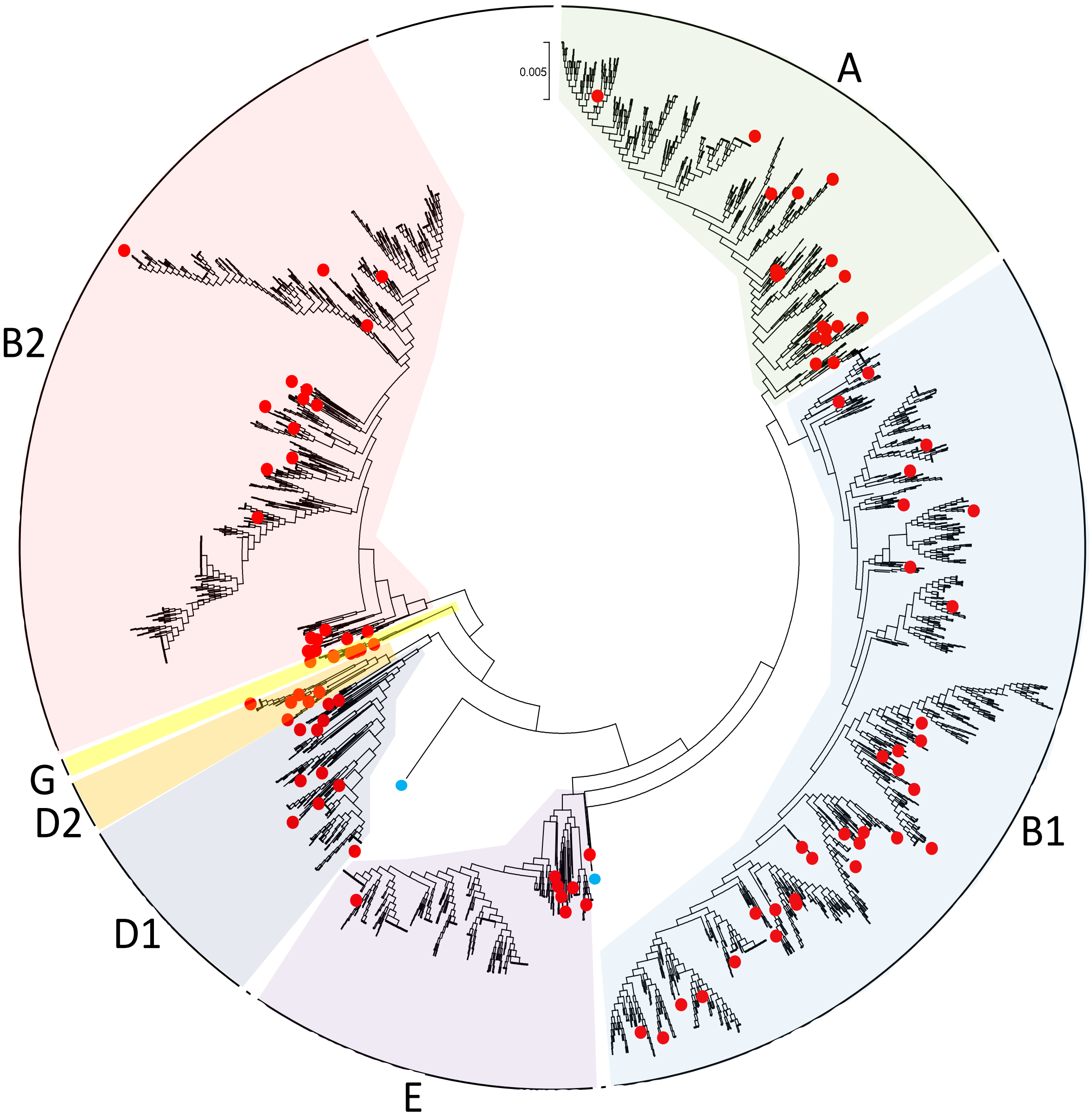


Figure 3.4. Unrooted RAxML maximum likelihood phylogeny constructed using an alignment of 120 core genes (167,376 bp) from 4,170 *E. coli* strain genomes obtained from GenBank with an N50 greater than 100,000 bp. The major phylogenetic groups are labelled A, B1, E, D1, D2, B2 in the outer ring, with gaps in this ring indicating group borders. 29 strains that did not cluster closely with those in A, B1, E, D1, D2, or B2, and are clustered as a sister group to B2 are labelled G. 50 commensal, environmental, and laboratory strains and 50 pathogenic strains were chosen from these strains to be a set of 100 A-G group strains that represent all *E. coli* phylogenetic diversity and are labelled here with a red dot. 2 strains which did not cluster with any group are labelled with a blue dot. The scale bar at the top shows the number of substitutions per site represented by branches of the indicated length.

Table 3.2. Strain information for the 50 commensal, environmental, and laboratory strains, and 50 pathogenic strains in the 100 *E. coli* strain set representing phylogenetic groups A-G.

Group	Strain name	Pathovar or environment	Genome length (bp)	GC content (%)	Accession
A	101-1	EAEC	4,979,723	50.63	AAMK00000000
A	1303	MPEC	4,948,797	50.73	CP009166
A	25	Water	4,766,232	50.79	CXYK00000000
A	53638	EIEC	5,066,886	51.10	AAKB00000000
A	ATCC_8739	Commensal (human)	4,746,218	50.87	CP000946
A	cattle16	Commensal (non human)	4,740,871	50.85	LVLZ00000000
A	CFSAN026836	Water	5,344,232	50.52	LDCY00000000
A	D6-117	MPEC	4,787,132	50.78	CCCP00000000
A	H1	Water	4,826,483	50.89	CP010160
A	H10407	ETEC	5,153,435	50.76	FN649414
A	H5	Water	4,833,228	50.75	CP010169
A	HS	Commensal (human)	4,643,537	50.82	CP000802
A	S1	Soil	4,707,208	50.84	CP010226
A	S30	Soil	5,072,853	50.67	CP010231
A	S43	Soil	5,043,711	50.64	CP010237
A	str. K-12 substr. MG1655	Laboratory	4,641,652	50.79	U00096
A	UMNK88	ETEC	5,186,406	50.72	CP002729
A	VL2732	MPEC	4,664,032	50.65	JTFD00000000
B1	3.5-R3	Commensal (human)	5,184,306	50.65	MOZF00000000
B1	APECO78	APEC	4,798,433	50.68	NC_020163
B1	C11	Commensal (non human)	5,414,571	50.83	CP010133
B1	C2	Commensal (non human)	4,818,237	50.73	CP010117
B1	C5	Commensal (non human)	5,633,965	50.45	CP010122
B1	D6	Commensal (non human)	4,910,852	50.92	CP010148
B1	E10019	EIEC	5,376,124	50.72	AAJW02000000
B1	E267	Commensal (human)	4,281,347	51.00	ADIN00000000
B1	ECC-1470	MPEC	4,803,751	50.78	CP010344
B1	ECOR29	Commensal (non human)	4,952,372	50.57	LYAH00000000
B1	ECOR45	Commensal (non human)	4,710,028	50.68	LYCD00000000
B1	ECOR58	Commensal (non human)	5,421,303	50.11	LYCY00000000
B1	ECOR67	Commensal (non human)	4,757,712	50.87	LYDN00000000
B1	ECOR68	Commensal (non human)	4,986,911	50.67	LYDF00000000
B1	H14	Water	4,735,489	50.76	CP010177
B1	H15	Water	4,857,969	50.82	CP010178
B1	H3	Water	4,679,162	50.81	CP010167
B1	M10	Commensal (non human)	4,954,801	50.81	CP010200
B1	M18 2	Commensal (non human)	5,160,136	50.52	CP010219
B1	O104:H4 str. 2009EL-2050	EAEC	5,438,174	50.59	CP003297
B1	O111:H- str. 11128	EHEC	5,371,077	50.62	AP010960
B1	O139:H28 str. E24377A	ETEC	5,249,287	50.56	JXRF00000000
B1	O1O3:H2 str. 12009	EHEC	5,524,860	50.63	AP010958
B1	O96:H19 CFSAN029787	EIEC	4,947,515	50.75	CP011416.1
B1	S10	Soil	4,886,210	50.77	CP010229
B1	S3	Soil	4,632,368	50.66	CP010228
B1	S42	Soil	4,838,808	50.86	CP010236
B1	S50	Soil	4,981,720	50.71	CP010238
B1	S56	Soil	4,992,522	50.84	CP010242
B1	St_Olav17	STEC	5,452,601	50.49	JYKT00000000

Table 3.2 continued.

Group	Strain name	Pathovar or environment	Genome length (bp)	GC content (%)	Accession
E	400654	EPEC	5,408,807	50.22	CYBM00000000
E	AF85	MPEC	5,305,698	50.58	MIVV00000000
E	B185	Commensal (human)	5,106,856	50.60	ACXF00000000
E	C161 11	EAEC	5,254,629	50.45	AIAI00000000
E	D6-113	MPEC	5,082,312	50.51	CCCO00000000
E	O157:H16 str. Santai	EHEC	5,104,557	50.60	CP007592
E	O157:H7 str. Sakai substr RIMD 0509952	EHEC	5,594,477	50.48	BA000007.2
E	O169:H41 str. F9792	ETEC	4,923,453	50.41	JHJJ00000000
D1	042	EAEC	5,241,977	50.56	FN554766
D1	B354	Commensal (human)	4,831,929	50.55	ACXG00000000
D1	C1	Commensal (non human)	4,843,023	50.54	CP010116
D1	C4	Commensal (non human)	4,989,757	50.57	CP010121
D1	EC2	ExPEC	5,018,127	50.55	JFJL0000000
D1	ECOR48	Commensal (human)	5,426,246	50.42	LYCA00000000
D1	TA255	Commensal (human)	4,883,612	50.62	ADJG00000000
D1	TA280	Commensal (human)	5,258,156	50.59	ADBA00000000
D1	UMN026	UPEC	5,202,090	50.72	CU928163
D1	upec-213	UPEC	4,946,350	50.46	JSKZ00000000
D2	24.1-R1	Commensal (human)	5,076,999	50.38	MOYU00000000
D2	BIDMC 19C	UPEC	5,523,477	50.43	AXLI01000000
D2	HVH 87_4	Bacteremia	5,502,156	50.54	AVVI01000000
D2	IAI39	UPEC	5,132,068	50.63	CU928164
D2	SMS35	Soil	5,068,389	50.50	CP000970
D2	swine65	Commensal (non human)	5,140,443	50.30	LVOP00000000
D2	UCI 57	UPEC	5,036,717	50.65	JMVT00000000
G	71	Commensal (non human)	4,836,112	50.87	CXXK00000000
G	APECO2-211	APEC	5,112,508	50.63	CP006834
G	cattle19	Commensal (non human)	5,372,330	50.42	LVMC00000000
G	CFSAN026806	STEC	5,160,057	50.69	LHCQ00000000
G	HVH 79 (4-2512823)	Bacteremia	5,113,135	50.72	AVVC01000000
G	KTE75	UPEC	5,734,748	50.62	ANUO01000000
G	MDR 56	Commensal (human)	4,978,170	50.82	CP019903
B2	173	Commensal (human)	4,927,547	50.63	LM996590
B2	536	UPEC	4,938,920	50.52	CP000247.1
B2	403128	EPEC	4,915,103	50.52	CXZV01000000
B2	200135 aEPEC	EPEC	4,887,438	50.62	CYBG01000000
B2	401480 aEPEC	EPEC	4,878,079	50.57	CYGR00000000
B2	B2 12-1-TI12	AIEC	5,017,481	50.51	1.
B2	B671	Commensal (human)	5,075,767	50.67	ADIP00000000
B2	blood-10-1310	Bacteremia	5,569,438	50.32	JSPY00000000
B2	C262 10	EAEC	4,727,165	50.53	AIAP00000000
B2	C796 10	EPEC	4,626,542	50.75	AIBS00000000
B2	CFT073	UPEC	5,231,148	50.48	AE014075.1
B2	ECOR65	Commensal (non human)	4,944,210	50.76	LYDD00000000
B2	H588	Commensal (human)	4,719,304	50.67	ADIQ00000000
B2	HVH 193 (4-3331423)	Bacteremia	4,998,133	50.63	AVYT01000000
B2	NMECO18	NMEC	5,002,781	50.75	CP007275
B2	O127:H6 str. E2348/69	EPEC	5,069,678	50.52	FM180568
B2	O83:H1 str. NRG857C	AIEC	4,894,875	50.69	CP001855
B2	SCB-11	NMEC	5,105,498	50.48	JSYT00000000
B2	SE15	Commensal (human)	4,717,338	50.74	AP009378
B2	TOP382-2	Commensal (human)	5,094,267	50.41	AOQD00000000

1. https://datacommons.anu.edu.au/DataCommons/rest/records/anudc:5410/data/B2_12-1-TI12.gbk

To confirm the phylogenetic diversity of all strains in the 100-strain genome set, it was necessary to conduct pan genome analysis and use the resultant core genome alignment to create a core genome phylogeny through conducting phylogenetic analysis. To ensure consistent annotations for the pan genome analysis, the 100 genomes were all reannotated using Prokka, and pan genome analysis was conducted using Roary using the 95% orthologue identity as an inputted parameter.

Roary produced a 2.34 Mb core gene alignment, and from this an *E. coli* core gene phylogeny was constructed using RAxML (Figure 3.5). In this phylogeny groups G and B2 clustered together as an outgroup to the remaining strains. Group D2 was the next to branch off, followed by group D1 and group E, with groups A and B1 clustering together. There was 100% bootstrap support for the basal branches of all 7 phylogenetic groups. The phylogeny was consistent with the results of the 120 gene phylogeny for all 4,170 strains, so the set of 100 genomes was confirmed for use in all subsequent analyses.

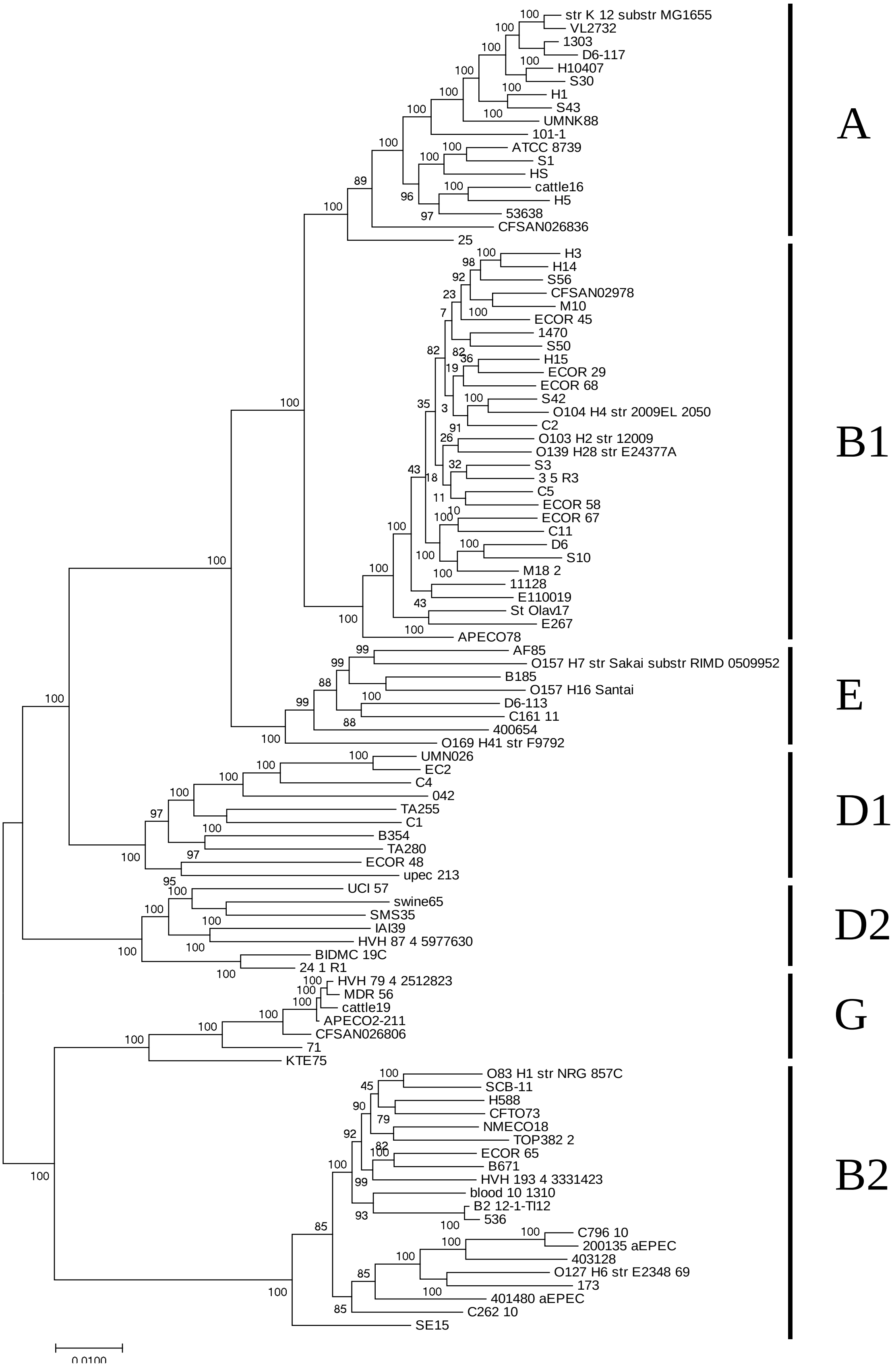

100
str K 12 substr MG1655
100
VL2732
1303
100
D6-117
100
H10407
100
S30
100
H1
100
S43
100
UMNK88
100
101-1
A
100
ATCC 8739
100
S1
100
HS
89
cattle16
100
H5
96
53638
100
97
CFSAN026836
25
100
H3
98
H14
92
S56
23
CFSAN02978
M10
100
7
ECOR 45
1470
S50
82
82
36
H15
19
ECOR 29
ECOR 68
100
S42
35
3
O104 H4 str 2009EL 2050
C2
91
26
O103 H2 str 12009
O139 H28 str E24377A
B1
32
S3
43
18
3 5 R3
C5
11
ECOR 58
10
100
ECOR 67
100
C11
100
D6
100
S10
100
100
M18 2
11128
E110019
43
St Olav17
100
E267
100
APECO78
99
AF85
99
O157 H7 str Sakai substr RIMD 0509952
88
B185
O157 H16 Santai
E
100
99
D6-113
C161 11
88
400654
100
O169 H41 str F9792
100
UMN026
100
EC2
100
C4
042
100
TA255
97
C1
100
B354
D1
TA280
100
97
ECOR 48
upec 213
95
UCI 57
100
swine65
100
SMS35
100
IAI39
D2
100
HVH 87 4 5977630
100
BIDMC 19C
24 1 R1
100
100
HVH 79 4 2512823
100
MDR 56
100
cattle19
100
APECO2-211
G
100
CFSAN026806
100
71
KTE75
100
O83 H1 str NRG 857C
45
SCB-11
90
H588
79
CFTO73
92
NMECO18
82
TOP382 2
100
ECOR 65
100
B671
B2
100
HVH 193 4 3331423
99
blood 10 1310
B2 12-1-TI12
93
536
100
100
C796 10
100
200135 aEPEC
100
403128
85
O127 H6 str E2348 69
173
100
100
401480 aEPEC
85
C262 10
SE15
0.0100

Figure 3.5. Midpoint rooted phylogeny of the 100 genome *E. coli* reference set. The tree was constructed by maximum likelihood analysis of a 2,338,727 bp core genome alignment obtained from 50 commensal, laboratory and environmental isolates and 50 pathogenic *E. coli* strains chosen to represent the full phylogenetic diversity of *E. coli*. Percentage bootstrap support values are shown on internal branches. The scale bar on the bottom left indicates the number of substitutions per site represented by the branch length shown. Major phylogenetic groups are labelled, including a sister group to group B2 which is putatively labelled group G.

The final set of 100 *E. coli* strains genomes exhibited broad diversity in terms of phylogenetic group, genome length, and type of isolate (pathogenic, commensal, laboratory or environmental). Pathogenic strains were selected to maximise the diversity of pathovars (Table 3.2, Figure 3.6).

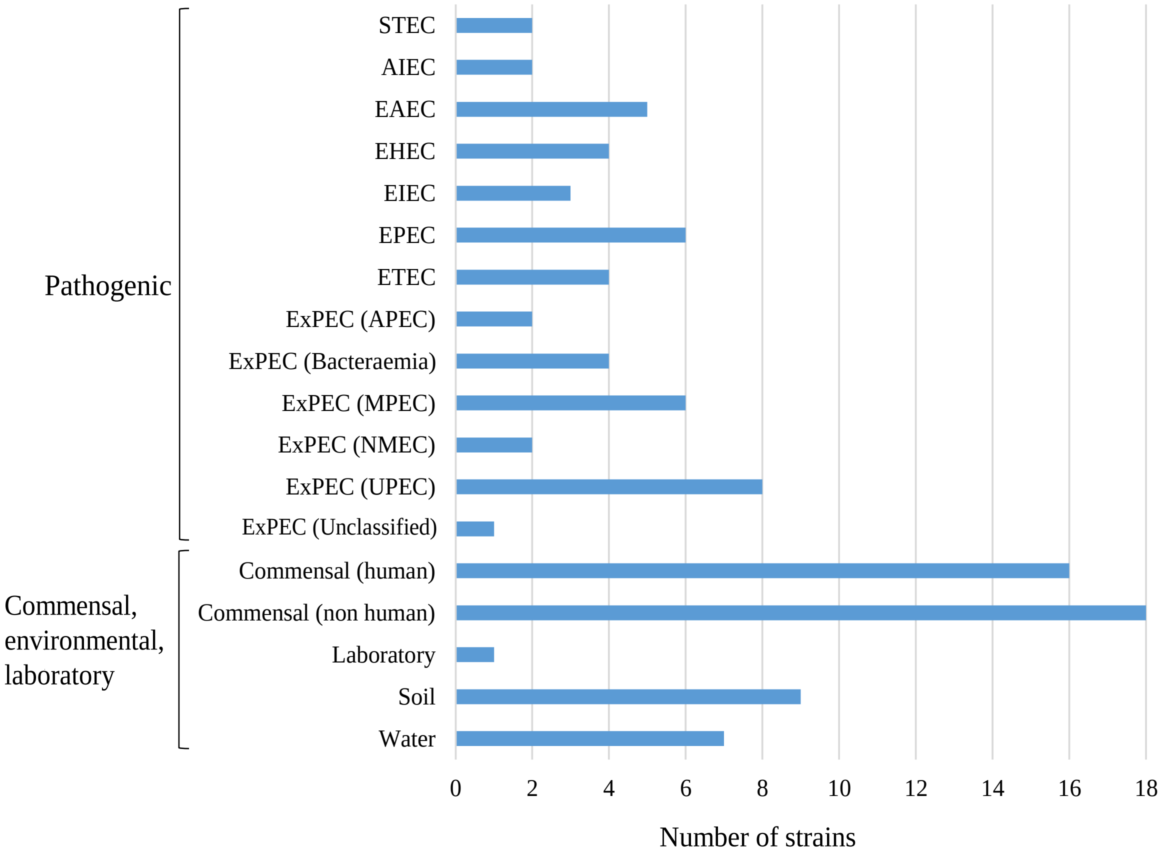
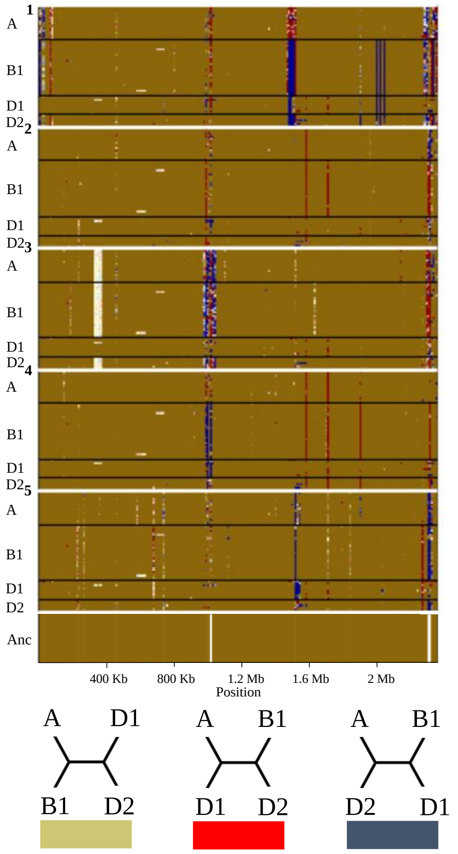


Figure 3.6. The number of strains from the 100 *E. coli* set of strain genomes from phylogenetic groups A-G grouped according to pathovar, or the commensal or environmental source of isolation. The set includes 50 pathogenic and 50 commensal, environmental, and laboratory strains. Pathogenic *E. coli* acronyms are as follows: STEC: Shiga toxigenic *E. coli*, AIEC: adherent-invasive *E. coli*, EAEC: enteroaggregative *E. coli*, EHEC: enterohaemorrhagic *E. coli*, EIEC: enteroinvasive *E. coli*, EPEC: enteropathogenic *E. coli*, ETEC: enterotoxigenic *E. coli*, APEC: avian pathogenic *E. coli*, MPEC: mammary pathogenic *E. coli*, NMEC: neonatal meningitis-associated *E. coli*, UPEC: uropathogenic *E. coli*.

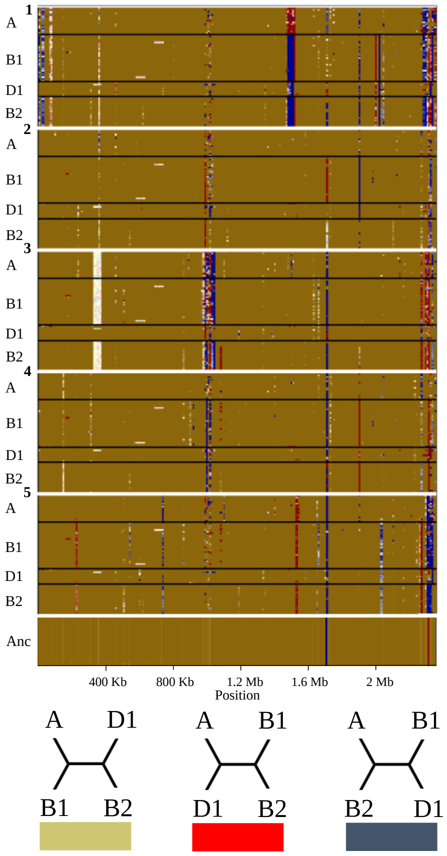
3.3.2. Estimating the *E. coli* clonal frame phylogeny

To infer the *E. coli* clonal frame phylogeny, an analysis named ‘quartet analysis’ was first applied, using the core gene alignment which was produced by Roary pan genome analysis using the 100 set of *E. coli* strain genomes, as input. The quartet analysis produced 35 plots, one for each quartet analysis, detailing the clustering patterns of 10 kb windows, between sets of 4 groups along the core gene alignment (Figure 3.7).

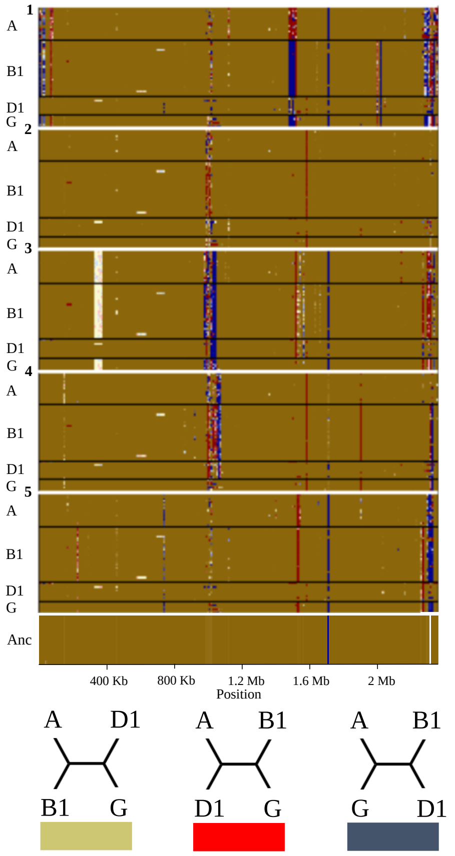
A, B1, D1, D2



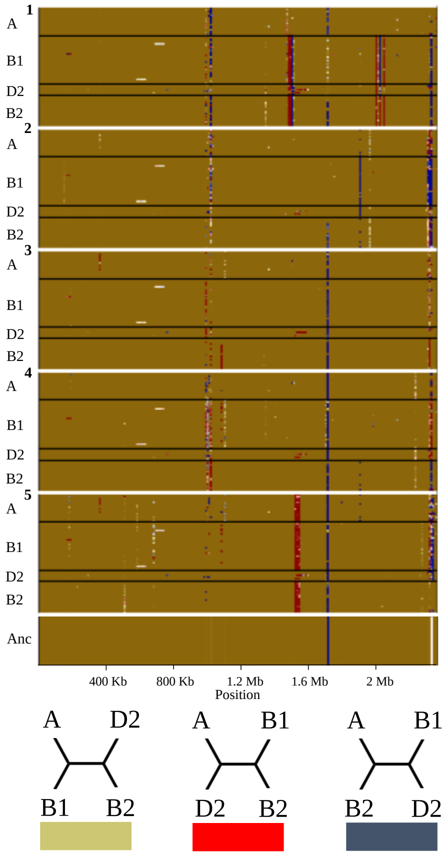
A, B1, D1, B2

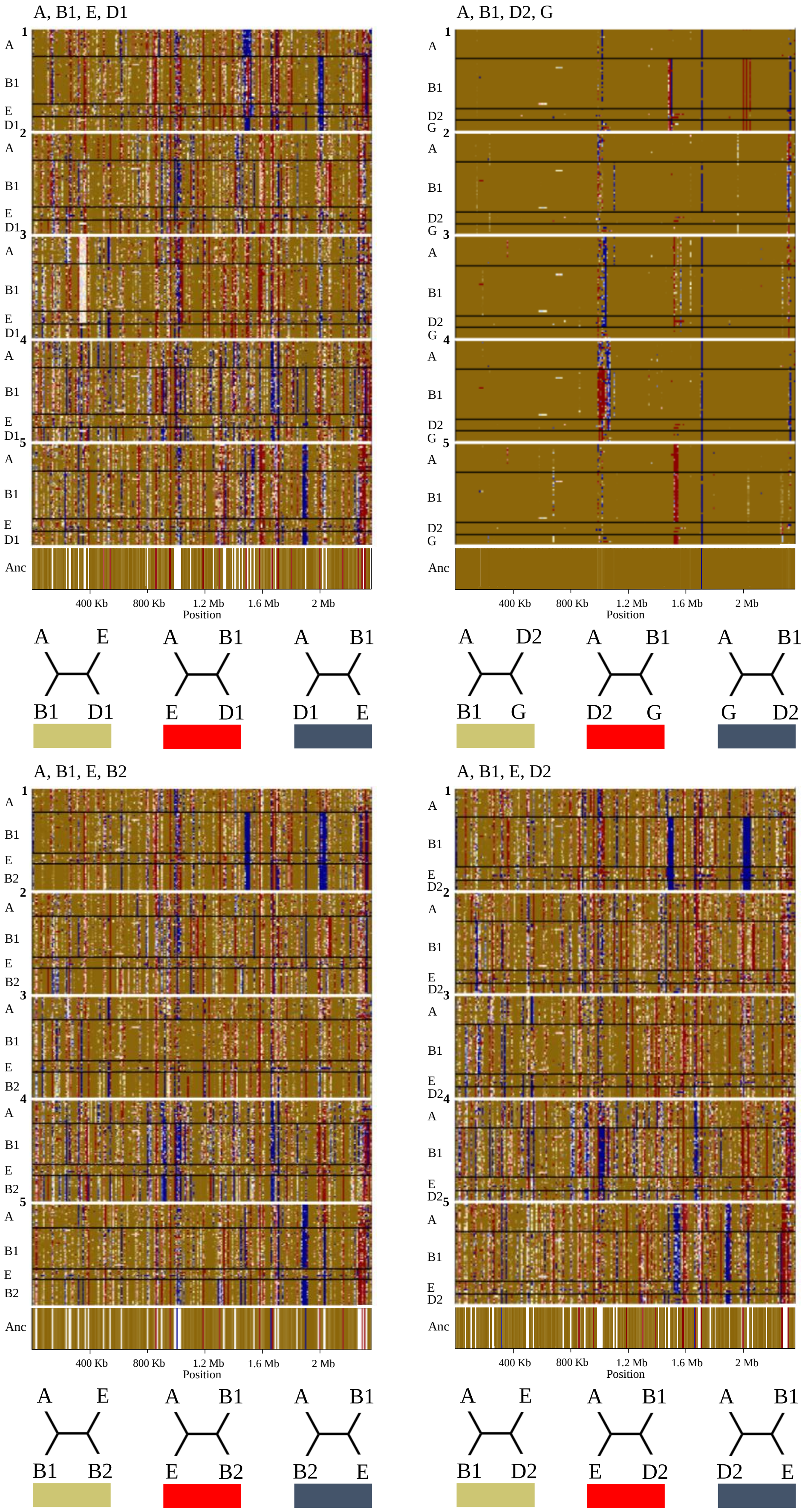


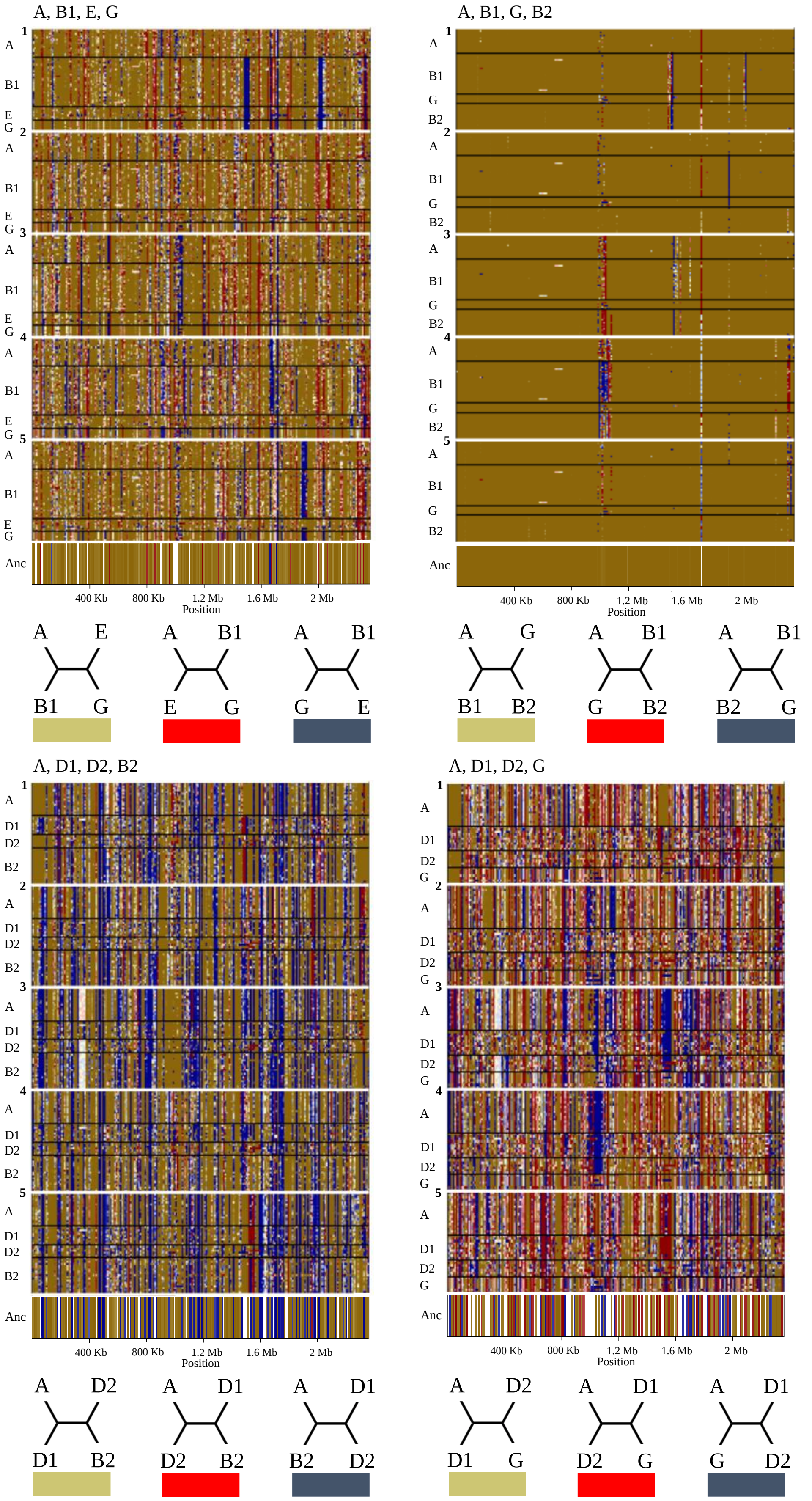
A, B1, D1, G

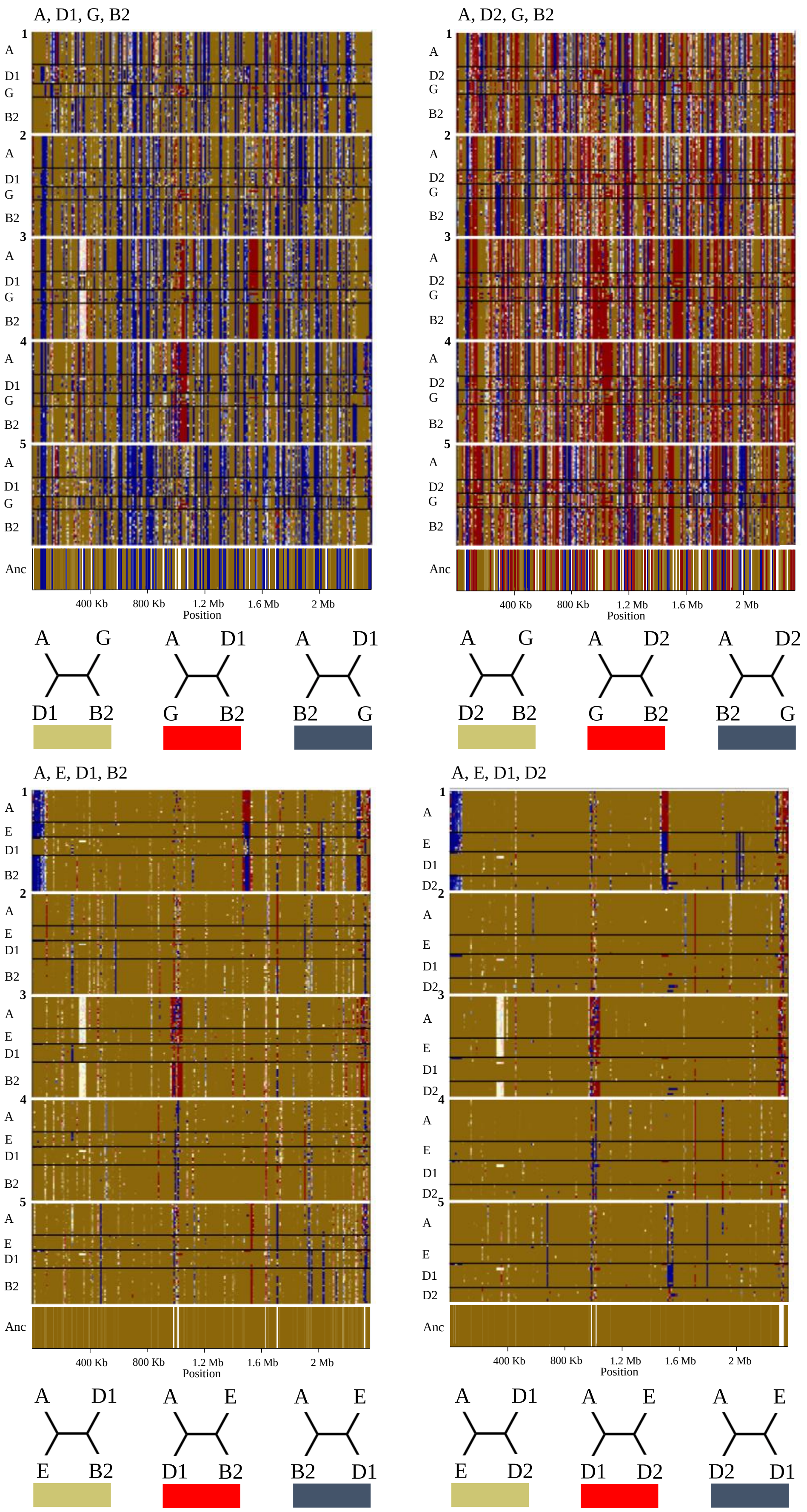


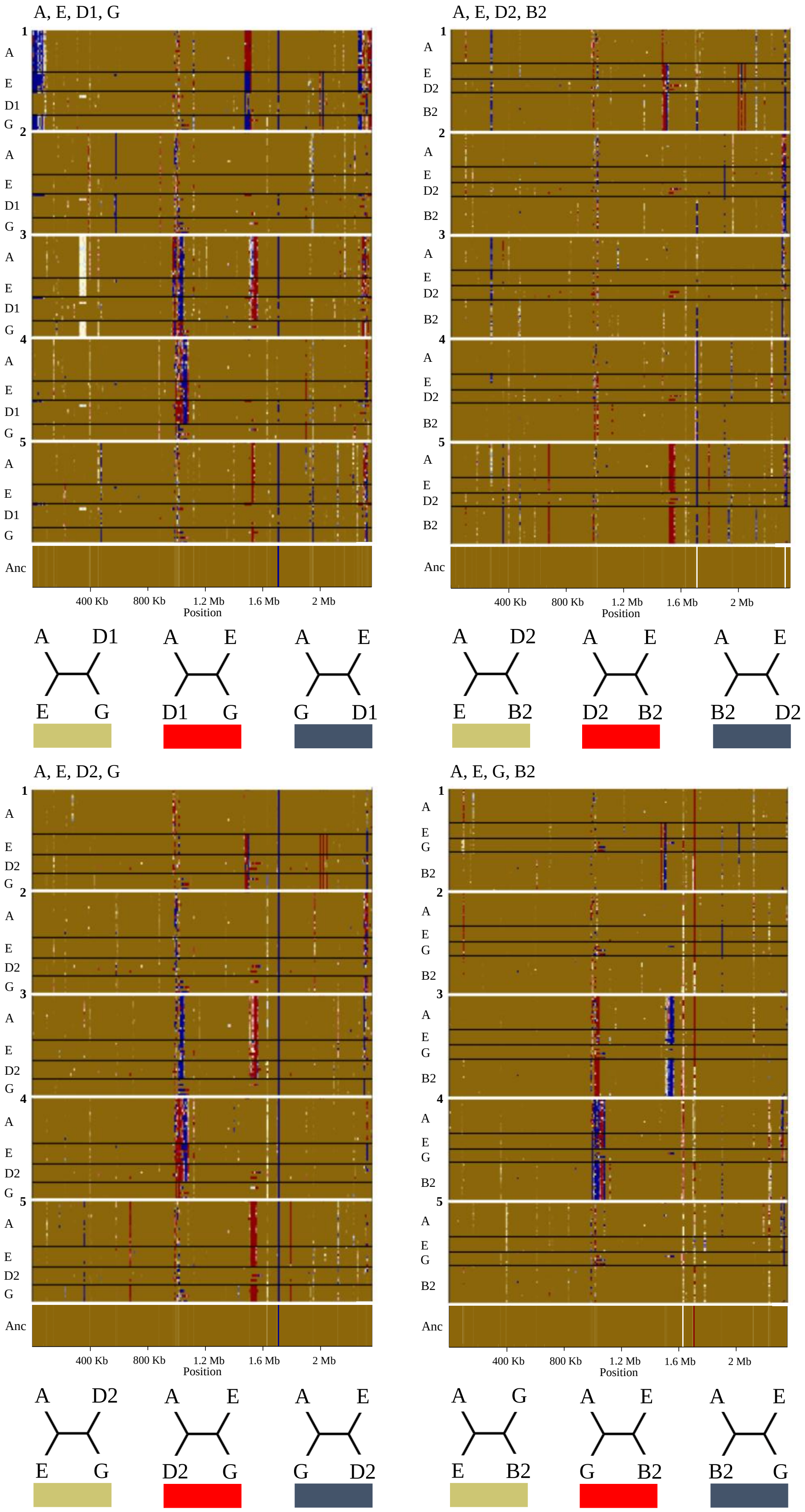
A, B1, D2, B2



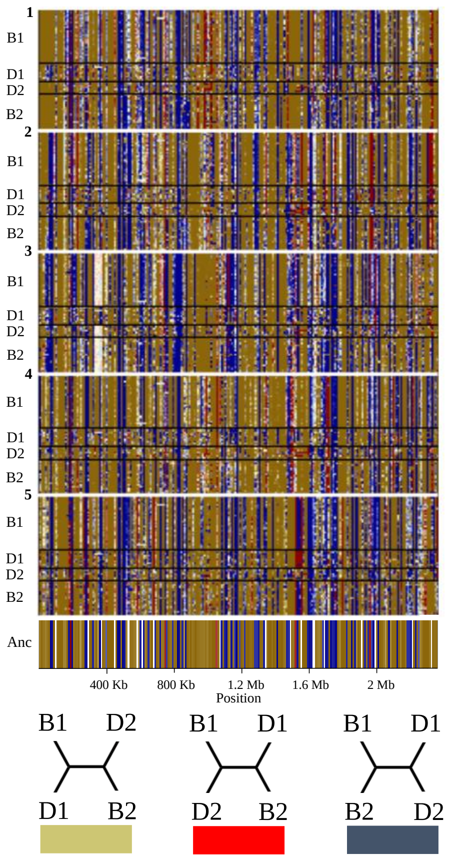

A, B1, E, D1
A, B1, D2, G
A, B1, E, B2
A, B1, E, D2
1
2
3
4
5
A
B1
E
D1
D2
G
B2
Anc
400 Kb
800 Kb
1.2 Mb
1.6 Mb
2 Mb
Position
A E B1 D1
A B1 E D1
A B1 D1 E
A D2 B1 G
A B1 D2 G
A B1 G D2
A E B1 B2
A B1 E B2
A B1 B2 E
A E B1 D2
A B1 E D2
A B1 D2 E


A, B1, E, G
1
A
B1
E
G
2
3
4
5
Anc
400 Kb
800 Kb
1.2 Mb
1.6 Mb
2 Mb
Position
A E B1 G
A B1 E G
A B1 G E
A, B1, G, B2
A
B1
G
B2
Anc
A G B1 B2
A B1 G B2
A B1 B2 G
A, D1, D2, B2
A
D1
D2
B2
Anc
A D2 D1 B2
A D1 D2 B2
A D1 B2 D2
A, D1, D2, G
A
D1
D2
G
Anc
A D2 D1 G
A D1 D2 G
A D1 G D2

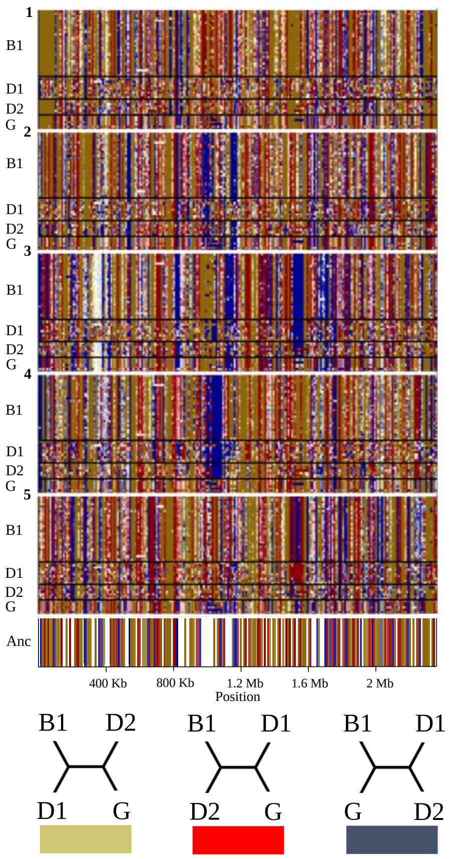

A, D1, G, B2
A, D2, G, B2
A, E, D1, B2
A, E, D1, D2
1
2
3
4
5
A
D1
D2
E
G
B2
Anc
400 Kb
800 Kb
1.2 Mb
1.6 Mb
2 Mb
Position
A G D1 B2
A D1 G B2
A D1 B2 G
A G D2 B2
A D2 G B2
A D2 B2 G
A D1 E B2
A E D1 B2
A E B2 D1
A D1 E D2
A E D1 D2
A E D2 D1


A, E, D1, G
A, E, D2, B2
A, E, D2, G
A, E, G, B2
1
A
E
D1
G
2
3
4
5
D2
B2
Anc
400 Kb
800 Kb
1.2 Mb
1.6 Mb
2 Mb
Position
A D1 E G
A E D1 G
A E G D1
A D2 E B2
A E D2 B2
A E B2 D2
A D2 E G
A E D2 G
A E G D2
A G E B2
A E G B2
A E B2 G

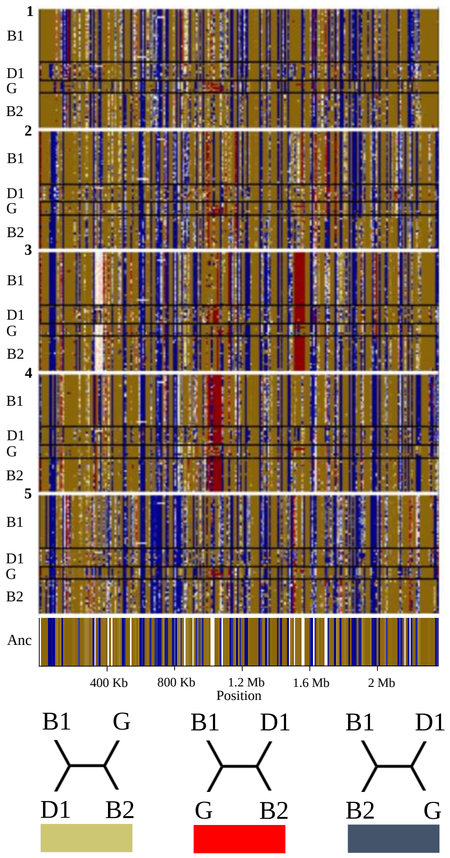
B1, D1, D2, B2



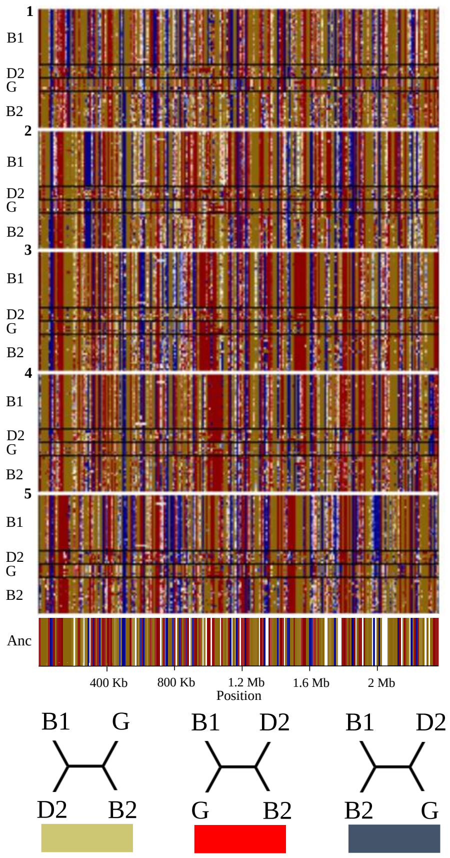
B1, D1, D2, G



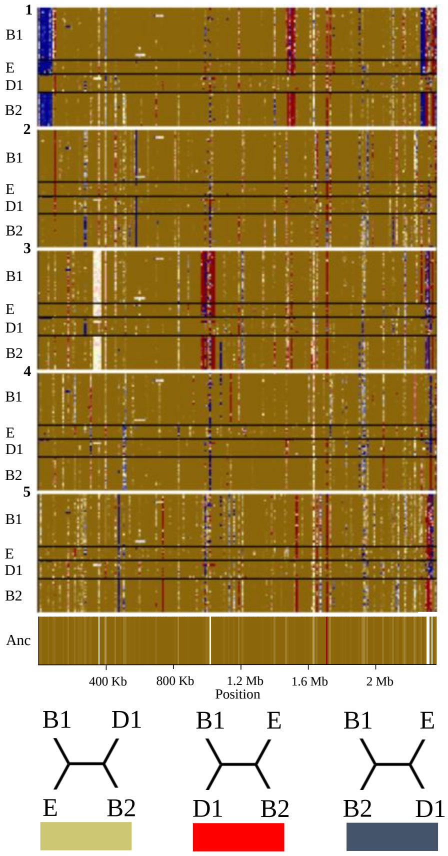
B1, D1, G, B2



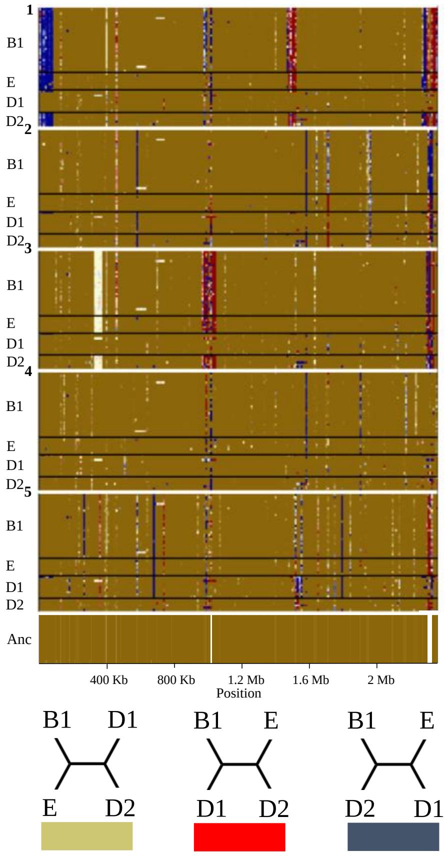
B1, D2, G, B2



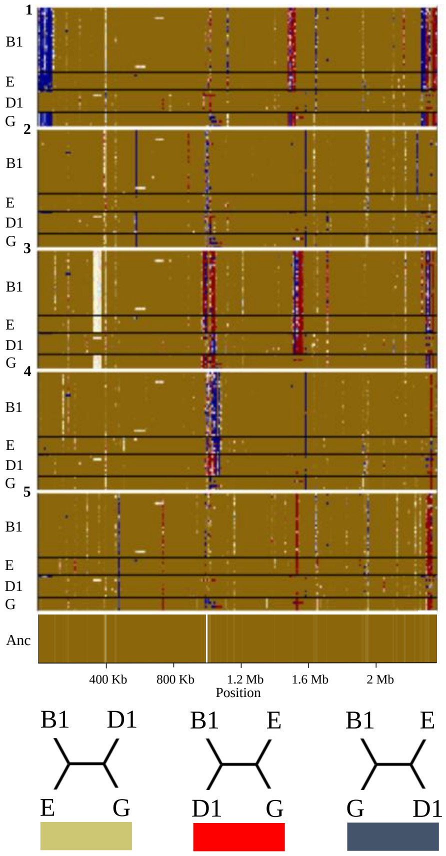
B1, E, D1, B2



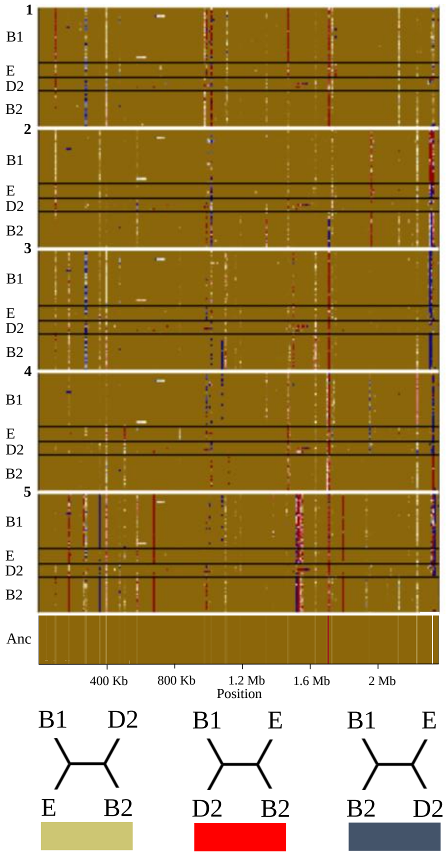
B1, E, D1, D2



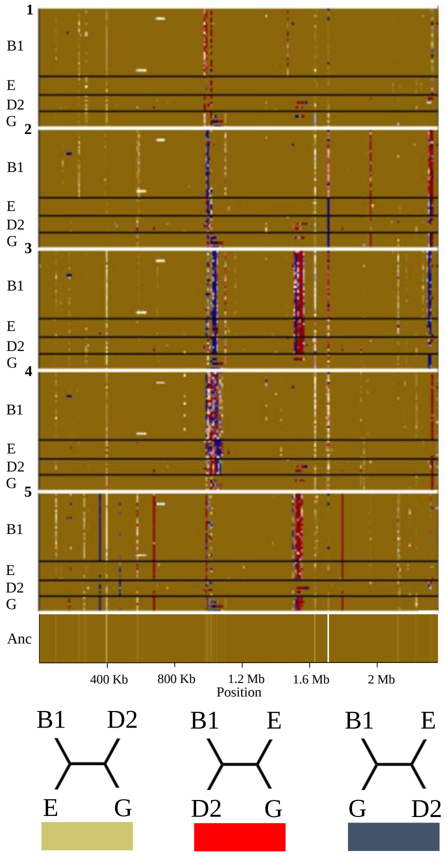
B1, E, D1, G



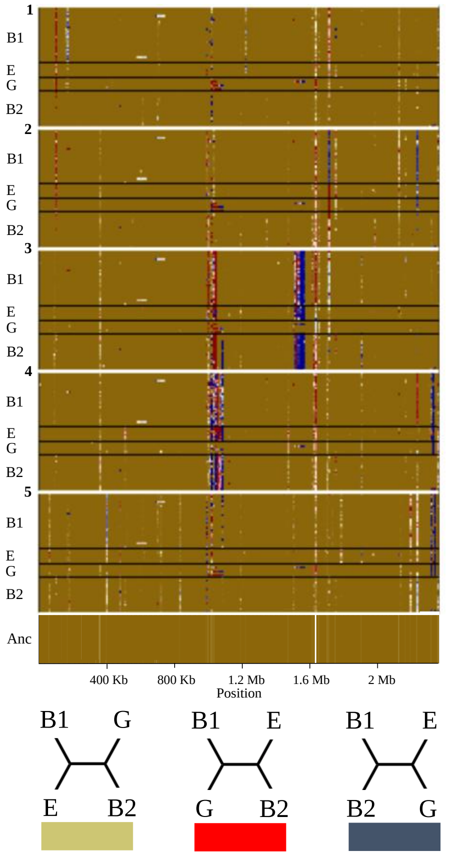
B1, E, D2, B2



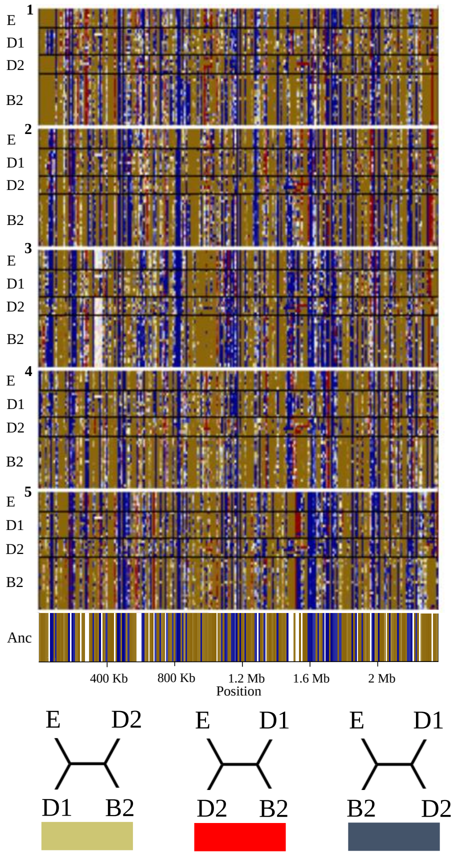
B1, E, D2, G



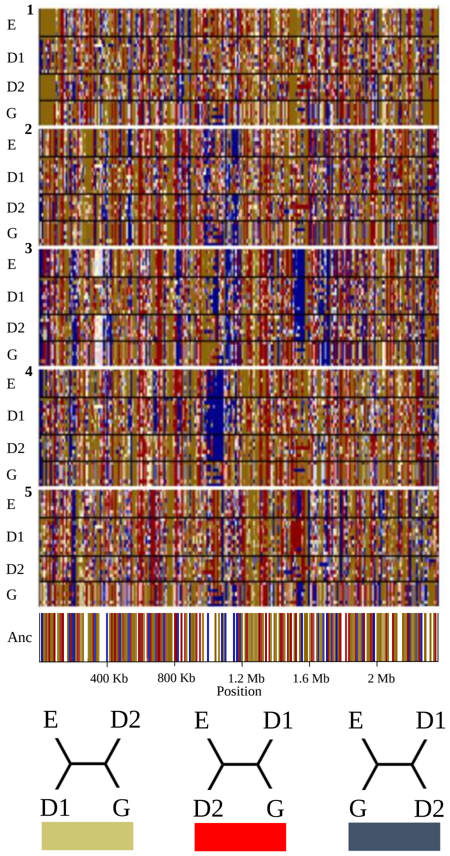
B1, E, G, B2

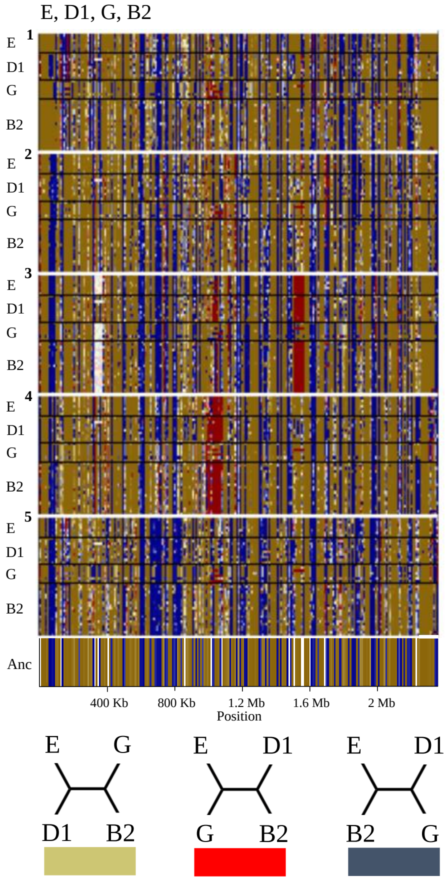


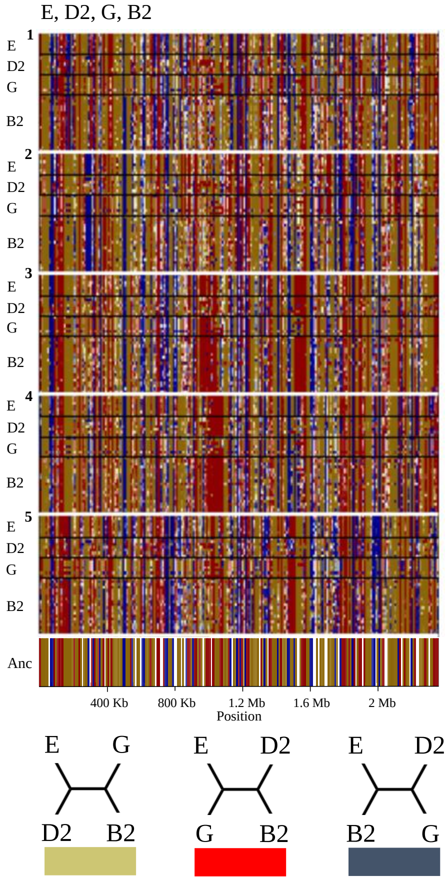
E, D1, D2, B2



E, D1, D2, G




E, D1, G, B2
1
E
D1
G
B2
2
E
D1
G
B2
3
E
D1
G
B2
4
E
D1
G
B2
5
E
D1
G
B2
Anc
400 Kb
800 Kb
1.2 Mb
1.6 Mb
2 Mb
Position
E
G
D1
B2
E
D1
G
B2
E
D1
B2
G


E, D2, G, B2
1
E
D2
G
B2
2
E
D2
G
B2
3
E
D2
G
B2
4
E
D2
G
B2
5
E
D2
G
B2
Anc
400 Kb
800 Kb
1.2 Mb
1.6 Mb
2 Mb
Position
E
G
D2
B2
E
D2
G
B2
E
D2
B2
G

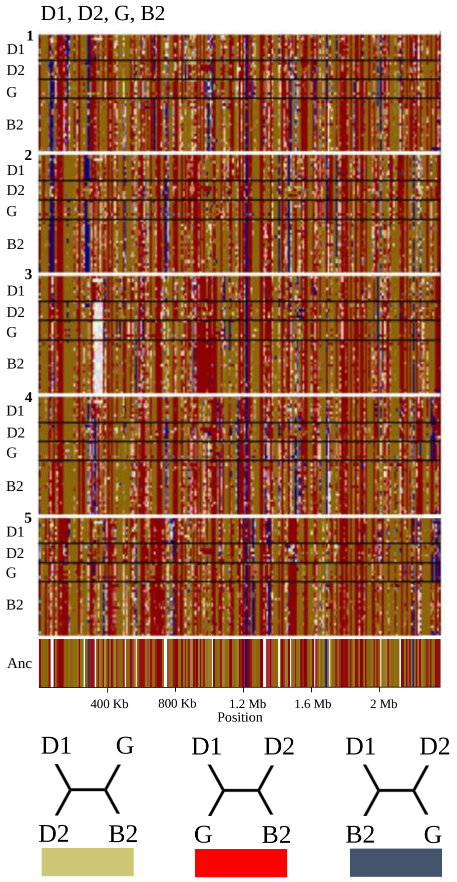

D1, D2, G, B2
1
D1
D2
G
B2
2
D1
D2
G
B2
3
D1
D2
G
B2
4
D1
D2
G
B2
5
D1
D2
G
B2
Anc
400 Kb
800 Kb
1.2 Mb
1.6 Mb
2 Mb
Position
D1
G
D2
B2
D1
D2
G
B2
D1
D2
B2
G

Figure 3.7. Quartet clustering patterns obtained for 10 kb windows across the core genome alignment. Plots are shown for all possible quartet combinations from *E. coli* phylogenetic groups A-G. The x axis represents the position along the core genome alignment (2,350,705 bp), ordered as in strain str. K-12 MG1655. Each row of the figure shows the topologies for a unique quartet of strains, for 235 non-overlapping windows across the core genome alignment. The topologies are coloured in accordance with their relationships as shown in the legend. The rows are separated into sections (separated by white lines), within which a consistent set of reference strains contribute three of each quartet. The fourth member of the quartet is varied in each row, with the subsections (separated by black lines) indicating which of the four taxa is varied. Quartets are coloured only if the middle branch is supported by ≥50% bootstrap support otherwise it is shown as white. The row labelled 'Anc' indicates the clustering pattern deduced to be pre-divergence for the four phylogenetic groups based on a single clustering pattern's presence in >50% of quartets for a section.

The quartet analysis plots showed that the 3 possible phylogenetic group pre-divergence clustering patterns were found to have all occurred across core genes in 14 different quartet analyses (Figure 3.7). Pre-divergence clustering pattern 1 (as depicted in Figure 3.2) was the pattern for 100% of core genes in 16 analyses, ≥ 90% of core genes in 22 analyses, ≥ 70% of core genes in 25 analyses, and ≥ 50% of core genes in all 35 analyses (Table 3.3). This was in comparison to cluster patterns 2 and 3 which were observed in 20-42% of core genes for 7 analyses for cluster pattern 2 for and 8 analyses for cluster pattern 3 (Table 3.3). Clustering pattern 1 was therefore found to be the clonal clustering pattern for *E. coli* phylogenetic groups A-G (Figure 3.8).

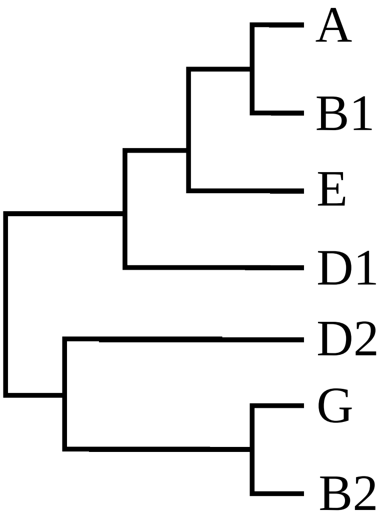


Figure 3.8. Diagrammatic tree indicating the inferred clonal frame clustering pattern for *E. coli* phylogenetic groups A-G.

Table 3.3. The number of core genome genes associated with each of the three pre-divergence clustering patterns in each of the 35 quartet phylogeny analyses. *

	Cluster pattern 1			Cluster pattern 2			Cluster pattern 3		
Quartet analysis	**Quartet phylogeny**	**Genes**	**Perc. genes**	**Quartet phylogeny**	**Genes**	**Perc. genes**	**Quartet phylogeny**	**Genes**	**Perc. genes**
1	(A,B1),(D1,B2)	2,484	100	(A,D1),(B1,B2)	0	0	(A,B2),(B1,D1)	0	0
2	(A,B1),(D1,D2)	2,475	100	(A,D1),(B1,D2)	0	0	(A,D2),(B1,D1)	0	0
3	(A,B1),(D1,G)	2,484	100	(A,D1),(B1,G)	0	0	(A,G),(B1,D1)	0	0
4	(A,B1),(D2,B2)	2,483	100	(A,D2),(B1,B2)	0	0	(A,B2),(B1,D2)	0	0
5	(A,B1),(D2,G)	2,493	100	(A,D2),(B1,G)	0	0	(A,G),(B1,D2)	0	0
6	(A,B1),(E,B2)	2,068	94	(A,E),(B1,B2)	98	4	(A,B2),(B1,E)	43	2
7	(A,B1),(E,D1)	2,040	94	(A,E),(B1,D1)	117	5	(A,D1),(B1,E)	22	1
8	(A,B1),(E,D2)	1,976	94	(A,E),(B1,D2)	105	5	(A,D2),(B1,E)	22	1
9	(A,B1),(E,G)	1,978	91	(A,E),(B1,G)	141	6	(A,G),(B1,E)	60	3
10	(A,B1),(G,B2)	2,493	100	(A,G),(B1,B2)	0	0	(A,B2),(B1,G)	0	0
11	(A,D1),(D2,B2)	1,498	67	(A,D2),(D1,B2)	35	2	(A,B2),(D1,D2)	703	31
12	(A,D1),(D2,G)	1,049	58	(A,D2),(D1,G)	497	27	(A,G),(D1,D2)	264	15
13	(A,D1),(G,B2)	1,522	66	(A,G),(D1,B2)	0	0	(A,B2),(D1,G)	782	34
14	(A,D2),(G,B2)	1,079	51	(A,G),(D2,B2)	613	29	(A,B2),(D2,G)	425	20
15	(A,E),(D1,B2)	2,467	100	(A,D1),(E,B2)	0	0	(A,B2),(E,D1)	0	0
16	(A,E),(D1,D2)	2,475	100	(A,D1),(E,D2)	0	0	(A,D2),(E,D1)	0	0
17	(A,E),(D1,G)	2,493	100	(A,D1),(E,G)	0	0	(A,G),(E,D1)	0	0
18	(A,E),(D2,B2)	2,493	100	(A,D2),(E,B2)	0	0	(A,B2),(E,D2)	0	0
19	(A,E),(D2,G)	2,478	100	(A,D2),(E,G)	0	0	(A,G),(E,D2)	0	0
20	(A,E),(G,B2)	2,493	100	(A,G),(E,B2)	0	0	(A,B2),(E,G)	0	0
21	(B1,D1),(D2,B2)	1,541	70	(B1,D2),(D1,B2)	47	2	(B1,B2),(D1,D2)	627	28
22	(B1,D1),(D2,G)	1,063	55	(B1,D2),(D1,G)	555	29	(B1,G),(D1,D2)	319	16
23	(B1,D1),(G,B2)	1,589	70	(B1,G),(D1,B2)	0	0	(B1,B2),(D1,G)	668	30
24	(B1,D2),(G,B2)	1,077	51	(B1,G),(D2,B2)	663	31	(B1,B2),(D2,G)	374	18
25	(B1,E),(D1,B2)	2,425	99	(B1,D1),(E,B2)	24	1	(B1,B2),(E,D1)	0	0
26	(B1,E),(D1,D2)	2,475	100	(B1,D1),(E,D2)	0	0	(B1,D2),(E,D1)	0	0
27	(B1,E),(D1,G)	2,486	100	(B1,D1),(E,G)	0	0	(B1,G),(E,D1)	0	0
28	(B1,E),(D2,B2)	2,483	100	(B1,D2),(E,B2)	0	0	(B1,B2),(E,D2)	0	0
29	(B1,E),(D2,G)	2,479	100	(B1,D2),(E,G)	0	0	(B1,G),(E,D2)	0	0
30	(B1,E),(G,B2)	2,479	99	(B1,G),(E,B2)	14	1	(B1,B2),(E,G)	0	0
31	(D1,D2),(G,B2)	1,254	57	(D1,G),(D2,B2)	932	42	(D1,B2),(D2,G)	32	1
32	(E,D1),(D2,B2)	1,482	68	(E,D2),(D1,B2)	22	1	(E,B2),(D1,D2)	682	31
33	(E,D1),(D2,G)	1,007	54	(E,D2),(D1,G)	601	32	(E,G),(D1,D2)	265	14
34	(E,D1),(G,B2)	1,599	70	(E,G),(D1,B2)	0	0	(E,B2),(D1,G)	686	30
35	(E,D2),(G,B2)	1,055	52	(E,G),(D2,B2)	561	28	(E,B2),(D2,G)	398	20

* The total number of core genes used in each quartet analysis of the 2,493 genes which made up the core gene alignment is the sum of the values in all 3 'genes' columns for each analysis. The number in the 'genes' column for a given clustering pattern is the number of core genes used in that analysis with quartet phylogeny topologies which exhibited that clustering pattern. The 'percentage genes' value for each clustering pattern for each analysis is the percentage of genes with quartet phylogenies which exhibited that clustering pattern, of the total number of core genes used in that quartet analysis.

To create a clonal frame phylogeny, core gene sequence alignments were extracted from the core gene alignment if they had an inferred history of the clonal frame clustering pattern (clustering pattern 1) across all 35 quartet analyses. This amounted to 250,000 bp of gene sequence alignment from 256 genes, which was then successfully used to construct an *E. coli* core gene phylogeny using RAxML (Figure 3.9). The clustering pattern in this robust phylogenetic analysis was identical to that predicted by the quartet analysis (Figure 3.8).

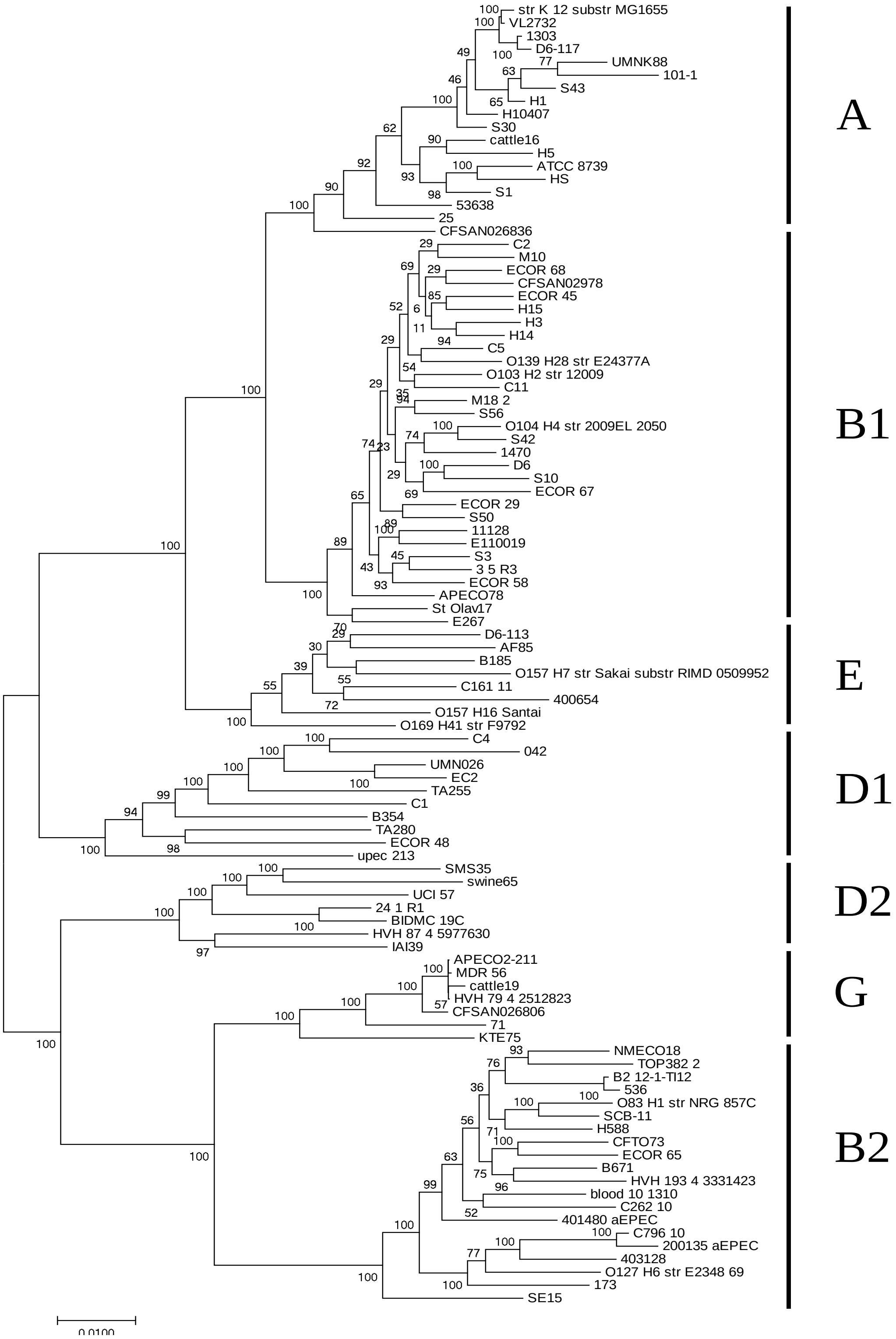


Figure 3.9. A robust maximum likelihood midpoint rooted phylogeny of the 100 *E. coli* strain reference set from phylogenetic groups A-G. The tree was constructed using RaxML from a 250,000 bp alignment of phylogenetically reliable core genome sequence derived from windows from 256 genes which showed a consistent cluster pattern 1 topology across all 35 quartet analyses, the clonal frame clustering pattern. Percentage bootstrap support values are shown on internal branches. The scale bar indicates the number of substitutions per site represented by the branch length shown.

3.3.3. Characterising pre-divergence and post-divergence recombination events between the major phylogenetic groups

To address objective 1, it was necessary to determine the number of *E. coli* core genes with a reported history of both phylogenetic group pre-divergence and post-divergence recombination. This information was taken from the results of the quartet analysis. Clustering pattern 1 represented the proposed clonal frame phylogeny (Figure 3.8), and clustering patterns 2 and 3 (as depicted in Figure 3.2) were evidence of pre-divergence inter-group recombination events affecting core genes. 19 of 35 quartet analyses collectively showed that 46% of core *E. coli* genes exhibited at least some evidence of pre-divergence recombination (figure breakdown not shown, events inferred from genes included in analysis for each quartet analysis for the recombinant patterns 2 and 3 shown in Table 3.3). The difference in recombinant compared to non-recombinant clustering relationships for a given set of four groups can be seen when comparing the former in Table 3.3 (patterns 2 and 3) to the latter in the *E. coli* clonal frame phylogeny (Figure 3.8).

By inspecting the results of the quartet analysis for evidence of post-divergence recombination between phylogenetic groups it was found that 94% of core *E. coli* genes exhibited evidence of post-divergence recombination (cluster patterns 2 and 3, as depicted in Figure 3.2) between strains of particular differing phylogenetic groups. The difference in recombinant compared to non-recombinant clustering relationships for a given set of four groups can be seen when comparing the former in Table 3.4 (patterns 2 and 3) to the latter in the *E. coli* clonal frame phylogeny (Figure 3.8).

Table 3.4. The number of core genes with a reported history of each of the three post-divergence recombination clustering patterns for four given phylogenetic groups in each of all 35 quartet phylogeny analyses. Each row is one quartet analysis. *

	Cluster pattern 1			Cluster pattern 2			Cluster pattern 3		
Quartet analysis	Quartet phylogeny	Genes	Perc. genes	Quartet phylogeny	Genes	Perc. genes	Quartet phylogeny	Genes	Perc. genes
1	(A,B1),(D1,B2)	9	1	(A,D1),(B1,B2)	556	50	(A,B2),(B1,D1)	552	49
2	(A,B1),(D1,D2)	18	2	(A,D1),(B1,D2)	440	45	(A,D2),(B1,D1)	520	53
3	(A,B1),(D1,G)	9	1	(A,D1),(B1,G)	477	52	(A,G),(B1,D1)	438	47
4	(A,B1),(D2,B2)	10	2	(A,D2),(B1,B2)	378	58	(A,B2),(B1,D2)	267	41
5	(A,B1),(D2,G)	0	-	(A,D2),(B1,G)	458	58	(A,G),(B1,D2)	327	42
6	(A,B1),(E,B2)	425	9	(A,E),(B1,B2)	2,064	45	(A,B2),(B1,E)	2,134	46
7	(A,B1),(E,D1)	453	10	(A,E),(B1,D1)	2,090	44	(A,D1),(B1,E)	2,169	46
8	(A,B1),(E,D2)	517	11	(A,E),(B1,D2)	2,058	43	(A,D2),(B1,E)	2,159	46
9	(A,B1),(E,G)	515	11	(A,E),(B1,G)	2,054	44	(A,G),(B1,E)	2,074	45
10	(A,B1),(G,B2)	0	-	(A,G),(B1,B2)	278	44	(A,B2),(B1,G)	358	56
11	(A,D1),(D2,B2)	936	23	(A,D2),(D1,B2)	1,625	40	(A,B2),(D1,D2)	1,519	37
12	(A,D1),(D2,G)	1,398	27	(A,D2),(D1,G)	1,831	35	(A,G),(D1,D2)	2,011	38
13	(A,D1),(G,B2)	913	26	(A,G),(D1,B2)	1,322	37	(A,B2),(D1,G)	1,332	37
14	(A,D2),(G,B2)	1,252	28	(A,G),(D2,B2)	1,649	37	(A,B2),(D2,G)	1,557	35
15	(A,E),(D1,B2)	26	1	(A,D1),(E,B2)	1,056	54	(A,B2),(E,D1)	858	44
16	(A,E),(D1,D2)	18	1	(A,D1),(E,D2)	498	41	(A,D2),(E,D1)	708	58
17	(A,E),(D1,G)	0	-	(A,D1),(E,G)	620	50	(A,G),(E,D1)	628	50
18	(A,E),(D2,B2)	0	-	(A,D2),(E,B2)	482	62	(A,B2),(E,D2)	295	38
19	(A,E),(D2,G)	15	2	(A,D2),(E,G)	499	58	(A,G),(E,D2)	343	40
20	(A,E),(G,B2)	0	-	(A,G),(E,B2)	438	55	(A,B2),(E,G)	356	45
21	(B1,D1),(D2,B2)	903	22	(B1,D2),(D1,B2)	1,588	39	(B1,B2),(D1,D2)	1,552	38
22	(B1,D1),(D2,G)	1,379	26	(B1,D2),(D1,G)	1,835	35	(B1,G),(D1,D2)	2,035	39
23	(B1,D1),(G,B2)	834	23	(B1,G),(D1,B2)	1,341	37	(B1,B2),(D1,G)	1,443	40
24	(B1,D2),(G,B2)	1,314	30	(B1,G),(D2,B2)	1,587	36	(B1,B2),(D2,G)	1,495	34
25	(B1,E),(D1,B2)	68	3	(B1,D1),(E,B2)	1,025	51	(B1,B2),(E,D1)	901	45
26	(B1,E),(D1,D2)	18	1	(B1,D1),(E,D2)	692	50	(B1,D2),(E,D1)	676	49
27	(B1,E),(D1,G)	7	1	(B1,D1),(E,G)	587	48	(B1,G),(E,D1)	618	51
28	(B1,E),(D2,B2)	10	1	(B1,D2),(E,B2)	525	59	(B1,B2),(E,D2)	355	40
29	(B1,E),(D2,G)	14	1	(B1,D2),(E,G)	515	53	(B1,G),(E,D2)	450	46
30	(B1,E),(G,B2)	14	2	(B1,G),(E,B2)	400	50	(B1,B2),(E,G)	383	48
31	(D1,D2),(G,B2)	1,132	27	(D1,G),(D2,B2)	1,429	35	(D1,B2),(D2,G)	1,575	38
32	(E,D1),(D2,B2)	925	23	(E,D2),(D1,B2)	1,500	38	(E,B2),(D1,D2)	1,565	39
33	(E,D1),(D2,G)	1,451	27	(E,D2),(D1,G)	1,805	34	(E,G),(D1,D2)	2,106	39
34	(E,D1),(G,B2)	830	23	(E,G),(D1,B2)	1,356	37	(E,B2),(D1,G)	1,446	40
35	(E,D2),(G,B2)	1,322	29	(E,G),(D2,B2)	1,715	37	(E,B2),(D2,G)	1,547	34

* The total number of core genes used in each quartet analysis of the 2,493 genes which made up the core gene alignment is the sum of the values in all 3 ‘genes’ columns for each analysis. The number in the ‘genes’ column for a given clustering pattern is the number of core genes used in that analysis with quartet phylogeny topologies which exhibited that clustering pattern. The ‘percentage genes’ value for each clustering pattern for each analysis is the percentage of genes with quartet phylogenies which exhibited that clustering pattern, of the total number of core genes used in that quartet analysis.

3.3.4. Determining the proportion of each *E. coli* major phylogenetic group's pan genome that is shared with that of each other group

To address objective 2, it was necessary to determine the number of genes in the pan genome of each group, and the number of core and accessory genes unique to each group or shared with other groups, by post-processing the results of the pan genome analysis carried out by Roary using a written in-house program. When *E. coli* core genes conserved across all strains of all phylogenetic groups were deducted, this analysis showed that group B1 had the largest group pan genome (14,221 genes) and group G the smallest (6,218 genes). The same groups also exhibited the fewest and most group core genes of all groups (339 and 1,132 for B1 and G respectively). Groups E, D1, and D2 exhibited similar number of group core genes (762, 748, and 775 respectively). Groups A, B1, and D1 did not exhibit any group-specific core genes, while groups B2 and G exhibited 19 and 16 respectively, and groups E and D2 each exhibited 8. Groups A, B1, and B2 exhibited the largest number of group accessory genes (9,330, 13,882, and 9,684 respectively) and group G exhibited the fewest across groups (5,086). Groups A, B1, and B2 also exhibited the greatest number of group-specific accessory genes (3,241, 5,414, and 3,347 respectively) and group D2 the fewest (1,227) (Table 3.5).

Table 3.5. The number of pan, core, and accessory genes present in each *E. coli* phylogenetic group A-G.

Category (PG = per genome)	**A**	**B1**	**E**	**D1**	**D2**	**G**	**B2**
Number of genomes in group	18	30	8	10	7	7	20
Group pan-genome size*	9,791	14,221	7,456	7,292	6,501	6,218	10,247
Number of group core genes *	461	339	762	748	775	1,132	536
- Of which unique to group	0	0	8	0	8	16	19
- Of which present in other groups	461	339	754	748	767	1,116	517
Number of group accessory genes *	9,330	13,882	6,694	6,544	5,727	5,086	9,684
- Of which unique to group	3241	5,414	1,683	1,556	1,227	1,329	3,347
- Of which present in other groups	6,089	8,468	5,011	4,988	4,500	3,757	6,337

* Note that *E. coli* core genes conserved across all strains of all phylogenetic groups are not included in these figures.

When investigating the proportion of group core genes (genes core to the group and core, non-core, or absent in other groups) shared between pairs of phylogenetic groups (Table 3.6), specific results can be highlighted. Group A was found to share the highest percentage of its core genes with groups B1 and E (99.57 and 94.36) and the lowest percentage with groups G and B2 (82 and 82.86). The relative greater proportion of genes shared between groups A, B1, and E compared to between groups A, G and B2 is consistent with groups A clustering more closely with groups B1 and E compared to G and B2 in the inferred clonal frame phylogeny (Figure 3.8). This consistency of shared group core genes with clonal frame phylogeny was also shown for other groups. Groups B1, E, D1, D2, and B2 shared most group core genes with groups they are closely related to and shared to fewest with more distantly related groups (Table 3.6, Figure 3.8). Of note was that group B1 shared the most group core genes with groups A and E (99.57 and 94.36). Group E shared the greatest percentage of its group core genes with groups B1 and D1 (94.49 and 96.59) and the lowest percentage with groups G and B2 (86.48 and 85.43). Group D1 shared the greatest percentage of its group core genes with groups E and D2 (96.52 and 97.53) and the lowest percentage with groups G and B2 (88.1 and 87.48). Group D2 shared the greatest percentage of its group core genes with groups D1 and E (96.26 and 90.06) and the lowest percentage with groups A and G (86.84 and 85.29). Group B2 shared the greatest percentage of its group core genes with groups D2, D1, and G (92.9, 92.72, and 92.72) and the lowest percentage with groups A and B1 (84.19 and 86.86) across groups. However, group G shared the greatest percentage of its group core genes with groups D1 and B2 (92.49 and 92.4) despite clustering closer to group D2 than D1, and the lowest percentage with groups A and E (84.72 and 84.19).

Table 3.6. The percentage of group core genes of each *E. coli* phylogenetic group which are present in other groups.

Group	**Core genes ***	**A** (18)	**B1** (30)	**E** (8)	**D1** (10)	**D2** (7)	**G** (7)	**B2** (20)
A	461	-	99.57	94.36	94.14	89.59	82	82.86
B1	339	100	-	93.51	91.74	89.68	76.4	76.99
E	762	93.44	94.49	-	96.59	93.83	86.48	85.43
D1	748	90.51	91.04	96.52	-	97.73	88.1	85.43
D2	775	86.84	88.77	90.06	96.26	-	85.29	87.48
G	1132	84.72	88.25	84.19	92.49	89.75	-	92.4
B2	536	84.19	86.86	88.28	92.72	92.9	92.72	-

* *E. coli* core genes conserved across all strains of all phylogenetic groups are not included in these figures.

When the proportion of group accessory genes shared between pairs of phylogenetic groups were determined (Table 3.7), specific results could also be highlighted. Unlike with core genes, the pattern of shared accessory genes was not consistent with the *E. coli* phylogenetic group relationships in the clonal frame phylogeny (Figure 3.10). Group A shared the largest percentage of its group accessory genes with groups B1 and B2 (51.67 and 38.47) despite clustering closer to E and D1 than B2. Group B1 shared the greatest percentage of its group accessory genes with groups A and B2 (35.59% and 33.94%) despite clustering closer to E and D1 than B2. Group E shared the greatest percentage of its group accessory genes with groups B1 and B2 (59.4% and 44.5%) despite clustering closer to groups A, and E than B2. Group D2 shared the greatest percentage of its group accessory genes with groups B1 and B2 (54.48% and 50.81%) despite clustering closer to G than B1, and the lowest percentage with groups E and G (40.72% and 40.02%) despite clustering in a separate clade to groups A B1 and D1 and adjacently to group G. Group G shares the greatest percentage of its group accessory genes with groups B1 and B2 (53.99% and 43.26%) despite clustering closer to group D2 than B1. Group B2 shared the greatest percentage of its accessory genes with groups A and B1 (36.11% and 46.3%) despite clustering adjacently to groups D2 and G and the lowest percentage with groups D2 and G (31.65% and 28.13%) despite clustering adjacently to them.

Table 3.7. The percentage of accessory genes present within strains of each phylogenetic group which are also present in other groups.

Group	**Accessory genes ***	**A** (18)	**B1** (30)	**E** (8)	**D1** (10)	**D2** (7)	**G** (7)	**B2** (20)
A	9,330	-	51.67	35.1	34.23	31.18	27.78	38.47
B1	13,882	35.59	-	31.55	29.08	25.24	25.11	33.94
E	6,694	44.79	59.41	-	41.69	34.58	33.43	44.5
D1	6,544	45.09	56.04	42.86	-	41.47	35.36	49.8
D2	5,727	46.25	54.48	40.72	47.13	-	40.02	50.81
G	5,086	39.54	53.99	38.22	37.87	38.08	-	43.26
B2	9,684	36.11	46.3	32.35	34.86	31.65	28.13	-

* Values are the number of accessory genes present across strains of each given phylogenetic group.

3.3.5. Determining genes unique and highly enriched (≥ 90% presence) in *E. coli* phylogenetic groups

84 genes were found to be unique to specific phylogenetic groups or groups of phylogenetic groups and present in at least 90% of group members (Figure 3.10, Table 3.8). Group B1-specific genes included genes annotated as encoding a pilus protein, an ABC transporter, and an ATP-binding transporter. Group E had 8 group-specific genes, 7 of which encoded proteins with environmental interaction functions located over three gene clusters (E-1, E-2, E-3) (here defined as sets of genes within 10 kb of each other). Group D1 had 8 group-specific genes, 2 of which encoded proteins of metabolism-related function. Group G exhibited 15 group-specific genes, 7 of which encoded proteins with environmental interaction functions located on 2 gene clusters (G-1, G-2). Group B2 exhibited 25 group-specific genes, including 13 which encoded proteins with metabolic-associated functions and 5 which encoded proteins with membrane transport-associated functions, all located across 4 gene clusters (B2-1, B2-2, B2-3, B2-4). Groups A+B1 together exhibited 5 specific genes, 2 of which encoded proteins with membrane transport-associated functions. Groups A+B1+E together exhibited 8 specific genes, 3 of which encoded proteins of metabolic-associated functions and 2 encoded proteins with environmental interaction-associated

functions. Groups A+B1+E+D1 exhibited 5 specific genes, one of which encoded a protein with metabolism associated function and another with membrane transport-associated function. Groups G+B2 exhibited 6 specific genes, 2 of which encoded proteins with metabolism-associated functions and 2 encoding proteins with regulation-associated functions.

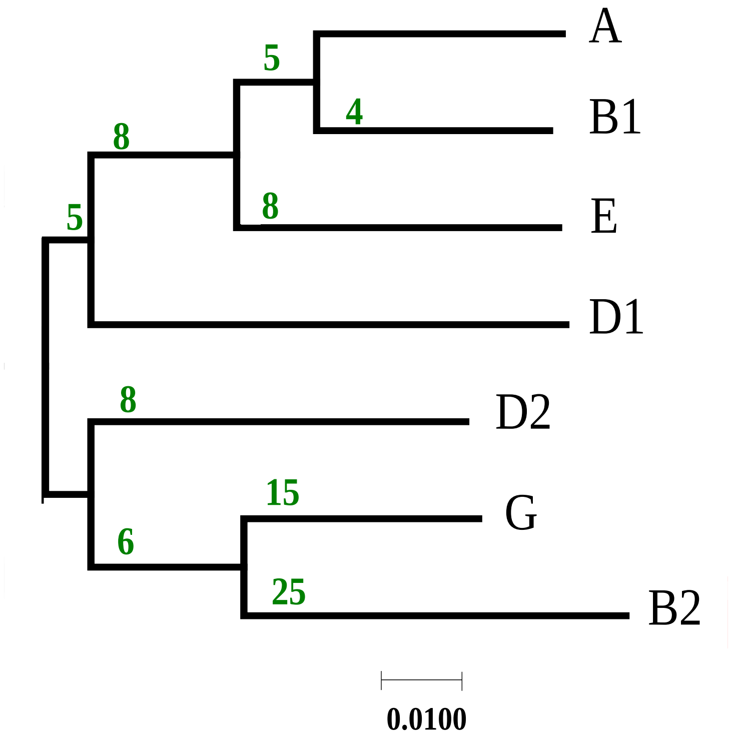


Figure 3.10. The *E. coli* clonal frame phylogeny with the number of genes which are unique to and present in ≥90% of strains for each phylogenetic group, superimposed onto that groups' pre-divergence lineage. Presence of the genes in that lineage alone is indicated with the number in green at the base of the lineage. The scale bar indicates the number of substitutions per site represented by the branch length shown.

Table 3.8. Genes found to be unique to specific phylogenetic groups or groups of phylogenetic groups and present in at least 90% of group members. *

Group	Reference strain	Reference locus ID	Cluster	Annotation	Functional group
B1	O104:H4 str. 2009EL-2050	O3M_03525	-	CblD like pilus biogenesis initiator	Env. interaction
		O3M_03750	-	Uncharacterised protein	-
		O3M_11160	-	ABC transporter, ATP-binding protein	Memb. transport
		O3M_14800	-	ATP-binding transport component	Metabolic
E	O157:H7 str. Sakai substr RIMD O509952	ECs0140	E-1	Fimbrial protein	Env. interaction
		ECs0146		methyldihydropteridine diphosphokinase	Metabolic
		ECs3220	E-2	Probable fimbrial chaperone protein papD	Env. interaction
		ECs3221		Fimbrial usher	Env. interaction
		ECs3222		Type 1 fimbiral protein	Env. interaction
		ECs4426	E-3	Probable fimbrial subunit LpfE	Env. interaction
		ECs4430		Probable fimbrial chaperone LpfB	Env. interaction
		ECs4431		Probable major fimbrial subunit LpfA	Env. interaction
D2	SMS35	EcSMS35_2123	-	Uncharacterised protein YccE	-
		EcSMS35_2178	-	Fimbrial-like adhesin protein	Env. interaction
		EcSMS35_2489	D2-1	ATP-binding protein	Metabolic
		EcSMS35_2490		TIGR02646 family protein	Metabolic
		EcSMS35_2659	-	RatA-like protein	-
		EcSMS35_3798	-	Uncharacterised protein YhiS	-
		EcSMS35_4085	-	Probable type III effector protein	Memb. transport
		EcSMS35_4475	-	Shikimate 5-dehydrogenase	Metabolic
G	MDR 56	B1200_03710	-	Uncharacterised protein	-
		B1200_04105	-	Uncharacterised protein	-
		B1200_12150	-	Uncharacterised protein	-
		B1200_18530	-	Uncharacterised protein	-
		B1200_24130	G-1	DUF2544 domain-containing protein	-
		B1200_24135		Fimbrial protein YfcP	Env. interaction
		B1200_24140		Fimbrial protein	Env. interaction
		B1200_24145		Fimbrial protein YfcR	Env. interaction
		B1200_24150		Fimbrial periplasmic chaperone protein	Env. interaction
		B1200_24155		Fimbrial assembly usher protein	Env. interaction
		B1200_24160		Fimbrial-like adhesin protein	Env. interaction
		B1200_26610	G-2	Rhs Vgr Type IV secretion protein	Env. interaction
		B1200_26615		N-acetylmuramoyl-L-alanine amidase	Metabolic
		B1200_26620		Uncharacterised protein	-
		B1200_26645	-	Uncharacterised protein	-

* Genes found in groups B1, E, D2, G, B2, A+B1, A+B1+E, A+B1+E+D1, and G+B2 are shown. For each gene, details of a reference strain, its locus ID, the name of the associated gene cluster (if any), the gene annotation, and the functional category of the encoded protein is shown.

Table 3.8 continued.

Group	Reference strain	Reference locus ID	Cluster	Annotation	Functional group
B2	O127:H6 str. E2348/69	E2348C_0562	B2-1	Citrate transporter family protein	Memb. transport
		E2348C_0563		Antitoxin family protein	Metabolic
		E2348C_0564		Glycosyl hydrolase	Metabolic
		E2348C_0566		Dihydrodipicolinate synthase	Metabolic
		E2348C_0567		Iron-containing alcohol dehydrogenase	Metabolic
		E2348C_0568		Inner membrane protein	Memb. structure
		E2348C_0569		Probable pyridoxine phosphate biosynthetic protein	Metabolic
		E2348C_0570		DeoR family transcriptional regulator	Regulation
		E2348C_1924	-	Uncharacterized protein YeaR	-
		E2348C_3671	-	Uncharacterised protein	-
		E2348C_3834	-	Probable decarboxylase	Metabolic
		E2348C_4317	B2-2	Malate synthase A	Metabolic
		E2348C_4322		Uncharacterised protein	-
		E2348C_4370	B2-3	2-oxoglutarate dehydrogenase E1 component	Metabolic
		E2348C_4371		succinyltransferase of 2-oxoglutarate dehydrogenase	Metabolic
		E2348C_4372		Dihydrolipoyl dehydrogenase	Metabolic
		E2348C_4373		Succinate--CoA ligase subunit beta , *sucC*	Metabolic
		E2348C_4374		Succinate--CoA ligase subunit alpha, *sucD*	Metabolic
		E2348C_4375		DASS family sodium-coupled anion symporter	Memb. transport
		E2348C_4376		L-lactate dehydrogenase	Metabolic
		E2348C_4377		C4-dicarboxylate transcriptional regulatory protein	Regulation
		E2348C_4378		Sensory histidine kinase in two-component system	Signalling
		E2348C_4402	B2-4	Probable ABC transporter family protein	Memb. transport
		E2348C_4404		Putative dipeptide/nickel transporterYddQ	Memb. transport
		E2348C_4405		ABC transporter membrane permease	Memb. transport
A+B1	str. K-12 subtr. MG1655	b1196	-	Uncharacterised protein	-
		b3715	-	6-phosphogluconate phosphatase YieH	Metabolic
		b4038	-	Probable type III effector protein	Memb. transport
		b4555	-	Uncharacterised protein YicS	-
		b4661	-	Outer membrane usher protein	Memb. transport
A+B1 +E	str. K-12 subtr. MG1655	b0608	-	Predicted oxidoreductase, Zn-dependent	Metabolic
		b1537	-	Nicotinamide-nucleotide amidohydrolase PncC	Metabolic
		b1877	-	PF07007 family protein	-
		b2824	-	DUF2509 domain-containing protein	-
		b3143	AB1E-1	Periplasmic pilin chaperone	Env. interaction
		b3145		Fimbrial protein	Env. interaction
		b3890	-	Antitoxin component, ribbon-helix-helix fold protein	Metabolic
		b4031	-	D-xylose transporter XylE	Memb. transport
A+B1 +E+D1	str. K-12 subtr. MG1655	b1001	-	Uncharacterised protein	-
		b1202	-	Autotransporter outer membrane beta-barrel protein	Memb. transport
		b1465	-	Nitrate reductase A subunit gamma	Metabolic
		b3516	-	YccE family protein	-
		b4045	-	UPF0337 protein YjbJ	-
G+B2	O127:H6 str. E2348/69	E2348C_0022	-	NhaR DNA-binding transcriptional activator	Regulation
		E2348C_2412	-	GNAT family N-acetyltransferase	Metabolic
		E2348C_4027	-	Inner membrane protein CbrB	Memb. structure
		E2348C_4248	-	Cytoplasmic protein	-
		E2348C_4528	-	Type II toxin-antitoxin system HipA family toxin	Metabolic
		E2348C_4631	-	Anti-adapter protein IraD	Regulation

3.4. Discussion

The purpose of the work presented in this chapter was to make use of *E. coli* genome sequence data to produce an up-to-date narrative describing *E. coli* evolution. The narrative was to be based on the results of addressing a hypothesis, an overall aim, and 3 objectives using a computer with up-to-date processing capabilities for genomic analysis.

Before main analyses were started it was necessary to determine the appropriate identity cut-off for use in a BLAST analysis to identify *E. coli* orthologues of the same gene. As reported, it was determined to be 95%. However, pan genome analysis of the reference set of 100 *E. coli* genomes using a lower value of 80% produced a comparable core gene phylogeny with an identical topology to that based on 95% identity (results not shown). There was also no difference in the reported results in the quartet analysis, results regarding inferred history of recombination, the proportion of shared genes between phylogenetic groups after rounding of numbers, and the specific genes which were unique and core to 90% of strain genomes in certain phylogenetic groups or groups of phylogenetic groups (results not shown). This indicated that the results of these analyses were robust to the slightly arbitrary choice of identity cut-off. However, for the results presented here the cut-off of 95% was applied for orthologue identification, because this value was more conservative and likely to have reduced the number of cases of false orthologue pairs being reported.

Creation of a set of *E. coli* strain genomes which represented full phylogenetic and sample diversity for use in analyses was approached by using filtering the available 5,623 strain genomes down to 100. After filtering for genome sequence assembly quality, phylogenetic diversity was maximised in the final set by selecting strain genomes which were separated by relative long phylogenetic branches as observed in a core gene phylogeny. Sample diversity was maximised in the final set by selecting genomes which represented the twelve major *E. coli* pathovars, together with commensals (human and varied non-human sources), laboratory strains, and environmental isolates from soil and water. This filtering process made it likely that the size of the *E. coli* pan genome was maximized for the 100 genome set and included *E. coli* pan genome genes which collectively encoded the widest possible range of functions. Creating this strain set of 100 addressed the first part of the principal aim, which was to create an up-to-date *E. coli* strain set representative of the available phylogenetic and sample diversity.

Genomes from cryptic clade groups C-I, C-III, C-IV, and C-V were not sufficiently abundant to represent their groups in the evolutionary studies which were carried out in this work. However, it is likely that in future the sequencing of more diverse samples will allow the creation of a phylogenetically diverse genome set for the cryptic clades.

Given the data produced in this chapter it is possible to propose an up to-date narrative of *E. coli* evolutionary history. After divergence away from the ancestor of the cryptic clade *E. coli*, it can be proposed that the first lineage subsequently diverged into group D2 and a lineage which later diverged into groups G and B2. The second lineage diverged into the outgroup D1, followed by group E, leaving a single lineage which diverged into groups A and B1 (Figure 3.9, see Figure 3.12 For the proposed *E. coli* clonal frame clustering pattern). Through creating a representative *E. coli* strain set and using it to construct this *E. coli* clonal frame phylogeny, the principal aim was addressed.

It can be inferred that the lineage ancestor genome of G+B2 contained 6 genes encoding proteins linked to environmental stress tolerance and increased adaptation to stress which were absent in other *E. coli* lineage ancestor genomes at the time of their respective diversifications. The presence of these genes, either gained in that lineage or more ancestrally, might have been the cause for divergence away from the ancestor of groups A+B1+E+D1. These included a gene encoding the type II toxin-antitoxin system HipA protein that may be involved in an altruistic cell death mechanism or DNA sequence stabilisation and stress tolerance processes (Schumacher et al. 2009). The G+B2 ancestor also encoded the anti-adapter protein IraD which has been found to work to increase the stability of the stigma stress factor RpoS during oxidative stress to contribute to overall increased oxidative stress resistance (Bougdour et al. 2008). The adaptations provided by these genes may have contributed to the divergence of the lineage ancestor of groups G and B2. The divergence of the group G ancestor from the group B2 ancestor may then have occurred as a result of the acquisition of 15 genes with metabolic, environmental interaction, or pathogenic functions. These genes were absent in other ancestral lineages at the time of their respective divergences. They included 6 fimbrial-function encoding genes on a 7-gene cluster which imply presence of a unique fimbrial system. The system can be proposed to have collectively provided an environmental motility or host colonisation function in the lineage ancestor, based on previous functional characterisation work on *E. coli* fimbriae (Korea et al. 2010). The lineage ancestor of group B2 was found to have exhibited the most of any phylogenetic group pre-divergence lineage; 25 genes, including four gene clusters (cluster B2-1, B2-2, B2-3, B2-4) which were absent in other ancestral lineages at the time of their respective divergences. These included 13 genes with metabolic-encoding functions including s*ucCD* homologues involved in the tricarboxylic acid pathway, and metabolic functions. It can be speculated that these genes provided the ancestor of group B2 with a suite of novel metabolic capabilities which likely allowed it to diversify itself significantly away from other *E. coli* in terms of environmental interaction, metabolism, cell structure, regulation, and habitat usage.

The ancestor of A+B1+E+D1 can be inferred to have exhibited 5 genes which contributed to diversification from the ancestor of groups D2+G+B2 and were absent in other ancestral lineages at the time of their respective divergences. One gene encodes a nitrate reductase A subunit gamma protein, which is a membrane-bound anaerobic respiratory enzyme which is used to process nitrate to nitrite and generate energy (MacGregor 1974). It is plausible that this gene provided novel nitrate-based energy yield metabolic capabilities, which along with other genes contributed to lineage diversification. The diversification of the ancestor of lineage of A+B1+E away from the ancestral lineage for group D1 lineage can be speculated to have occurred as a result of the presence of 8 genes unique to the lineage and absent in other lineage ancestors at the time of their respective divergences. These included a periplasmic pilin chaperone and fimbrial protein which likely enhanced *E. coli* to bacteria or host cell contact to enhance survival and allowed exchange of DNA in the case of the former (Giron et al. 1991), and a D-xylose transporter protein which enable the accumulation of sugar against a concentration gradient (Sun et al. 2012); a potentially crucial adaptation for surviving when environmental sugar resources are lower than within the cell.

The divergence of group E away from the lineage ancestor of groups A+B1 can be proposed to be due in part to the presence of 7 fimbrial function-encoding genes which were absent in other ancestral lineages at the time of their respective divergences. Fimbriae have been implicated as important for environmental motility and host infection in numerous studies and specifically the *lpf*ABE genes which include these have previously been found to enable interaction with eukaryotic host cell through assisting in microcolony formation (Torres et al. 2002). It can be suggested that group E diverged in order to occupy previously uncolonized habitat in the form of animal host cells most through use of unique fimbriae structures not present in related *E. coli*. Group B1 divergence away from group A can be proposed to have occurred through the presence of 4 genes which that were absent in other ancestral lineages at the time of their respective divergences, and include a gene encoding a pilus biogenesis initiator protein. Pilus systems in *E. coli* have been previously found to increase the frequency of horizontal gene transfer (Marklund et al. 2002) so the ability of the ancestral lineage colony to uptake genes more regularly than other colonies can be speculated to have contributed to this divergence. By inferring which gene gain events are likely have been important contributors to the divergence of each *E. coli* phylogenetic group A-G, the third objective was addressed.

The quartet recombination analyses indicated that both post-divergence recombination (occurring after the divergence of major phylogenetic groups) and pre-divergence recombination (occurring before the divergence of major phylogenetic groups) can be inferred to have played a role in shaping the core gene genetic diversity between phylogenetic groups A-G. 46% of core genes can be hypothesised to have undergone recombination between at least 2 phylogenetic group ancestral lineage strains before they each diversified into their respective phylogenetic groups (pre-divergence recombination). After the divergence of phylogenetic groups (post-divergence recombination), 96% of core genes can be deduced to have undergone recombination between at least 2 strains of different phylogenetic groups. This difference can be speculated to be possibly the result of increased closer physical proximity and clonal expansion of populations, which provided increased opportunity for genetic exchange via recombination between groups in more recent evolutionary history after groups diverged. By producing these findings regarding the prevalence of pre-divergence and post-divergence recombination between phylogenetic group ancestral lineages, the first objective was addressed.

It can be proposed on the basis of analysis that *E. coli* groups exchanged genes via horizontal transfer to the extent where 21%-39% of group accessory genes are now unique to a given phylogenetic group. Groups B2, B1, and A can be inferred to have become the most genetically diverse groups as each were found to exhibit the greatest number of group accessory genes unique to their groups and the largest group pan genomes. In contrast, groups D2 and G diversified least in terms of gaining unique genes and exhibited the smallest pan genomes. This is likely to indicate that groups A, B1, and B2 evolved to survive and grow in number in a wider range of habitats and environments than groups D2 and G. This diversification can be speculated to have been possible because of genes gained in the respective ancestral lineages of phylogenetic groups A+B1, B1, and B2. Obtaining these findings regarding the determined proportion of shared accessory and core genes between strain genomes of phylogenetic groups addressed the second objective.

In summary, the hypothesis, principal aim and objectives were addressed which resulted in the creation of an up-to-date narrative describing the major events occurring during *E. coli* evolution. Non-cryptic clade *E. coli* were found to have diverged into 7 instead of 6 major phylogenetic groups (Figure 3.9). The divergence of 5 of these groups (all but A and D2) was associated with the presence of between 4 and 25 genes with functions associated with metabolism, stress tolerance, and virulence. Since group divergence, groups B2, B1, and A were found to be the most genetically diverse groups with the largest pan genomes indicating these groups exhibit adaptations to a broader range of habitats and environments than the other groups. 46% of core *E. coli* genes can be inferred to have undergone pre-divergence inter-group recombination. This was compared to 96% of genes exhibiting post-divergence inter-group recombination. This difference is possibly the result of increased physical proximity and clonal expansion of colonies, which increased opportunity for genetic exchange via recombination between strains from different evolutionary groups post-divergence. However, testing this hypothesis would require further investigation.

Chapter 4: Novel *E. coli* phylogenetic group assignment methods

4.1. Introduction

Phylogenetic group assignment began with the development of multi-locus sequence typing to study phylogenetic relationships of *Neisseria meningitidis* strains (Maiden et al. 1998). *E. coli* phylogenetic group assignment has become a standard approach for characterising genetic relationships between *E. coli* strains since MLST was first developed (Tenaillon et al. 2010, Maiden et al. 1998). For *E. coli*, the traditional principal methods, which employ multiple sequence loci for phylogenetic group assignment, are multi-locus sequence typing (MLST) (Maiden et al. 1998) (7-15 loci, Table 4.1) and the 4-loci Clermont quadruplex PCR method (referred to as “Clermont PCR” in this thesis; Clermont et al. 2013, Table 4.2). Their use is still commonplace, as demonstrated by recent use of the MLST schemas in studies by Matamouros et al. (2018), Janecko et al. (2018), Carter and Pham (2018), and use of the Clermont multiplex by Cho et al. (2018), Zahara et al. (2018), and Garcia et al. (2018). However, I predicted that the fewer number of loci involved and the moderate likelihood of recombination affecting the loci included in each method means multiplex and 7-15 loci MLST schemas may not reliably provide the same clonal group assignment as a core gene phylogeny provides. This was because a core gene phylogeny provides the most accurate representation of clonal inheritance patterns within a species other than if recombinant sequences are removed prior to phylogeny construction (Tenaillon et al. 2010). This is as with a core gene phylogeny all possible loci are considered which minimises the potential influence of recombination (Tenaillon et al. 2010).

Table 4.1. Details of the four most commonly used *E. coli* MLST schemas.

MLST schema	**Genes**	**Origin**	**Website**
Achtman	*adk, fumC, gyrB, icd, mdh, purA, recA*	Warwick Medical School	http://enterobase.warwick.ac.uk/species/ecoli/ download_7_gene
Pasteur	*dinB, icdA, pabB, polB, putP, trpA, trpB, uidA*	Pasteur Institute	http://bigsdb.pasteur.fr/perl/bigsdb/bigsdb.pl?db=pub mlst_ecoli_seqdef_public&page=downloadAlleles
EcMLST: 7 genes	*aspC, clpX, fadD, icdA, lysP, mdh, uidA*	Michigan State University	http://shigatox.net/ecmlst/cgi-bin/da
EcMLST: 15 genes	*aspC, clpX, fadD, icdA, lysP, mdh, uidA, mtlD, mutS, rpoS, grpE, dnaG, cyaA, arcA, aroE*	Michigan State University	http://shigatox.net/ecmlst/cgi-bin/da

Table 4.2. Details of the Clermont multiplex method, showing the presence/absence of the four loci in each phylogenetic group.

Fragment name	**Length (bp)**	**A**	**B1**	**D1/E**	**D2**	**B2**	**B2**	**E/C-I**	**C-I/ C-III**	**C-III/ C-IV/C-V**
arpA	400	+	+	+	-	-	-	+	-	-
chuA	288	-	-	+	+	+	+	+	-	+ (476bp)
yjaA	211	-	-	-	-	+	-	+	+	-
TspE4.C2	152	-	+	+/-	-	+/-	+	-	-	-

E. coli researchers are increasingly using a phylogeny constructed from all core gene sequences in the form of a core gene MLST (cgMLST) instead of a traditional MLST employing 7-15 reference loci (Maiden et al. 2013) to conduct *E. coli* clonal group assignment. Clonal group assignment using a cgMLST schema can be done using the core genes from a specific set of *E. coli* strain genomes under study to construct a core gene phylogeny (Grönthal et al. 2018, Pietsch et al. 2018, Zhou et al. 2017, Allen et al. 2017). It can also be done through using a set of reference *E. coli* core genes from an established cgMLST schema for public use, such as that offered by the Enterobase database hosted at the Warwick Medical School (http://enterobase.warwick.ac.uk), or for commercial use as is offered by '1928' (2,500 gene cgMLST, https://1928diagnostics.com/product_resources/ escherichia-coli/#) and Ridom (cgMLST based on 3,152 genes from the *E. coli* Sakai strain genome, https://www.cgmlst.org/ ncs/schema/8896773/). Further to this, the Bacterial Isolate Genome Sequence Database (BIGSdb) can be used which combines available cgMLST reference sequences with database bacterial genome sequence data and database information regarding isolate sample, phenotype, and encoded protein functional annotations to provide a BISGdb sequence type (Jolley and Maiden 2015). BIGSdb sequence types can be used to assign *E. coli* clonal groups and membership of a clonal complex like a cgMLST does, but also be used to provide information about likely phenotypes based on BLAST database matches to gene sequences.

The increased use of whole-genome sequence data with cgMLST schemas and BIGSdb is most likely due to decreasing costs of whole-genome sequencing each year (Quinoo et al. 2017). However, the cgMLST employs most or all core reference genes and so some are likely to be included which have a history of recombination. This means existing cgMLST methods cannot be used to reliably place strains phylogenetically (including providing a clonal group and clonal complex assignment). In Chapter 3 it was reported that 256 genes (comprising of 250,000 bp) of core *E. coli* gene sequence were free of recombination between phylogenetic groups and could be used to construct an inferred *E. coli* clonal frame phylogeny using phylogenetic construction. At its current stage of testing based on the work presented in Chapter 3, it can be inferred that these 256 genes would produce 100% *E. coli* clonal group assignment if used as a standalone cgMLST and employ fewer reference genes than cgMLST which are typically used by researchers. This means it would provide greater phylogenetic accuracy and would require a lesser computational time to run due to the reduced number of sequences to analyse, compared to existing standard cgMLSTs. With further testing it can also be determined if the cgMLST could be used to assign clonal complexes at the sub-phylogenetic group level. I predicted that this cgMLST could be used to assign *E. coli* clonal groups reliably either as a standalone cgMLST or implemented as part of BIGSdb to provide the *E. coli* clonal group assignment component of a BIGSdb sequence type.

The purpose of the work in this chapter was to explore evidence to support the preferential use of this cgMLST through addressing a hypothesis, an overall aim, and objectives. Evidence was explored because, as the cgMLST employs a relatively fewer number of recombination-free genes over alternative cgMLST schemas, it is likely to exhibit greater reliability of clonal group assignment compared to the 7-15 loci MLST and multiplex schemas. It also has the potential to be developed into an *in-silico* stand-alone schema or implemented as part of a program such as BIGSdb to be used as an efficient *E. coli* clonal group tool.

4.1.1. Hypothesis

The hypothesis of the work presented in this chapter was designed to support the preferential use of the proposed 256 cgMLST for *E. coli* clonal group assignment through comparing it to novelly created alternative MLST schemas comprised of the same or a fewer number of genes which have a history of both recombination and no recombination:

Created novel alternative MLST schemas must be comprised of a greater number of randomly selected core *E. coli* genes with a history of both no recombination and recombination, than the number of recombination-free *E. coli* core genes used in the cgMLST (256), to reliably achieve 100% correct clonal group assignment of all *E. coli* strain genomes.

4.1.2. Overall aim and objectives

The overall aim was to explore evidence to support the preferential use of the proposed 256 cgMLST for *E. coli* clonal group assignment through addressing the hypothesis. It was to also address the following objective, designed to determine if it is possible to create an MLST and multiplex schema using a limited number of 7-15 recombination-free loci based on those in the proposed cgMLST for use in the place of the proposed cgMLST. These schemas would be designed for stand-alone use when time is limited. This is because, if hundreds or thousands cgMLST sequences (1 per *E. coli* strain genome) require processing, it could take an impractical amount of computational time with most modern computers (Quainoo et al. 2017):

1. To develop a novel *in-silico* 7-15 gene MLST schema and *in-silico* 4-gene novel multiplex schema devised using genes identified as free from recombination in Chapter 3.

4.2. Methods

4.2.1. Obtaining reference gene sequences for an *E. coli* MLST schema

An in-house program was used which identified *E. coli* core gene sequences genes (≥99% presence) in a reference strain genome using a gene presence and absence file generated from a Roary analysis as a guide. The reference genome was a PROKKA-annotated *E. coli* strain K-12 substr. MG166 (GenBank accession U00096) genome sequence. After identification of all *E. coli* core genes present in strain K-12 genome sequence, the program then randomly sorted the genes and printed out a user-specified number (N) of them. These printed N gene sequences were then taken to be the reference gene sequences for an MLST schema of size N reference gene sequences.

4.3. Results

4.3.1. Novel alternative MLST schema analysis

In order to support the use of the proposed cgMLST over alternative MLST schemas comprised of a similar or fewer number of genes, comprising of genes with a history or recombination and no recombination, it was necessary to test the accuracy of such MLST schemas at group assignment. For *E. coli* groups A-G this was clonal group assignment as clonal groups had already been established for the reference set of 100 *E. coli* strains in Chapter 3. Assigning *E. coli* cryptic clade groups C-I, C-III, C-IV, and C-V strains was also tested but for phylogenetic groups instead of clonal groups as no clonal group analysis was carried out in Chapter 3 for the cryptic clade *E. coli*. To do this, a set of 20 genomes from the 4 cryptic clade phylogenetic groups C-I to C-V were defined. They were selected as a representative sample from the 27 genomes identified as belonging to the 4 groups in Chapter 3. This was done firstly by carrying out pan genome analysis with Roary using the 27 strain genomes and constructing a phylogeny using RAxML with the resulting 2.04 Mb core gene alignment (Figure 4.1). Strains which were separated from one another by long branches were then selected for the set of 20 strains (highlighted with a red dot in Figure 4.1, Table 4.3). To observe the phylogenetic placement of the 20 cryptic clade strains in relation to those from groups A-G, a core gene phylogeny was required. This was obtained by conducting pan genome analysis using the 20 strains with the 100 *E. coli* set of strains representing groups A-G defined in Chapter 3. The resulting 1.86 Mb core gene alignment was then used to construct a phylogeny using RAxML (Figure 4.2). Cryptic clade group C-I was shown to cluster as an out group to the remaining strains of groups A-G. A clade comprising groups C-III and C-IV was the next to branch off, followed by group C-V. There was 100% bootstrap support for the basal branches of all 11 phylogenetic groups.

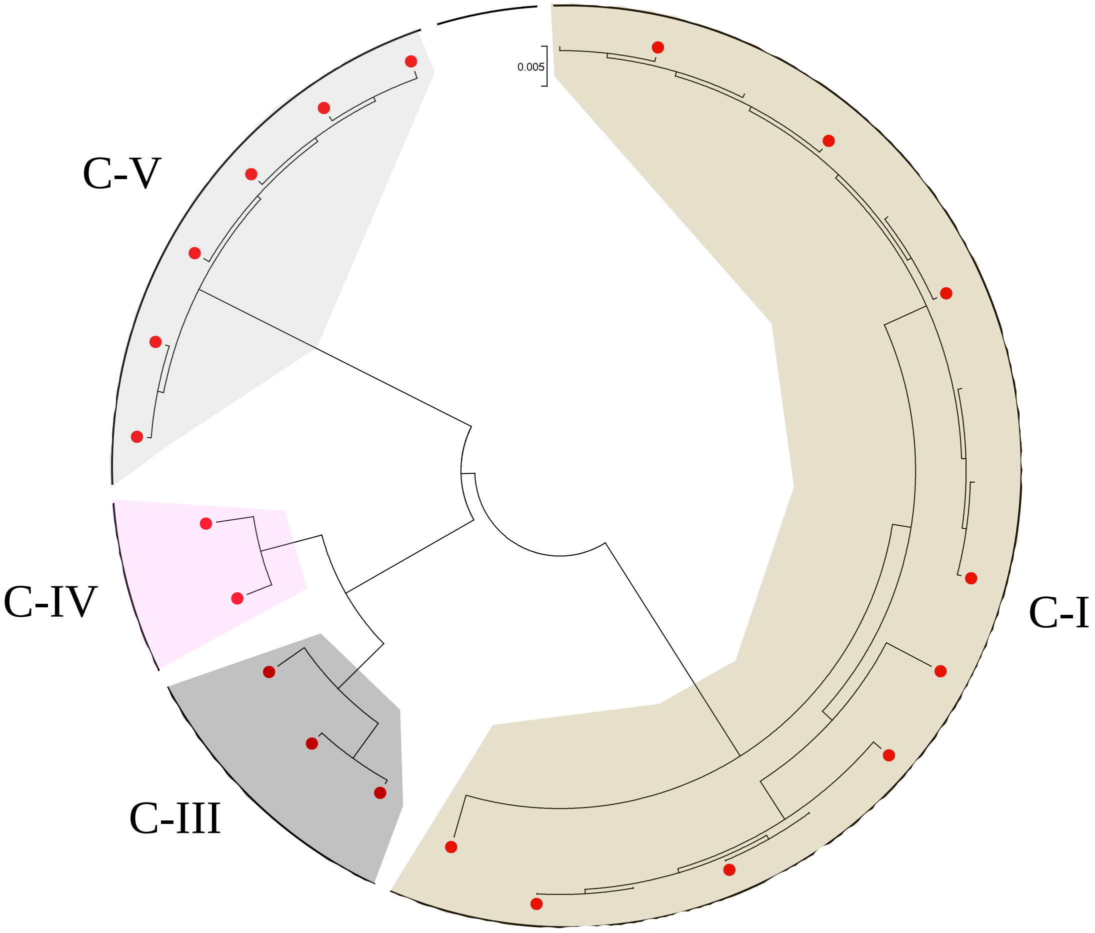
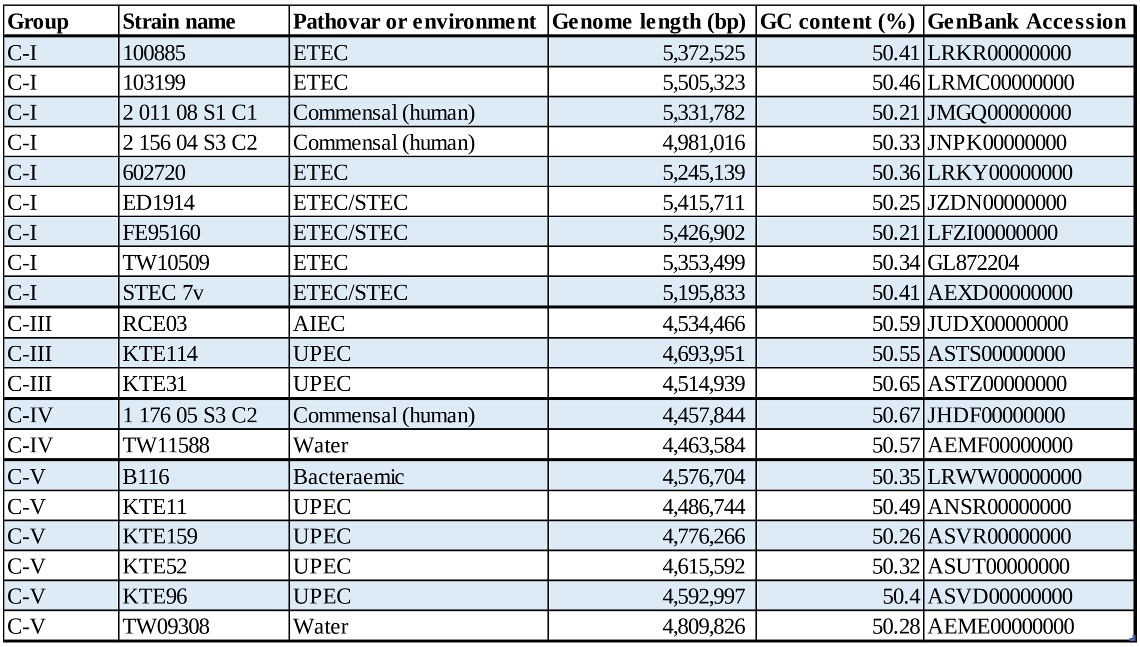


Figure 4.1. Unrooted RAxML maximum likelihood phylogeny constructed using a core gene alignment (2,039,093 bp) from 27 *E. coli* cryptic clade strain genomes obtained from GenBank with an N50 greater than 100,000 bp. Major Phylogenetic groups are labelled C-I, C-III, C-IV, C-V in the outer ring, with gaps in the ring indicating group borders. 20 strains chosen to represent the phylogenetic diversity are labelled with a red dot. The scale bar at the top indicates the number of substitutions per site associated with the indicated branch length.

Table 4.3. Phylogenetic group, strain name, pathovar or environment of isolation, genome length, % GC content, and GenBank accession of the 20 strains chosen for the set of phylogenetically diverse cryptic clade strain representatives.



Group	Strain name	Pathovar or environment	Genome length (bp)	GC content (%)	GenBank Accession
C-I	100885	ETEC	5,372,525	50.41	LRKR00000000
C-I	103199	ETEC	5,505,323	50.46	LRMC00000000
C-I	2 011 08 S1 C1	Commensal (human)	5,331,782	50.21	JMGQ00000000
C-I	2 156 04 S3 C2	Commensal (human)	4,981,016	50.33	JNPK00000000
C-I	602720	ETEC	5,245,139	50.36	LRKY00000000
C-I	ED1914	ETEC/STEC	5,415,711	50.25	JZDN00000000
C-I	FE95160	ETEC/STEC	5,426,902	50.21	LFZI00000000
C-I	TW10509	ETEC	5,353,499	50.34	GL872204
C-I	STEC 7v	ETEC/STEC	5,195,833	50.41	AEXD00000000
C-III	RCE03	AIEC	4,534,466	50.59	JUDX00000000
C-III	KTE114	UPEC	4,693,951	50.55	ASTS00000000
C-III	KTE31	UPEC	4,514,939	50.65	ASTZ00000000
C-IV	1 176 05 S3 C2	Commensal (human)	4,457,844	50.67	JHDF00000000
C-IV	TW11588	Water	4,463,584	50.57	AEMF00000000
C-V	B116	Bacteraemic	4,576,704	50.35	LRWW00000000
C-V	KTE11	UPEC	4,486,744	50.49	ANSR00000000
C-V	KTE159	UPEC	4,776,266	50.26	ASVR00000000
C-V	KTE52	UPEC	4,615,592	50.32	ASUT00000000
C-V	KTE96	UPEC	4,592,997	50.4	ASVD00000000
C-V	TW09308	Water	4,809,826	50.28	AEME00000000

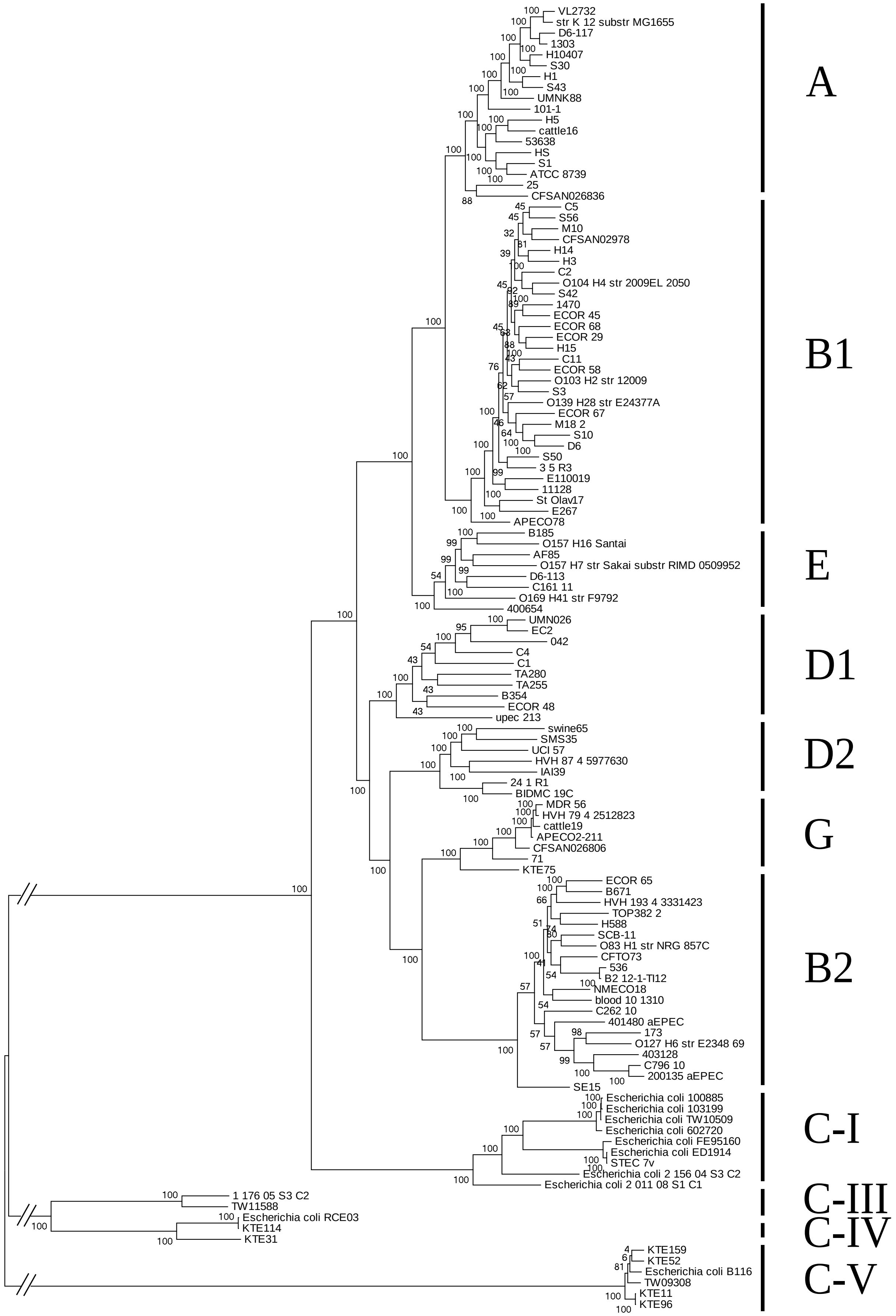

VL2732
str K 12 substr MG1655
D6-117
1303
H10407
S30
H1
S43
UMNK88
101-1
H5
cattle16
53638
HS
S1
ATCC 8739
25
CFSAN026836
C5
S56
M10
CFSAN02978
H14
H3
C2
O104 H4 str 2009EL 2050
S42
1470
ECOR 45
ECOR 68
ECOR 29
H15
C11
ECOR 58
O103 H2 str 12009
S3
O139 H28 str E24377A
ECOR 67
M18 2
S10
D6
S50
3 5 R3
E110019
11128
St Olav17
E267
APECO78
B185
O157 H16 Santai
AF85
O157 H7 str Sakai substr RIMD 0509952
D6-113
C161 11
O169 H41 str F9792
400654
UMN026
EC2
042
C4
C1
TA280
TA255
B354
ECOR 48
upec 213
swine65
SMS35
UCI 57
HVH 87 4 5977630
IAI39
24 1 R1
BIDMC 19C
MDR 56
HVH 79 4 2512823
cattle19
APECO2-211
CFSAN026806
71
KTE75
ECOR 65
B671
HVH 193 4 3331423
TOP382 2
H588
SCB-11
O83 H1 str NRG 857C
CFTO73
536
B2 12-1-TI12
NMECO18
blood 10 1310
C262 10
401480 aEPEC
173
O127 H6 str E2348 69
403128
C796 10
200135 aEPEC
SE15
Escherichia coli 100885
Escherichia coli 103199
Escherichia coli TW10509
Escherichia coli 602720
Escherichia coli FE95160
Escherichia coli ED1914
STEC 7v
Escherichia coli 2 156 04 S3 C2
Escherichia coli 2 011 08 S1 C1
1 176 05 S3 C2
TW11588
Escherichia coli RCE03
KTE114
KTE31
KTE159
KTE52
Escherichia coli B116
TW09308
KTE11
KTE96
A
B1
E
D1
D2
G
B2
C-I
C-III
C-IV
C-V

Figure 4.2 A RAxML maximum likelihood midpoint rooted phylogeny constructed using 1,865213 bp of core genome sequence from 100 *E. coli* strains of phylogenetic groups A-G defined in Chapter 3, and 20 *E. coli* strains of cryptic clade phylogenetic groups. Percentage bootstrap support values are shown on internal branches. The scale bar on the bottom left indicates the number of substitutions per site represented by the branch length shown. Major phylogenetic groups are labelled, including the sister group to group B2 which is putatively labelled group G.

To conduct novel alternative MLST schema analysis using the 100 strains from groups A-G and 20 from the cryptic clade groups, core gene alignments were obtained from the previously conducted pan genome analysis using these strain's genomes. Using these core gene alignments, a total of 1500 unique MLST schemas were next created for use with the 120 strain genomes. This was firstly done by defining their reference gene sequences, which took the place of loci in the Achtman, Pasteur, and EcMLST schemas. MLST schemas were created with reference gene sequences numbering 7, 15, 25, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000. To obtain reference gene sequences for a single schema, the method to obtain reference gene sequences for an *E. coli* MLST schema was carried out. For this, the output of the Roary analysis was used as input, the number of reference gene sequences to be included for the MLST schema, such as '7', were defined as a parameter. The method was repeated 100 times for MLST schemas of each of the sizes, where each time the desired number of randomly selected genes to include in the MLST schema was specified to the program. The result was the creation of 1500 unique MLST schemas. These alternative MLST schemas were created novelly and consisted of fewer *E. coli* core gene sequences (as in the Achtman, Pasteur, and EcMLST schemas) or the same or more *E. coli* core gene sequences (as in typically used cgMLST schemas) than the proposed 256 gene cgMLST. MLST schema genes were also randomly selected *E. coli* core reference genes as would occur in an existing cgMLST and MLST schema, so reference genes have both a history of no recombination and recombination.

Next, orthologues for gene sequences of each MLST schemas were obtained from the 2 *E. coli* strain sets representing *E. coli* phylogenetic groups A-G (100 strain genomes) and the cryptic clade groups C-I to C-V (20 strain genomes) using BLAST to obtain a set of orthologue sequences. After this, a 120-strain phylogeny was constructed for each of the 1500 MLST schemas. This was carried out by first concatenating the 120 orthologue sequences for a given MLST reference gene sequence to one another and aligning them using Muscle. Next, gene alignments were concatenated together to produce 1500 separate alignments, one for each MLST schema (where 120 orthologue sequences were present for each MLST reference gene sequence in the schema). Phylogenetic construction of each alignment using RAxML then occurred.

Next, topological consistency of each MLST phylogeny to the clonal frame phylogeny was assessed using ETE Toolkit (Huerta-Cepas et al. 2016). Consistency to the clonal frame phylogeny was recorded if the phylogeny adhered to two criteria:

i) The clade containing all strains of a given phylogenetic group must include zero (strict analysis; 100% consistency) or no more than 10%, 30%, and 50% (permissive analyses; ≤ 100% consistency) strains of other phylogenetic groups.

ii) Inferred ancestral nodes of phylogenetic groups must exhibit the same topology as in the inferred clonal frame phylogeny from Chapter 3 for phylogenetic groups A-G (Figure 4.3) and as in the core gene phylogeny from Chapter 3 for cryptic clade groups C-I to C-V (Figure 4.2).

Topologies of MLST schemas which fulfilled these criteria were then reviewed manually.

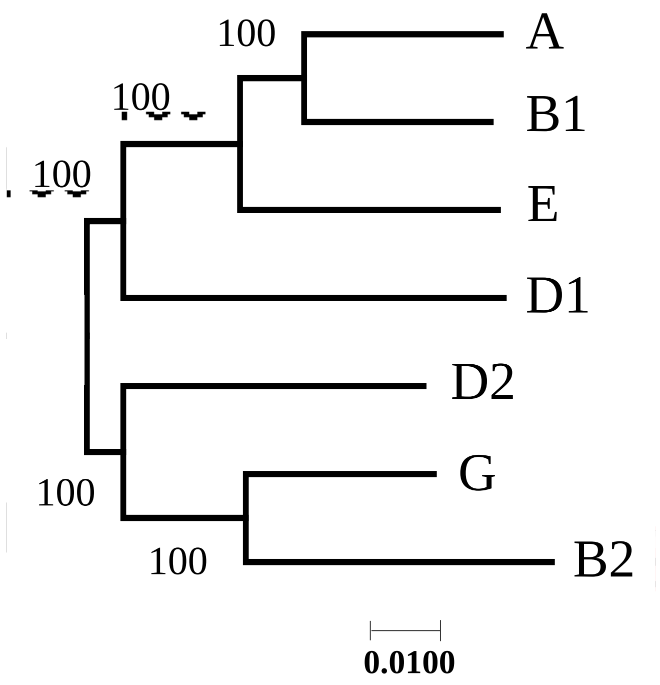


Figure 4.3. The inferred *E. coli* clonal frame phylogeny as reported in Chapter 3, showing the phylogenetic relationships of the 7 major phylogenetic groups A-G. Percentage bootstrap support values are shown on internal branches. The scale bar indicates the number of substitutions per site represented by the branch length shown. Major phylogenetic groups are labelled, including the sister group to group B2 which is putatively labelled group G.

The analysis reported that the percentage of MLST schema phylogenies with a phylogenetic group topology that was 50%, 70%, 90%, and 100% consistent with that of the clonal frame phylogeny for *E. coli* strains from phylogenetic groups A-G and consistent with the core gene phylogeny for strains from the cryptic clade groups C-I to C-V (Figure 4.2), increased with the number of reference gene sequences employed by the MLST schema (Table 4.4). This was the case up until 800 genes were employed by an MLST schema for both 90% and 100% consistency analyses. At this point and with the addition of further reference gene sequences employed by the MLST schema, 100% of phylogenies exhibited a topology which was 100% consistent to the intergroup topology of the clonal species phylogeny. 50% and 70% consistency in all MLST schema phylogenies was reached with the use of 200 and 600 genes respectively. No 5-gene MLST phylogenies exhibited 50% consistency to the clonal frame phylogeny (phylogenetic groups A-G) or core gene phylogeny (cryptic clade groups C-I to C-V) (Table 4.4).

Table 4.4. Results of the novel alternative MLST schema analysis.

	Percentage consistency with core genome phylogeny (N=100)			
Genes in MLST	**>= 50%**	**>= 70%**	**>= 90%**	**100%**
5	0	0	0	0
7	3	2	1	1
15	14	5	5	5
25	35	25	21	21
50	65	59	48	46
100	91	87	81	81
200	100	99	90	90
300	100	97	93	93
400	100	99	95	95
500	100	98	96	96
600	100	100	100	100
700	100	100	99	99
800	100	100	100	100
900	100	100	100	100
1000	100	100	100	100

4.3.2. Creating a 7-15 locus MLST schema for clonal group assignment

The objective of this chapter included creating a 7-15 locus MLST schema for use in *E. coli* clonal group assignment when the option of using a cgMLST is unavailable. To do this, the results of the novel alternative MLST schema analysis were inspected. Of the 1500 unique MLST schemas reported by the novel alternative MLST schema analysis, MLST schemas were selected which placed 100% of 100 strain genomes from phylogenetic groups A-G into their correct clonal group and 100% of cryptic clade strain genomes into their phylogenetic group as determined in the core gene phylogeny (Figure 4.2). These schemas were also selected on the basis that they exhibited relatively higher phylogenetic diversity than schemas employing the same number of reference gene sequences. This phylogenetic diversity was determined manually by inspecting phylogenetic branch lengths separating pairs of phylogenetic groups in the MLST schema phylogeny. Of the MLST schemas which fitted these criteria, a single schema was chosen based on the low number of 7 reference gene sequences that it employed (Table 4.5, Figure 4.4). The selection of a 7-gene schema was decided on the logic that analysis with an MLST schema employing fewer reference gene sequences would be completed in a shorter time period compared to using an MLST schema which employed more genes, given the same number of strain genomes to be used in MLST analysis. To determine if the quantity of gene sequence bases used in the chosen MLST schema could be reduced further, a version of the chosen MLST schema was created where each reference gene sequence was reduced to a locus sequence with a length of a maximum size of 400 bases. These locus sequences represented the most phylogenetically diverse gene sequence region for each gene across the 120 strain genomes. However, the resulting 7-locus sequence phylogeny was determined to provide inferior phylogenetic resolution to the complete gene phylogeny (Figure 4.3) as clonal groups B2 and G, and D1 and D2 were indistinguishable (Figure 4.5).

Table 4.5. Gene and locus name, sequence length, and gene product of the 7 genes in *E. coli* strain str. K-12 MG 1655 (GenBank accession U00096) which make up the proposed novel *in-silico* MLST schema.

Gene name: str. K-12 MG 1655	**Locus name: str. K-12 MG 1655**	**Sequence length (bp)**	**Gene product**
ruvA	b1861	612	Component of RuvABC resolvasome
tcdA	b2812	807	tRNA threonylcarbamoyladenosine dehydratase
ybgS	b0753	381	Putative periplasmic protein
ydjZ	b1752	708	TVP38/TMEM64 family inner membrane protein
yhjK	b3529	1,989	Cyclic-di-GMP phosphodiesterase
yidH	b3676	348	DUF202 family inner membrane protein
ypfG	b2466	1,044	DUF1176 family protein

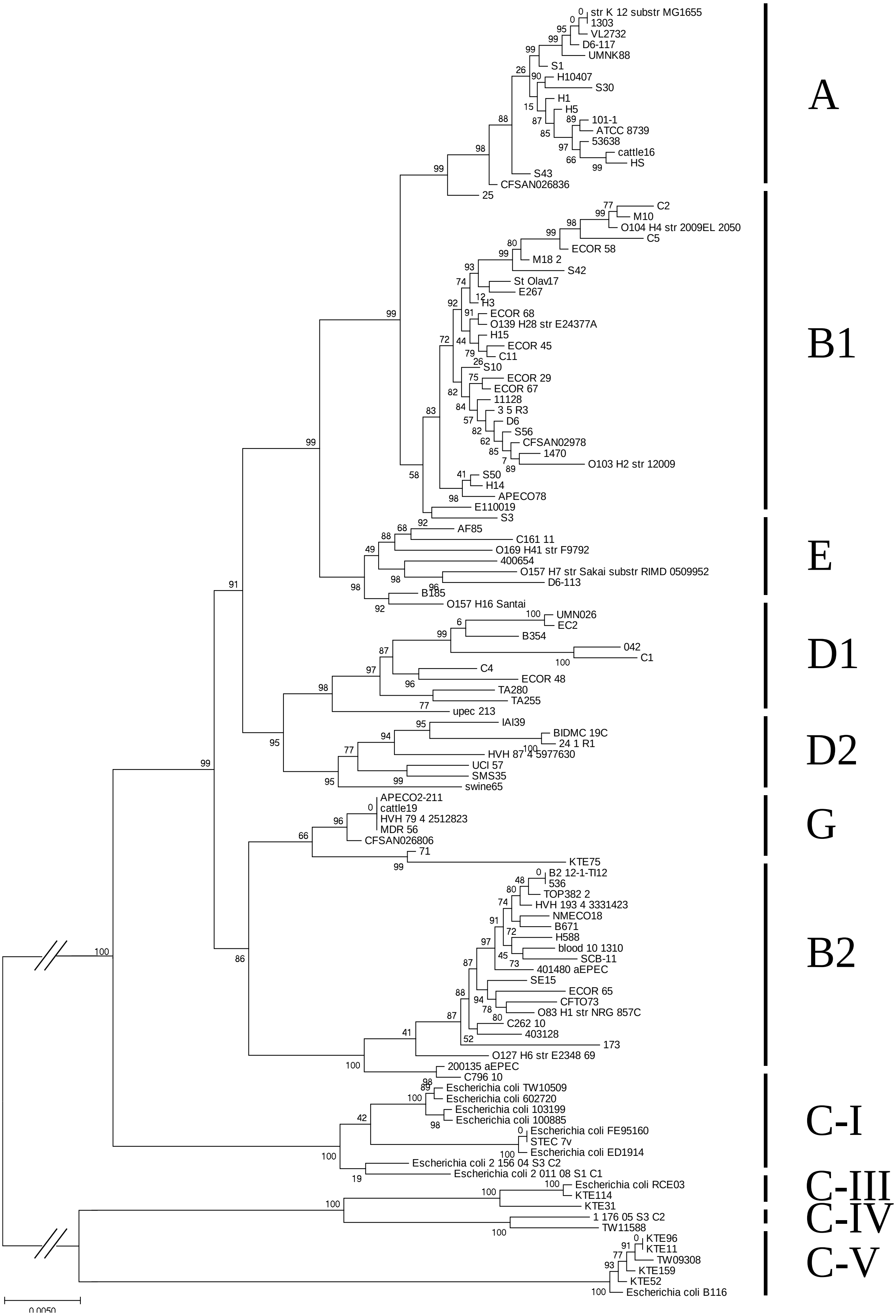

str K 12 substr MG1655
1303
VL2732
D6-117
UMNK88
S1
H10407
S30
H1
H5
101-1
ATCC 8739
53638
cattle16
HS
S43
CFSAN026836
25
C2
M10
O104 H4 str 2009EL 2050
C5
ECOR 58
M18 2
S42
St Olav17
E267
H3
ECOR 68
O139 H28 str E24377A
H15
ECOR 45
C11
S10
ECOR 29
ECOR 67
11128
3 5 R3
D6
S56
CFSAN02978
1470
O103 H2 str 12009
S50
H14
APECO78
E110019
S3
AF85
C161 11
O169 H41 str F9792
400654
O157 H7 str Sakai substr RIMD 0509952
D6-113
B185
O157 H16 Santai
UMN026
EC2
B354
042
C1
C4
ECOR 48
TA280
TA255
upec 213
IAI39
BIDMC 19C
24 1 R1
HVH 87 4 5977630
UCI 57
SMS35
swine65
APECO2-211
cattle19
HVH 79 4 2512823
MDR 56
CFSAN026806
71
KTE75
B2 12-1-TI12
536
TOP382 2
HVH 193 4 3331423
NMECO18
B671
H588
blood 10 1310
SCB-11
401480 aEPEC
SE15
ECOR 65
CFTO73
O83 H1 str NRG 857C
C262 10
403128
173
O127 H6 str E2348 69
200135 aEPEC
C796 10
Escherichia coli TW10509
Escherichia coli 602720
Escherichia coli 103199
Escherichia coli 100885
Escherichia coli FE95160
STEC 7v
Escherichia coli ED1914
Escherichia coli 2 156 04 S3 C2
Escherichia coli 2 011 08 S1 C1
Escherichia coli RCE03
KTE114
KTE31
1 176 05 S3 C2
TW11588
KTE96
KTE11
TW09308
KTE159
KTE52
Escherichia coli B116
0.0050
A
B1
E
D1
D2
G
B2
C-I
C-III
C-IV
C-V

Figure 4.4. A maximum likelihood phylogeny of the novel proposed in-silico 7-gene MLST schema used with 100 *E. coli* from phylogenetic groups A-G and 20 from cryptic glade groups C-I to C-V with labelled major phylogenetic groups to the right. The phylogeny was constructed using a concatenated alignment of 5,889 bp from the gene sequences of *ruvA*, *tcdA*, *ybgS*, *ydjZ*, *yhjK*, *yidH*, and *ypfG*. Percentage bootstrap support values are shown on internal branches. The scale bar on the bottom left indicates the number of substitutions per site represented by the branch length shown.

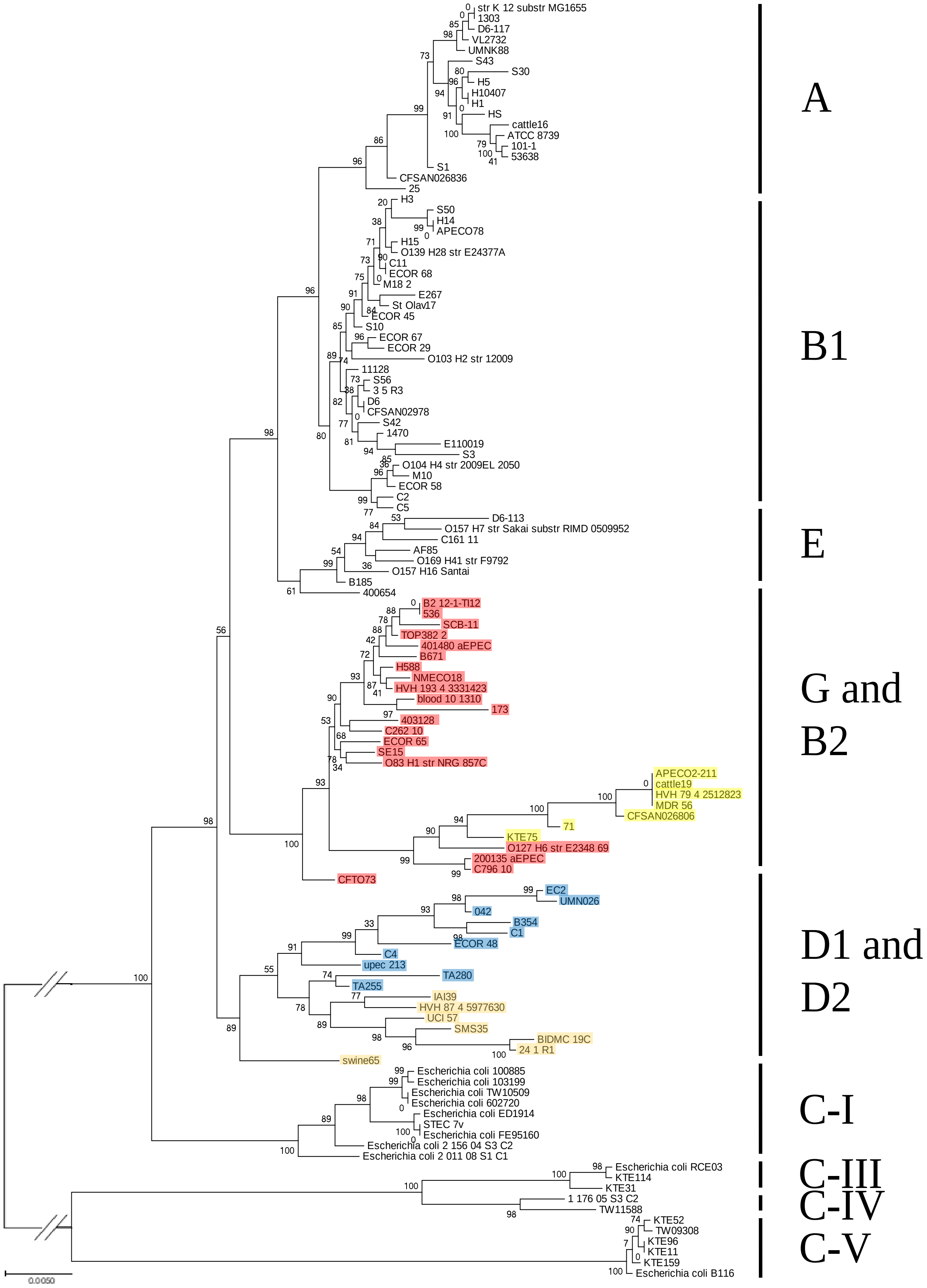

str K 12 substr MG1655
1303
D6-117
VL2732
UMNK88
S43
S30
H5
H10407
H1
HS
cattle16
ATCC 8739
101-1
53638
S1
CFSAN026836
25
H3
S50
H14
APECO78
H15
O139 H28 str E24377A
C11
ECOR 68
M18 2
E267
St Olav17
ECOR 45
S10
ECOR 67
ECOR 29
O103 H2 str 12009
11128
S56
3 5 R3
D6
CFSAN02978
S42
1470
E110019
S3
O104 H4 str 2009EL 2050
M10
ECOR 58
C2
C5
D6-113
O157 H7 str Sakai substr RIMD 0509952
C161 11
AF85
O169 H41 str F9792
O157 H16 Santai
B185
400654
B2 12-1-TI12
536
SCB-11
TOP382 2
401480 aEPEC
B671
H588
NMECO18
HVH 193 4 3331423
blood 10 1310
173
403128
C262 10
ECOR 65
SE15
O83 H1 str NRG 857C
APECO2-211
cattle19
HVH 79 4 2512823
MDR 56
CFSAN026806
71
KTE75
O127 H6 str E2348 69
200135 aEPEC
C796 10
CFTO73
EC2
UMN026
042
B354
C1
ECOR 48
C4
upec 213
TA280
TA255
IAI39
HVH 87 4 5977630
UCI 57
SMS35
BIDMC 19C
24 1 R1
swine65
Escherichia coli 100885
Escherichia coli 103199
Escherichia coli TW10509
Escherichia coli 602720
Escherichia coli ED1914
STEC 7v
Escherichia coli FE95160
Escherichia coli 2 156 04 S3 C2
Escherichia coli 2 011 08 S1 C1
Escherichia coli RCE03
KTE114
KTE31
1 176 05 S3 C2
TW11588
KTE52
TW09308
KTE96
KTE11
KTE159
Escherichia coli B116
0.0050
A
B1
E
G and
B2
D1 and
D2
C-I
C-III
C-IV
C-V

Figure 4.5. RAxML maximum likelihood phylogeny constructed using 400 bp sequence sections from each of the 7 genes *ruvA*, *tcdA*, *ybgS*, *ydjZ*, *yhjK*, *yidH*, and *ypfG* of the 7-gene *in-silico* MLST. Strains of groups D1 (dark blue) D2 (orange), G (yellow), and B2 (red) are highlighted showing the close clustering and relative short branches connecting strains of groups D1 and D2, and G and B2 to the point that strains of group D1 are indistinguishable from those of D2 and those of G from those of B2. Percentage bootstrap support values are shown on internal branches. The scale bar on the bottom left indicates the number of substitutions per site represented by the branch length shown.

4.3.3. The novel multiplex phylogenetic-group assignment schema

Part of the objective was to determine if it was possible to create a novel multiplex phylogenetic-group assignment schema with 4 loci. To address this, the output of the gene presence and absence file generated in the Roary pan genome analysis was inspected using an in-house program. The gene presence and absence file detailed gene presence for 100 strain genomes from *E. coli* phylogenetic groups A-G and 20 strain genomes from *E. coli* cryptic clade groups C-I to C-V (11 phylogenetic groups total). The program was used to identify genes in the file which were present in all members of each phylogenetic group (group core genes). The program then reported which of these genes were either present in at least 2 phylogenetic groups and absent in all other groups, or absent in at least 2 phylogenetic groups and present in all other groups (as in Clermont PCR). The program then combined the presence and absence combinations of these genes to report 4 genes which could be used in combination to assign a given strain genome its correct clonal group based on the combined presence and absence patterns of the 4 genes.

The analysis of genes across the pan genome of all *E. coli* phylogenetic groups revealed that a novel schema of 4 genes to correctly assign strains to each of the 11 phylogenetic groups was not possible. This was because no single gene was unique to 100% of strain genomes in group A and the following multiple phylogenetic group combinations: A+B1+E+D1+D2, D1+D2+G+B2, D2+G+B2, G+B2+C-I, G+B2+C-I+C+III, G+B2+C-

I+C-III+C-IV, G+B2+C-I+CIII+C-IV+C-V, and C-I+C-III. It was determined that 100% strain genome presence in groups A and or at least one of these multiple phylogenetic group combinations would have been required for a 4-gene multiplex schema to be possible. As an alternative to a 4-gene multiplex schema, 10 gene markers were reported which were found to be effective at assigning strains to their correct clonal phylogenetic group, with the exception of distinguishing A and B1, when searched for in genomes in specific combinations (Table 4.6).

Table 4.6. A set of 10 markers which has the same format as the Clermont PCR method by Clermont et al. (2013) designed for determining the major phylogenetic group of an *E. coli* strain or isolate to be A+B1, E, D1, D2, G, B2, C-I, C-III, C-IV, or C-V. Each marker accounts for recombination of marker sequences, was developed using a phylogenetically diverse set of *E. coli* genomes, and identifies strains belonging to the previously unreported phylogenetic group ‘G’. Use of 1-3 of the 10 markers can be used to assign a strain to one of 10 *E. coli* phylogenetic groups (groups A and B1 are indistinguishable using the method). Sequence markers are each derived from a gene without recombination in its phylogenetic history and which are core to a specific group or groups meaning the determined group is the *E. coli*’s clonal phylogenetic group. Identifying sequence markers can be carried out through a nucleotide BLAST analysis using at least a 50% identity and 80% length cut-off thresholds of a whole genome sequence, or through PCR of the sequence markers from an un-sequenced sample. For each sequence marker the given name and sequence length is provided. The reference strain name and GenBank accession where the sequence can be isolated are also provided with genomic base coordinates and the gene name and gene product information. The last 10 columns are the presence and absence patterns for each marker for *E. coli* from each major phylogenetic clonal group.

Given marker name	Marker length (bp)	Reference strain	GenBank accession	Marker sequence base coordinates	Strand	Reference Gene name	Product of full gene	A	B1	E	D1	D2	G	B2	C-I	C-III	C-IV	C-V
AB1E	420	str. K-12 substr. MG1655	U00096	642088 - 642507	+	*ybdR*	Putative Zn-dependent oxidoreductase	+	+	+	-	-	-	-	-	-	-	-
E	300	O157:H7 str. Sakai substr RIMD 0509952	BA000007.2	4452784 - 4453083	-	ECs4426	Putative fimbrial protein precursor	-	-	+	-	-	-	-	-	-	-	-
ED1D2	360	042	FN554766	1018435 - 1018794	+	EC042_0957	Putative type III effector protein	-	-	+	+	+	-	-	-	-	-	-
D2	480	SMS35	CP000970	2536010 - 2536489	+	EcSMS35_2489	SMC domain protein	-	-	-	-	+	-	-	-	-	-	-
G	330	MDR 56	CP019903	4943539 - 4943868	+	B1200_26620	Hypothetical protein	-	-	-	-	-	+	-	-	-	-	-
B2	155	O127:H6 str. E2348/69	FM180568	633134 - 633288	-	E2348C_0567	Predicted alcohol dehydrogenase	-	-	-	-	-	-	+	-	-	-	-
C-I	230	TW10509	GL872204	1497139 - 1497368	+	ERFG_01361	Isochorismatase	-	-	-	-	-	-	-	+	-	-	-
C-III	200	RCE03	JUDX00000000	2064054 - 2064253	-	AAW06_10000	Pilus assembly protein	-	-	-	-	-	-	-	-	+	-	-
C-IV	270	1 176 05 S3 C2	JHDF00000000	1556008 - 1556277	+	AC26_1524	Pili assembly chaperone	-	-	-	-	-	-	-	-	-	+	-
C-V	510	KTE52	ASUT00000000	2176779 - 2177288	-	A1SC_02111	Hypothetical protein	-	-	-	-	-	-	-	-	-	-	+

4.4. Discussion

The purpose of the work in this chapter was to explore evidence to support the preferential use of the proposed 256 gene 250,000 bp cgMLST. This cgMLST provided 100% correct clonal group assignment to the set of 100 strain genomes representing *E. coli* groups A-G as a result of employing reference gene sequence without a history of recombination between phylogenetic groups, and provided 100% correct phylogenetic assignment of cryptic clade groups, as based on a core gene phylogeny (see Figure 4.2). This 256 gene cgMLST schema also employed a reduced number of reference gene sequences than that employed by typically used ad hoc, publicly and privately available cgMLST schemas.

In the novel alternative MLST schema analysis, each unique MLST schema was created novelly and either consisted of less *E. coli* core gene sequences than the proposed 256 gene cgMLST (as in the Achtman, Pasteur, and EcMLST schemas) or the same or more *E. coli* core reference gene sequences than the proposed 256 gene cgMLST (as in typically used cgMLST schemas). MLST schema genes were also randomly selected *E. coli* core reference genes as would occur in existing cgMLST and MLST schemas, so reference genes have both a history of no recombination and recombination. The analysis showed that after the inclusion of any randomly selected 800 reference gene sequences in any given schema, strain genomes were assigned the correct clonal or cryptic clade phylogenetic group 100% of the time. In contrast, ≥ 50% correct clonal or cryptic clade phylogenetic group assignment was achieved with the inclusion of at least 200 genes. This finding indicated that a typically used cgMLST, either created based on a given set of strain's core genes, or one used from a public or private database, would need to employ at least 800 genes to provide correct clonal group assignment to *E. coli* strain genomes from phylogenetic groups A-G or correct phylogenetic group assignment to cryptic clade strain genomes. As the random novel MLST schema analysis showed 800 genes must be used in a schema for reliable correct clonal group assignment, the hypothesis was addressed and accepted as this is greater than the 256 used in the cgMLST. It can be inferred that use of the proposed cgMLST either as a standalone computational cgMLST within its own publicly available downloadable application or use as part of in silico sequence typing analysis for BIGSdb would be superior to typical cgMLST methods in terms of accuracy in phylogenetic and clonal group assignment and computational analysis time required to complete analysis. At the current stage of testing, it can therefore be suggested that the proposed 256 gene cgMLST be used preferentially for *E. coli* clonal group assignment for strains from phylogenetic groups A-G. When a greater number of strain genome sequences representing the *E. coli* cryptic clade groups are available, further testing can be carried out to determine if the cgMLST reference genes are free of recombination in cryptic clade groups C-I to C-V, but at the current stage of testing the proposed cgMLST does adequately assign phylogenetic groups, as determined by construction of a core gene phylogeny (see Figure 4.2). The cgMLST could either be implemented as a part of BIGSdb sequence typing for *E. coli* genome sequences, or as a standalone MLST with functionality in a publicly available program, where a user provides an *E. coli* genome sequence and BLAST is implemented to determine the correct clonal group assignment for the genome sequence. Further testing can also be carried out in future to determine if the proposed cgMLST can reliably assign to clonal complexes (clades at the sub-phylogenetic group level). Exploring evidence to support the preferential use of the proposed cgMLST for *E. coli* clonal group assignment addressed the first part of the principal aim.

The second part of the overall aim was to address an objective, designed to determine if a novel *in-silico* 7-15 gene MLST schema and *in-silico* 4-gene novel multiplex schema *E. coli* clonal group assignment schemas could be developed using genes identified as free from recombination in Chapter 3. The first part of the objective was to develop the 7-15 gene MLST. The novel 7 gene MLST schema presented provides 100% correct clonal group assignment for strain genomes from phylogenetic groups A-G and 100% correct assignment of phylogenetic groups for strain genomes from *E. coli* cryptic clade groups C-I to C-V. At the current stage of testing, it can therefore be proposed that the 7 gene MLST schema can be used either as a standalone computational cgMLST within its own publicly available downloadable application or use as part of in silico sequence typing analysis for BIGSdb. It can also be inferred to be suitable as an alternative for when the proposed 256 gene cgMLST cannot be used due to computational analysis time limitations. By creating a 7 gene MLST which can be used to assign *E. coli* phylogenetic groups, the first part of the objective was addressed.

The second part of the objective was to determine it was possible to create a novel 4-loci multiplex schema to correctly assign clonal or phylogenetic groups to *E. coli* strain genomes. As a 4-locus novel multiplex schema could not be determined, it can be suggested that the 10 markers determined which can correctly assign clonal or phylogenetic groups to *E. coli* strain genomes be used with Clermont PCR, to resolve unresolved group assignments or as an extra checking measure using 1-3 loci out of the available 10. Given this, additional characterisation would be required to differentiate groups A and B1 with this approach. However, further work to identify intergenic regions might reveal appropriate sequence regions which can be used as a multiplex schema to differentiate all clonal and phylogenetic groups using only 4 loci. By investigating whether it was possible to create a 4-locus novel multiplex schema for use as an alternative to the proposed 256 gene cgMLST, the second part of the objective was addressed.

In summary, the random novel MLST schema analysis revealed that a significant proportion of the core genome (800 genes) must be obtained to reliably assign strains to the correct clonal or phylogenetic group in the resulting phylogeny. Based on these factors it can therefore be recommended that the proposed 256 cgMLST should be used preferentially over typical cgMLST methods and approaches, all of which employ a greater number of reference gene sequences. However, if the number of strain or isolate genome sequences which require clonal or phylogenetic group assignment numbers hundreds or thousands, then it can be proposed that the determined 7-gene MLST can be used as an alternative approach to assigning *E. coli* clonal or phylogenetic groups. Like the proposed 256 gene cgMLST, the 7-gene MLST could either be implemented as a part of BIGSdb sequence typing for *E. coli* genome sequences sequence typing, or as a standalone MLST with functionality in a publicly available program, where a user provides an *E. coli* genome sequence and BLAST is implemented to determine the correct clonal group assignment for the genome sequence. Although development of a 4-loci novel multiplex for use as an alternative to the 256 gene MLST was not possible, 10 gene loci were reported which could be used to differentiate strains of all groups apart from those belonging to groups A and B1. It was proposed that these could be used as an option for computationally assigning clonal groups to *E. coli* from phylogenetic groups A-G and phylogenetic groups, as determined in the core gene phylogeny, to cryptic clade groups C-I to C-V as an addition to either the cgMLST or 7-gene MLST approaches, and carried out in a similar computational *in-silico* manner as them also.

Chapter 5: The evolutionary dynamics of bacterial genes important for urinary tract infection in a murine model

5.1. Introduction

Extraintestinal pathogenic *E. coli* (ExPEC) is an *E. coli* pathovar which represents a broad range of disease phenotypes and undergoes HT to spread VAFs (Foxman 2002). Uropathogenic *E. coli* (UPEC) is a subgroup of ExPEC which has attracted research attention for several decades as the subgroup are the principal and most significant causes of bacterial-mediated urinary tract infections (UTIs) (Wiles et al. 2013, Bingen et al. 1997, Picard et al. 1999, Salipante et al. 2015). UTIs are a highly ubiquitous set of disorders caused mostly by UPEC and represent a serious public health problem (reviewed in Foxman 2010). UPEC are the principal cause of UTIs for up to 90% of non-hospitalized patients and up to 50% of hospital-acquired (nosocomial) UTIs (Srinivasan *et al.* 2003, Tartof *et al.* 2005). Around 150 million people are affected worldwide each year which costs care centres an estimated 6 billion USD (equivalent to 4.9 billion GBP) per year (Stamm and Norrby 2001).

UPEC are believed to enter the body first via the faecal-oral route and are thought to maintain a reservoir population in the human intestine (Russo et al. 1995) where they faecally exit the body and can come into contact with the entrance to the urinary tract (Wiles et al. 2008). UPEC typically ascend the urethra and colonise the bladder and lower urinary tract causing cystitis, and in some cases, they subsequently ascend the ureters to colonise the kidneys causing pyelonephritis (Wiles et al. 2008). Bacterial recognition receptors on immune cells may then detect the presence of the UPEC structures including flagellum and peptidoglycan and lipopolysaccharide structures before initiating bacterial-induced signalling pathways (Mulvey et al. 2000). This detection stimulates the host immune system to respond with an inflammatory response which includes recruitment of neutrophils and immune cells capable of carrying out phagocytosis of bacterial cells (Haraoka et al. 1999), cytokine production (Mulvey et al. 2000), exfoliation of epithelial cells which may be colonised by UPEC (Floyd et al. 2010), and production of reactive oxygen such as nitric oxide which contribute to bacterial cell death through mediating cell DNA degradation by inhibition of oxidative phosphorylation (Floyd et al. 2012, Mulvey et al. 2000). However, UPEC have evolved mechanisms to evade these responses in many cases (Foxman 2002, Floyd et al. 2010).

UTIs can be contracted in people of any age or gender (Magliano et al. 2011, Litwin et al. 2005), but are more common in females; an estimated 33% of women develop UTI infections by age 24 (Foxman 2002). UTI incidence has also been reported to range between 10% to 30% in elderly hospitalised people and also be a major cause of morbidity in infant boys (Cove Smith and Almond 2007). UTIs are clinically described as either uncomplicated or complicated (Hooton and Stamm 1997). Uncomplicated UTIs include lower UTIs (cystitis) and upper UTIs (pyelonephritis) (Ronald 2002), and typically occur in individuals with healthy urinary tracts where there are no blockages (Foxman 2010). Complicated UTIs are associated with physiological abnormalities which obstruct urine flow such as renal failure, presence of polyps, or valve damage (Melekos and Naber 2000), or the presence of a catheter which can be used as a substrate for colonisation by UPEC (Warren 2001).

The virulence of UPEC is determined by the presence, number, and type of specific VAFs (Dobrindt 2005), which provide heightened fitness in a host environment and allows them to outcompete other bacteria, and colonise key extraintestinal anatomical regions (Wiles et al. 2008). Well studied UPEC VAFs include adhesins (Slavchev et al. 2009), toxins (Russo et al. 1995), immune evasion factors (Davis et al. 2006), iron acquisition factors (Crosa

1989), flagella (Haiko and Westerlund Wikström 2013), and surface polysaccharide capsules (Kusecek et al. 1984).

UPEC can use adhesin proteins to adhere to host anatomical structures such as the intestine epithelium when colonising the intestines (Mulvey et al. 2002), or capillary endothelium, urethra or ureter epithelium as a primary step in UTI pathogenesis (Klemm and Schembri 2000). This prevents their removal by the natural blood or urine flow of the host (Klemm and Schembri 2000). UPEC adhesion to a range of host tissues is mediated by fimbrial and afimbrial adhesins which exhibit overlapping roles in pathogenesis (Slavchev et al. 2009). P and type 1 fimbriae are the principal identified fimbriae adhesin types associated with UPEC (Slavchev et al. 2009), but other adhesins also exist including F1C and S adhesins and afrimbrial types (Lane and Mobley 2007) such as the Dr antigen-binding adhesins Afa and DraBC (Servin 2005). Another important adhesin is Antigen 43, which was found to significantly promote aggregation and biofilm growth and is associated with long-term persistence of UPEC in a murine model (Schembru et al. 2001). The adhesive property of UPEC means colonisation is initiated by adhesion to host cells (Katouli 2010), followed by development of local inflammation via induction of immune cell responses which the colonising UPEC next encounter (Johnson 1991).

The most studied UPEC toxins are α-haemolysin (HlyA) (Velasco et al. 2018), the cytotoxic necrotising factor 1 (CNF1) (Reppin et al. 2017) and the secreted autotransporter toxin (SAT) (Toloza et al. 2015). HlyA self-assembles transmembrane pores in host cells which directly results in the lysis or apoptosis of the affected cells as consequence of pore leakage. This leads to iron release which is crucial for the growth and proliferation of the bacterial cell that expressed the HlyA (Velasco et al. 2018). HlyA has also been implicated in host immune evasion (Dhakal and Mulvey 2012) and shown to stimulate the clearance of bladder urothelial cell surfaces and mediate rapid colonisation of the bacterial cell with expressed the HlyA in the recently cleared area (Smith et al. 2008, Floyd et al. 2010). CNF1 is a toxin used in renal cell invasion and expressed by up to 30% of pyelonephritis-causing UPEC strains (Bien et al. 2012). It has previously been found to stimulate the formation of actin stress fibres and membrane ruffles in renal cells thereby facilitating intracellular host cell access (Bien et al. 2012). It also prevents host immune activity by initiating apoptosis of bladder epithelial cells (Mills et al. 2000) and contributing to the down-regulation of phagocytosis the through targeting of the Rho family GTPase Rac2 which is critical for phagocytosis regulation in immune cells (Davis et al. 2005). The SAT toxin is an important toxin for pyelonephritis-causing UPEC (Guyer et al. 2002) and has previously been found to alter host cell signalling cascades (Dhakal et al. 2008), alter the host inflammatory response (Dhakal et al. 2008), and cause toxicity to renal and bladder cells (Bien et al. 2012).

In mammalian hosts, extracellular iron is low as it is sequestered in cells or attached to proteins such as lactoferrin (Masson et al. 1969, Litwin et al. 1993) and transferrin on mucosal surfaces (Anderson and Vulpe 2009), and haemoglobin in erythrocytes (Porcheron et al. 2013). Iron is crucial for UPEC growth and proliferation and employing iron acquisition factors such as siderophores to obtain iron from the environment are an important part of pathogenesis (Litwin and Calderwood 1993). Siderophores compete with host immune defences to uptake iron released from environmental lactoferrins and transferrins (Litwin and Calderwood 1993). In a study of 221 UPEC isolate genomes, Salipante et al. (2013) found 18 iron metabolism-associated ExPEC VAFs present in 80-100% of UPEC strains. Siderophores that UPEC use include the salmochelin, yersiniabactin, aerobactin, and enterobactin systems (Johnson et al. 2007). The first of which has been found to be important for UPEC pathogenesis for strain UTI89 communities growing in bladder epithelial cells in mice (Reigstad et al. 2007).

Yersiniabactin was found to be important for cystitis and pyelonephritis infection, as preventing yersiniabactin uptake was found to reduce infection development (Braumbaugh et al. 2015). Lastly, the aerobactin and enterobactin siderophore systems have previously been found to significantly contribute to of UPEC strain 83972 in an iron-deficient medium (Watts et al. 2012).

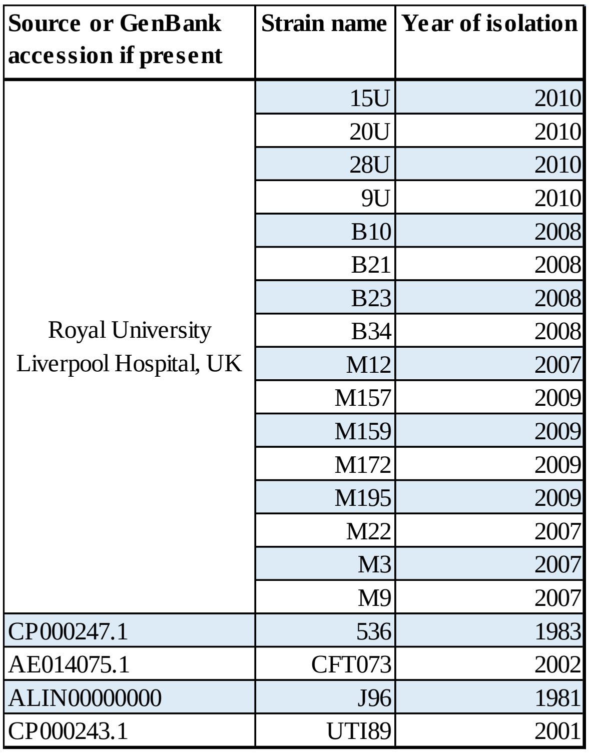
Flagella are macromolecular filamentous organelles used for bacterial motility through fluids or across surfaces (Terashima et al. 2008) and are thought to be used for moving into and up the urinary tract in UTIs (Floyd et al. 2010). They also play a role in biofilm formation (Pratt and kolter 1998), adhesion to host cells (Girón et al. 2002), and protein export (Young et al. 1999, Haiko and Westerlund Wikström 2013). They are comprised of three subunits: a basal support, hook, and filament (Terashima et al. 2008, Wright et al. 2015, Pratt and Kotler 1998, Bien et al. 2012). After entry to the urinary tract, flagella also allow bacteria to ascend from the lower urinary tract to the kidney renal duct cells (Bens et al. 2014) and previous studies indicated that at least 70% and up to 90% of UTIs are caused by UPEC expressing flagella when in contact with epithelial cell surface of the urinary tract (Bien et al. 2012).

On the cell surface, UPEC express polysaccharide structures including lipopolysaccharides (LPS) (Schilling et al. 2001), and polysaccharide capsules (Anderson et al. 2010) that are linked to pathogenesis (Bien et al. 2012). LPS stimulates the proinflammatory response in UTIs (Säve et al. 2010). However, their role in inducing ascending UTIs is unclear (Bien et al. 2012). Mutations in LPS gene *dsbA* reduced UPEC attachment to form biofilms, suggesting a role for LPS in biofilm formation (Genevaux et al. 1999), an important part of host-defence resistance for UPEC (Pratt and Kolter 1998). Polysaccharide capsules have been implicated in preventing phagocytosis through impairment of antibody protein binding (Howard and Glynn 1971) and protecting cells from bactericides present in human blood (Raksha et al. 2003). In a previous study, 7 genes encoding polysaccharide capsules were also found to be prevalent in 50%-80% of 221 UPEC strains analysed by Salipante et al. (2013).

Uropathogenic *E. coli* (UPEC) have received research attention for several decades (Yamamoto 2007, Gomez-Cruz et al. 2018). Despite this there remains a lack of understanding regarding UPEC infection of the upper urinary tract, specifically the mechanism by which UPEC impair contraction prior to colonisation (Floyd et al. 2012). Mammalian ureters naturally contract via peristalsis to transport urine from the kidneys to the bladder and prevent infection but UPEC-mediated impairment of contraction has been observed (Grana et al. 1968, Teague and Boyarsky 1968). This UPEC-mediated impairment causes ureter contraction to weaken and become less regular (Grana et al. 1968, Teague and Boyarsky 1968). The consequence of the ureters reduced movement means the UPEC can then more easily attach to and colonise the ureter cell surface (Grana et al. 1968, Teague and Boyarsky 1968). After ureter colonisation, UPEC are typically then able to ascend the ureter and colonise the kidneys, resulting in a longer term and more severe UTI and potentially the development of pyelonephritis (Grana et al. 1968, Teague and Boyarsky 1968). Colonising the ureter is a crucial part of UPEC pathogenesis and an increased understanding of it has the potential to contribute to the development of novel UTI therapeutics.

Dr Rachel Floyd and Professor Craig Winstanley of the Institute of Infection and Global Health at the University of Liverpool have previously investigated the UPEC-associated impairment of ureter contraction by developing a model which experimentally measures UPEC-mediated decreases in rat ureter contractility and has been shown to be comparable in response to human ureters (Floyd et al. 2010). Ureter contraction occurs through excitation from propagated action potentials across the epithelial cells which causes calcium ion (Ca^{2+}) transit into ureteric cells. The model involves isolating and mounting rat ureters to steel hooks in a physiological saline solution (pH 7.4 with composition (in mM) including 154 NaCl, 5.6 KCl, 1.2 MgSO4, 2 CaCl2, 8 glucose, and 10.9 HEPES) which provide 5-7V electrical pulses. Ureter contractions are measured as they naturally respond to electrical pulses with and without inoculation of specific UPEC strains. Previously the model has revealed time-dependent contractility impairment over 5 hours with the UPEC strains J96 and 536 by 89% and 87% respectively (Floyd et al. 2010) and by 96.75, 87.93, 78.03, 75.98, 42.18, 9.47 % for UPEC strains UTI89, CFT073, EC958, M160, M9, and M12, respectively (Floyd et al. 2012), relative to the sterile control level inhibition of 6.00% and 8.77% in each respective study. The latter study produced evidence implicating the *hlyCABD* operon and the gene *fimH* in UPEC-mediated contractility impairment, but it was recognised that a study employing a greater number of strains was needed to fully understand the genetic basis behind ureter contractility impairment phenotypes (Floyd et al. 2012). To address this need, 20 UPEC strains of were collated, 16 of which were isolated from the Royal Liverpool University Hospital (Table 5.1).

Table 5.1. Uropathogenic strain genomes used in this chapter



Source or GenBank accession if present	Strain name	Year of isolation
Royal University Liverpool Hospital, UK	15U	2010
	20U	2010
	28U	2010
	9U	2010
	B10	2008
	B21	2008
	B23	2008
	B34	2008
	M12	2007
	M157	2009
	M159	2009
	M172	2009
	M195	2009
	M22	2007
	M3	2007
	M9	2007
CP000247.1	536	1983
AE014075.1	CFT073	2002
ALIN00000000	J96	1981
CP000243.1	UTI89	2001

The 20 strains were subjected to the rat ureter contractility experimental model over 9 hours and the 16 hospital strains were whole genome sequenced, then all genome sequences were provided to me for analysis. ‘TG2’: a laboratory and non-pathogenic strain of *E. coli* called K-12 TG2 was used as a first control and a sterile version of the experiment model denoted ‘Krebs’ was used as a second control. The phenotype analysis provided by Dr Floyd and Professor Winstanley illustrated decreases in ureter contractility associated with the 21 strains and the sterile control over 9 hours (data for strains associated with contractility decreases of 0%-30% and 30%-100% at hour 9 are shown in Figures 5.1a and Figure 5.1b respectively).

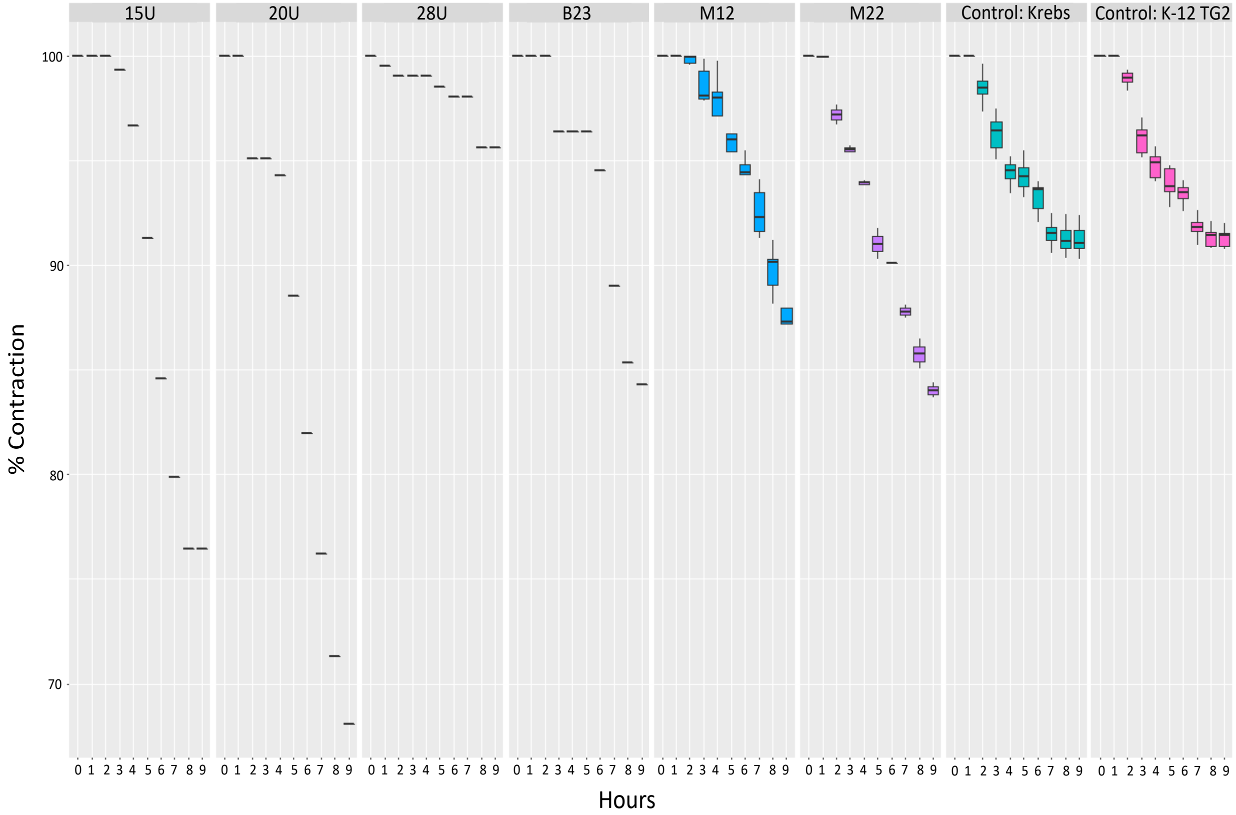


Figure 5.1a. Results of 20 closely related UPEC strains subjected to the rat ureter contractility model. Experiments are shown where over 9 hours the percentage ureter contraction associated with strains ranged from 100%-70% (also interpreted as a decrease in ureter contractility of 0%-30% associated with each strain). For strains where replicates were conducted a boxplot showing the mean and standard deviation of values at each hour is shown instead of a single line. Positive control runs of the model are labelled ‘Krebs’: a sterile physiological saline solution (pH 7.4 with composition (in mM) including 154 NaCl, 5.6 KCl, 1.2 MgSO4, 2 CaCl2, 8 glucose, and 10.9 HEPES), and ‘TG2’: a laboratory and non-pathogenic strain of *E. coli* called K-12 TG2. Data produced and provided by Dr Rachel Floyd and Professor Craig Winstanley.

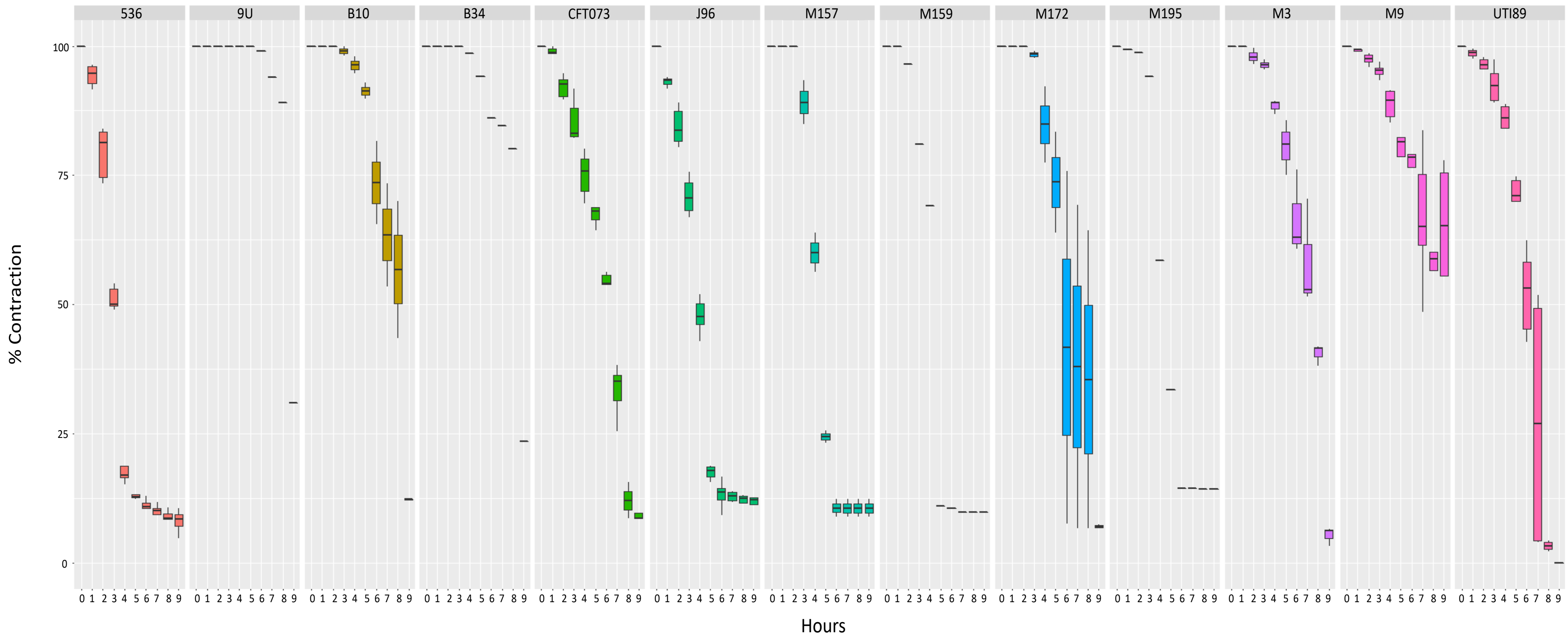


Figure 5.1b. Results of 20 closely related UPEC strains subjected to the rat ureter contractility model. Experiments are shown where over 9 hours the percentage ureter contraction associated with strains ranged from 100%-0% (also interpreted as a decrease in ureter contractility of 0%-100% associated with each strain). Only strains associated with ureter contractility of ≥70% (contractility decreases of 30%-100%) at hour 9 are shown, the remaining strains are shown in Figure 5.1a. For strains where replicates were conducted a boxplot showing the mean and standard deviation of values at each hour is shown instead of a single line. Positive control runs of the model are labelled ‘Krebs’: a sterile physiological saline solution (pH 7.4 with composition (in mM) including 154 NaCl, 5.6 KCl, 1.2 MgSO4, 2 CaCl2, 8 glucose, and 10.9 HEPES), and ‘TG2’: a laboratory and non-pathogenic strain of *E. coli* called K-12 TG2. Data produced and provided by Dr Rachel Floyd and Professor Craig Winstanley.

The purpose of the work in this chapter was to use the phenotype and genome sequence data for the 20 UPEC strains to provide insights into the genetic and evolutionary basis of *E. coli* ureter contractility inhibition phenotypes.

5.1.1. Hypothesis

The hypothesis of this chapter was designed to determine the genetic differences between UPEC strains which are associated with the observed phenotypic patterns of ureter contractility inhibition:

Genetic differences between UPEC strains are significantly associated with the observed phenotypic patterns of ureter contractility inhibition:

5.1.2. Overall aim and objectives

The overall aim was to determine genetic differences which are significantly associated with observed differences in ureter contractility inhibition phenotypes across the 20 UPEC strains. The aim addresses two objectives, designed to produce specific information about the evolution of these ureter contractility inhibition phenotypes:

4. To determine if HT has contributed to the phenotypic differences observed across strains.
5. To use phenotype-associated gene information to infer the mechanism of action underlying each observed phenotype pattern.

5.2. Methods

5.2.1. Preparing strain cultures for use in the phenotype experiment

To obtain standardised isolate cultures for the 16 UPEC isolates obtained from the Royal University Hospital Liverpool, Dr Rachel Floyd and Professor Craig streaked strain isolates onto agar plates, which were then incubated overnight at 37°C to ensure optimal colony growth. Following static serial passage, strain cultures were pelleted at 5,000 xg for 5 minutes at 4°C. The pellet was resuspended in sterile physiological saline (pH 7.4 with composition (in mM) including 154 NaCl, 5.6 KCl, 1.2 MgSO4, 2 CaCl2, 8 glucose, and 10.9 HEPES) and was diluted until the OD600 for each strain corresponded to 1–2 × 107 colony-forming units (CFU) per 50 µL. The resulting solution for each strain was then used in the phenotype experiment

5.2.2. Defining phenotypic groups

Phenotypic groups reflecting the observed patterns of decreases in ureter contractility were defined based on two criteria. Firstly, the number of hours following infection, for which two time points were considered, 5h and 9h post-infection, to distinguish "early" and "late" effects. Secondly the percentage decrease in the amplitude of ureter contraction grouped into the phenotypes: "weak" (a decrease in contractility of 8% - 100% (control level was 7%)), "mild" (a decrease in contractility of 20% - 100%), "moderate" (a decrease in contractility of 40% - 100%), "strong" (a decrease in contractility of 60% - 100%) and "severe" (a decrease in contractility of 80% - 100%). The strains exhibiting stronger phenotypes were included in the weaker groups to account for the possibility of genes with cumulative effects on contractility.

5.2.3. Identifying genes significantly associated with phenotype groups

Genes significantly associated with strains of the phenotype groups were identified in gene enrichment analysis using Scoary (Brynildsrud et al. 2016), which took the output of Roary and conducted a Fisher’s exact test (Fisher 1922, Agresti 1992) for each gene to determine whether its presence was significantly associated with one or more of the defined phenotypic traits.

5.3. Results

5.3.1. Selection of 20 UPEC strains and obtaining cultures and genome sequences

The 16 UPEC strains from the Royal University Liverpool Hospital were chosen for study because they have been isolated from patient urinary tract infections of varying severity (Table 5.1). Based on this the 20 strains were collectively thought by Dr Rachel Floyd and Professor Craig Winstanley to have an accurate representation of the range of UTI virulence-associated *E. coli* genes observable which cause ureter contractility inhibition phenotypes. The four additional strains 536, CFT073, J96, and UTI89 were also chosen for inclusion in the study due to their previously reported association with severe UPEC infections (Knapp et al. 1986, Kao et al. 1997, Blum et al. 1995, Mulvey et al. 2001). Strain isolates and genome sequences were previously available for strains UTI89, CFT073, 536, and J96, but the remaining 16 strains were sampled from the Royal Liverpool University Hospital and whole genome sequenced at the University of Liverpool (Table 5.1).

5.3.2. Pan genome analysis and creation of core gene phylogeny

To identify shared genes across the 20-strain set the assembled genome sequences were standardly annotated with Prokka and the pan genome determined by Roary with a 95% amino acid identity value. To determine the phylogenetic relationships of the 20 strain genomes in relation to the set of 100 *E. coli* genomes representing phylogenetic groups A-G (Chapter 3), the set of 100 were also included in the pan genome analysis. The core gene alignment generated by Roary was then used to construct a phylogeny using RAxML (Figure 5.2). The core gene phylogeny showed that the 20 strains clustered within phylogenetic group B2, 18 of which clustered into two clusters each containing 9 strains (Figure 5.2). Strains 536 and J96 each clustered adjacently to one of the clusters. To determine the relatedness of the 20 strains one another based on gene content, a dendrogram depicting strains clustered by the size of shared gene contents was also constructed using the post-analysis tools provided by Roary (Figure 5.3). The most notable difference between the core gene phylogeny and the dendrogram clustering strains by shared accessory gene content was the altered position of strain J96 between the two as it clusters more closely with 9 strains in the latter which it does not cluster with in the former (Figure 5.3).

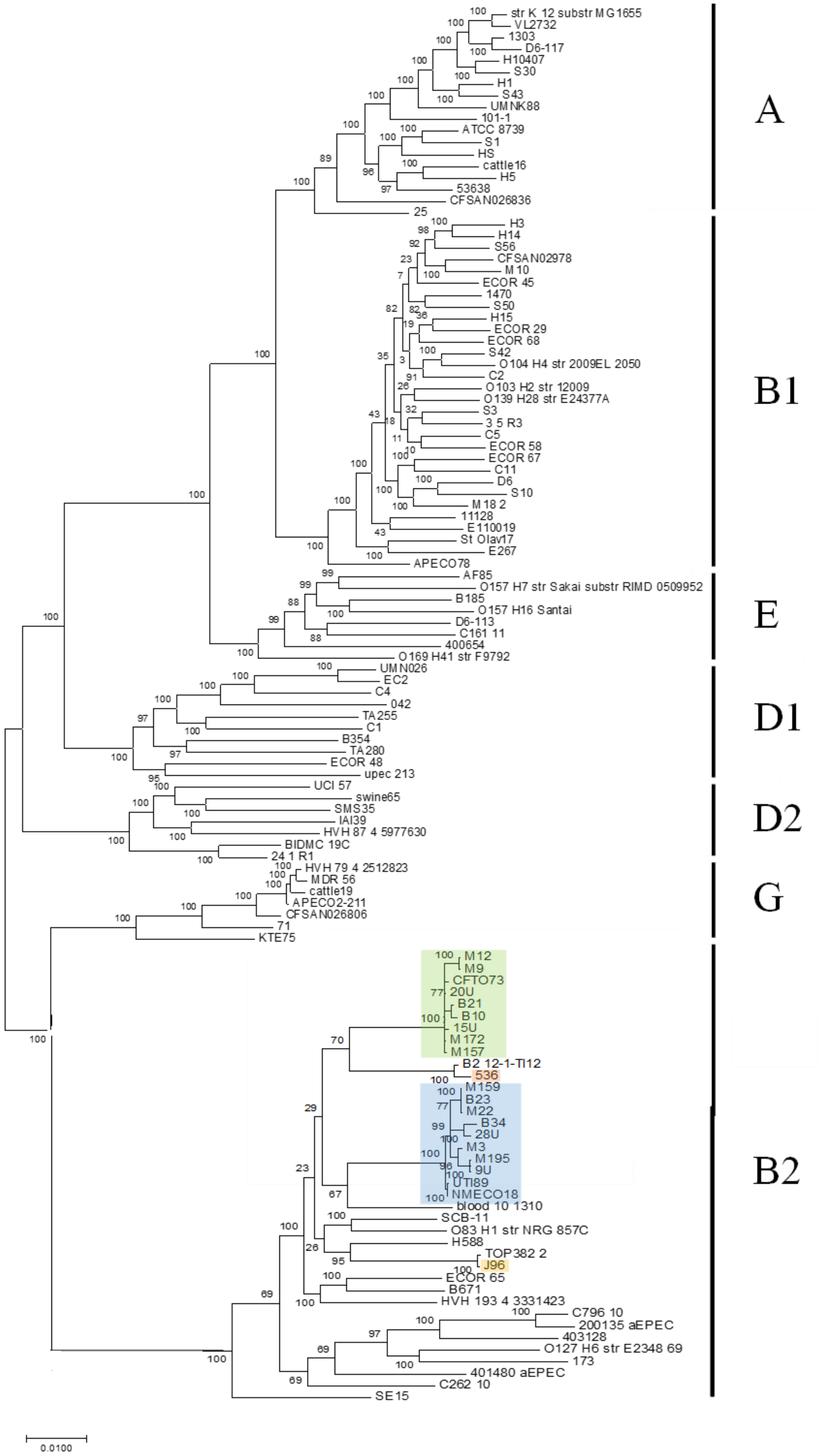

str K 12 substr MG1655
VL2732
1303
D6-117
H10407
S30
H1
S43
UMNK88
101-1
ATCC 8739
S1
HS
cattle16
H5
53638
CFSAN026836
25
H3
H14
S56
CFSAN02978
M10
ECOR 45
1470
S50
H15
ECOR 29
ECOR 68
S42
O104 H4 str 2009EL 2050
C2
O103 H2 str 12009
O139 H28 str E24377A
S3
3 5 R3
C5
ECOR 58
ECOR 67
C11
D6
S10
M18 2
11128
E110019
St Olav17
E267
APECO78
AF85
O157 H7 str Sakai substr RIMD 0509952
B185
O157 H16 Santai
D6-113
C161 11
400654
O169 H41 str F9792
UMN026
EC2
C4
042
TA255
C1
B354
TA280
ECOR 48
upec 213
UCI 57
swine65
SMS35
IAI39
HVH 87 4 5977630
BIDMC 19C
24 1 R1
HVH 79 4 2512823
MDR 56
cattle19
APECO2-211
CFSAN026806
71
KTE75
M12
M9
CFTO73
20U
B21
B10
15U
M172
M157
B2 12-1-TI12
536
M159
B23
M22
B34
28U
M3
M195
9U
UTI89
NMECO18
blood 10 1310
SCB-11
O83 H1 str NRG 857C
H588
TOP382 2
J96
ECOR 65
B671
HVH 193 4 3331423
C796 10
200135 aEPEC
403128
O127 H6 str E2348 69
173
401480 aEPEC
C262 10
SE15
A
B1
E
D1
D2
G
B2
0.0100

Figure 5.2. RAxML maximum likelihood midpoint rooted phylogeny constructed using a 1.8 Mb core genome alignment. The tree shows the phylogenetic placement of strains of the 20 UPEC strains across phylogenetic group B2 clustered with the *E. coli* strains representative of phylogenetic groups A-G (Chapter 3). Different colours highlight the topological position of the 20 UPEC strains spread over four distinct lineages of the group. The blue cluster also includes the UPEC strain NMECO18 which is not in the set of 20 UPEC strains. Percentage bootstrap support values are shown on internal branches. The scale bar on the bottom left indicates the number of substitutions per site represented by the branch length shown.

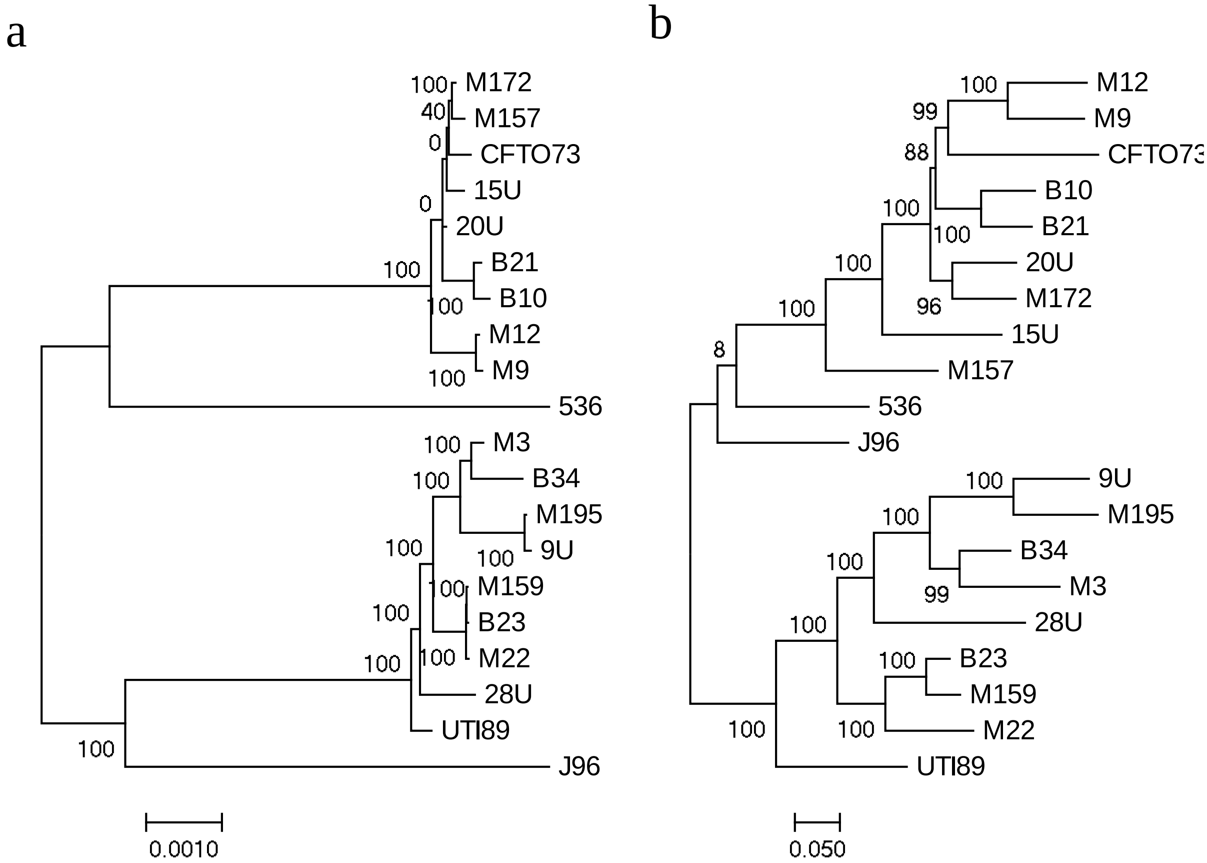


Figure 5.3. Midpoint rooted phylogenies constructed using (a) a 3.62 Mb core gene alignment (RAxML maximum likelihood phylogeny) (and (b) shared accessory gene content for the 20 strain UPEC set. Percentage bootstrap support values are shown on internal branches. The scale bar indicates the number of substitutions per site (a) and the distance in terms of number of shared genes (b), represented by the branch length shown.

5.3.3. Assigning phenotype groups to each of the 20 UPEC strains

In order to the identify genes responsible for the range of observed phenotypes, strains were grouped based on the ureter contractility inhibition phenotypes they were associated with. This enabled genes which were significantly enriched within the groups of strain genomes to be identified. Such genes were then inferred to be involved in the presentation of the group's exhibited phenotype. It was decided that strain groups should be defined based on their mean percentage ureter contractility inhibition phenotypes between ranges observed at 5-hour and 9-hour time points to account for "early" and "late" effects. The grouping system was designed so phenotype groups overlapped, and strains were assignable to multiple phenotype groups so that genes significantly associated with multiple phenotype groups could be identified. As both the sterile positive control 'Krebs' and laboratory strain *E. coli* K-12 TG2 were associated with a mean ureter contractility decrease of up to 7% after 9 hours, control level was determined to be 7%. Due to this, the lowest percentage ureter contractility decrease assigned to a phenotype was 8% (the case for the phenotypes; 1A and 2A) (Table 5.2).

Table 5.2. Details of the ureter contractility decrease phenotype groups to which the 20 UPEC strains were assignable to based on their exhibited phenotypes.

Group 1: mean percentage contractility after 5 hours into ureter phenotype experiment

Phenotype Group	Phenotype: ureter contractility decrease	Strains in group - exhibit phenotype	Strains not in group - do not exhibit phenotype
1A	8% - 100% (control level 7%)	15U, 20U, 536, B10, B21, CFT073, J96, M3, M9, M22, M157, M159, M172, M195, UTI89	9U, 28U, B23, B34, K-12 TG2, M12
1B	20% - 100%	536, CFT073, B21, J96, M157, M159, M172, M195, UTI89	9U, 15U, 20U, 28U, B10, B23, B34, K-12 TG2, M3, M9, M12, M22
1C	60% - 100%	536, J96, M157, M159, M195	9U, 15U, 20U, 28U, B10, B21, B23, B34, CFT073, K-12 TG2, M3, M9, M12, M22, M172, UTI89
1D	80% - 100%	536, J96, M159	9U, 15U, 20U, 28U, B10, B21, B23, B34, CFT073, K-12 TG2, M3, M9, M12, M22, M157, M159, M172, UTI89

Group 2: mean percentage contractility after 9 hours into ureter phenotype experiment

Phenotype Group	**Phenotype: ureter contractility decrease**	**Strains in group - exhibit phenotype**	**Strains not in group - do not exhibit phenotype**
2A	8% - 100% (control level 7%)	9U, 15U, 20U, 536, B10, B21, B23, B34, CFT072, J96, M3, M9, M12, M22, M157, M159, M172, M195, UTI89	28U, K-12 TG2
2B	20% - 100%	9U, 15U, 20U, 536, B10, B21, B34, CFT073, J96, M3, M9, M157, M159, M172, M195, UTI89	28U, B23, K-12 TG2, M12, M22
2C	40% - 100%	9U, 536, B10, B21, B34, CFT073, J96, M3, M9, M157, M159, M172, M195, UTI89,	15U, 20U, 28U, B23, K-12 TG2, M12, M22
2D	60% - 100%	9U, 536, B34, CFT073, J96, B10, B21, M3, M157, M159, M172, M195, UTI89	15U, 20U, 28U, B23, K-12 TG2, M9, M12, M22
2E	80% - 100%	536, B10, B21, CFT073, J96, M3, M157, M159, M172, M195, UTI89	9U, 15U, 20U, 28U, B23, B34, K-12 TG2, M9, M12, M22

5.3.4. Identifying and reporting information about phenotype-associated genes

To determine the presence of genes significantly associated with specific phenotype groups, gene enrichment analysis was performed using the output provided by Roary pan genome analysis. The result of this was a list of genes significantly associated with strains in each of the defined phenotype groups. Genes significantly associated with a specific phenotype were identified for phenotypes 1A (8%-100% contractility decrease after 5 hrs (control level 7%)), 1B (20%-100% contractility decrease after 5 hrs), 1C (60% - 100% contractility decrease after 5 hrs), 1D (80% - 100% contractility decrease after 5 hrs), 2A (8%-100% contractility decrease after 9 hrs (control level 7%)), 2B (20%-100% contractility decrease after 9 hrs), 2C (40%-100% contractility decrease after 9 hrs), 2D (60%-100% contractility decrease after 9 hrs), 2E (80%-100% contractility decrease after 9 hrs). Genes identified as significantly associated with strains from a certain phenotype group by Scoary were separated into those clustered (within 10kb) on the reference strain chromosome, and those which were not. To identify the functional roles for proteins encoded by each of the identified genes it was necessary to obtain functional annotations from multiple sources which were additional to those already provided using Prokka. To carry this out, reference protein sequences for genes previously identified as important for ExPEC and UPEC virulence with functional annotations were collated for use in a BLAST analysis to determine if any were orthologues or homologues of the proteins encoded by the identified genes. The reference protein sequences of 116 genes encoding known ExPEC VAFs and 2,238 putative UPEC virulence fitness genes (PFGs) present in seven different published data sets (Phan et al. 2013, Wiles et al. 2013, and Subashchandrabose et al. 2013, Salipante et al. 2014) were collated in total. Protein sequences from these sets were selected for this study as collectively they could provide unparalleled insight into the genetic basis to the phenotypes exhibited by UPEC. The known ExPEC VAF sequences were provided by Salipante et al. (2013). The PFGs were taken from studies using Transposon sequencing (TnSeq) or Transposon Directed Insertion Sequencing (TraDIS) (Landridge et al. 2009, van Opijnen 2009) to identify genes putatively associated with infection in UPEC models of infection. These include genes identified by Phan et al. (2013), Wiles et al. (2013), and Subashchandrabose et al. (2013). Phan et al. (2013) used TraDIS to identify 56 fitness genes associated with UPEC strain EC958 survival in human blood serum survival, and Subashchandrabose et al. (2013) used TraDIS to identify 334 fitness genes associated UPEC strain CFT073 survival in a murine model of bacteraemia. Similarly, Wiles et al. (2013) used TnSeq to identify 1,940 associated with UPEC strain F11 survival in zebrafish, of which 970 were associated with survival in the zebrafish embryo, 772 survival in blood, 122 survival in the pericardial cavity (PC), and 76 survival within multiple niches.

Reference protein sequences for the 116 ExPEC VAF-encoding genes were obtained from Supplementary dataset 9 published by Salipante et al. (2013). Reference protein sequences for PFGs listed in Table 2 of Phan et al. (2013) were extracted from the *E. coli* strain EC958 genome (GenBank accession HG941718), and the encoded protein sequence determined. Similarly, PFGs listed in Supplementary Table 2 of Subashchandrabose et al. (2013) were extracted from the *E. coli* strain CFT073 genome (AE014075), and PFGs listed in Supplementary Table 2 of Wiles et al. (2013) were extracted from the genome of *E. coli* strain

F11 (GenBank accession: AAJU00000000). All protein sequences across the 7 data sets were then concatenated together into a single file to simulate protein coding sequences of a genome sequence. Annotations for protein sequences in the file were then obtained using Prokka and non-Prokka methods (use of Sanger manually-annotated *E. coli* genomes and the KEGG database).

Genes associated with phenotypes observed after 5 hours

The gene enrichment analysis revealed 33 genes to be significantly associated phenotype groups observed after 5 hours into the phenotype experiment. For group 1A - strains exhibiting 8%-100% contractility decrease after 5 hrs (control level 7%), 13 genes were found to be significantly associated with strains in this group (Table 5.3). The gene of locus ID ECP_2028 in str. 536 (inferred hypothetical protein) was present in 9 of 15 strains (and present in 6 non-1A strains and absent in 9 non-1A strains, p=0.0186, Fisher's exact test (Fisher 1922, Agresti 1992)). Two gene clusters were inferred to be associated with the phenotype group 1A, one consisting of 9 genes denoted cluster 1A1 in Table 5.3, and one of 2 genes denoted cluster 1A2. The 9 genes comprising the cluster denoted 1A1 are present together on the chromosome in 8 of 15 phenotype group 1A strains (20U, 536, CFT073, J96, M157, M172, M9, UTI89) and absent in all other strains (p = 0.0456, Fisher's exact test). The 9 genes form a 12,318 bp gene cluster which has a conserved structure (Figure 5.4). These genes (locus IDs c3564-c3574 in str. CFT073) were found to be the *hlyCABD* operon, which have previously been associated with ExPEC virulence (Velasco et al. 2018). A gene also associated with phenotype group 1A strains (locus ID c3564 in str. CFT073, inferred two-component sensor protein KdpD) was also found to be a homologue of a putative fitness gene (PFG) (locus ID EcF11_0725 in str. F11) reported as associated with infection of multiple bodily niches of zebrafish by Wiles et al. (2013). 3 other genes were present in 11 of 15 phenotype group 1A strains and in 1 non-1A strain, two of which were denoted as cluster 1A2 in Table 5.3 and included a gene (locus ID ECP_4581 in str. 536) encoding haemolysin transport protein ShlB.

Table 5.3. Genes significantly associated with strains from phenotype group 1A: mean ureter contractility decrease after 5 hrs of 8%-100%.

Reference strain	Gene locus in reference strain	Grouped gene cluster	Group strains present in	Group strains absent in	Non-group strains	Non-group strains	Naïve p value	Inferred annotation based on identity to genes of known function	Gene and protein name based on identity to genes of known
536	ECP_2028	-	9	6	0	6	0.0186	Hypothetical protein	-
CFT073	c3564	1A1	8	7	0	6	0.0456	Two-component sensor protein	*kdpD*, KdpD
	c3565		8	7	0	6	0.0456	Putative two-component response regulator, alkaline phosphatase synthesis transcriptional regulatory protein	*phoP*, PhoP
	c3566		8	7	0	6	0.0456	Prokaryote cytochrome b561	-
	c3567		8	7	0	6	0.0456	Oxidoreductase	-
	c3568		8	7	0	6	0.0456	Hypothetical protein	-
	c3569		8	7	0	6	0.0456	Hemolysin C, Hemolysin-activating lysine-acyltransferase HlyC	*hlyC*, HlyC
	c3570		8	7	0	6	0.0456	Hemolysin A	*hlyA*, HlyA
	c3573		8	7	0	6	0.0456	Hemolysin B, Alpha-hemolysin translocation ATP-binding protein HlyB	*hlyB*, HlyB
	c3574		8	7	0	6	0.0456	Hemolysin D, Hemolysin secretion protein D	*hlyD*, HlyD
536	ECP_4584	1A2	11	4	1	5	0.0464	Putative DNA-binding protein	-
	ECP_4581		11	4	1	5	0.0464	Hemolysin transporter protein	*shlB*, ShlB
	ECP_1138	-	11	4	1	5	0.0464	RNA polymerase-binding transcription factor	*dksA*, DksA

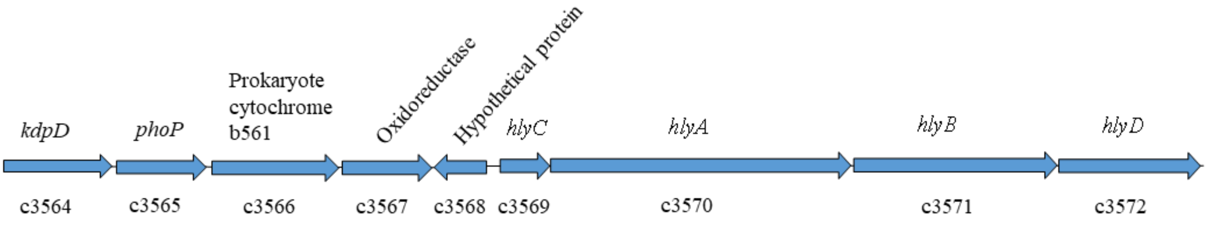


Figure 5.4. A 9 gene cluster identified as significantly associated with strains exhibiting phenotype 1A: ureter contractility decrease of 8% - 100% (control level 7%) exhibited after 5 hours as it occurs in strain CFT073 (shown as gene cluster 1A1 in Table 5.3). Gene locus IDs as in strain 536 (bottom). Gene names or descriptions of protein functions where a gene name is unavailable were provided based on protein sequence identity to known proteins (top).

Phylogenies for an example gene (locus ID ECP_4581 in str. 536) encoding haemolysin transporter protein ShlB (Figure 5.5) and the 9 gene cluster denoted 1A1 (Figure 5.6) respectively each displayed topologies which were inconsistent with that for the same strains in their core genome phylogeny. For gene ECP_4581 (inferred *shlB*), strain 536 and UTI89 and strains 536 and J96 clustered closer to one another than in the core genome phylogeny.

The remaining strains were clustered together in a manner consistent for both phylogenies. For the 9 gene cluster strains M172, M159, M9, and 20U clustered together in the cluster phylogeny as in the core genome phylogeny. However, strain CFT073 clustered more closely to strain J96 than the former strains. Strains TOP382 2, SCB-11, B2 12-1-TI12, and ECOR 48 were all also closely clustered with strain UTI89 and with one another in the cluster phylogeny compared to the core genome phylogeny, the latter of which, strain ECOR 48 was the outgroup strain.

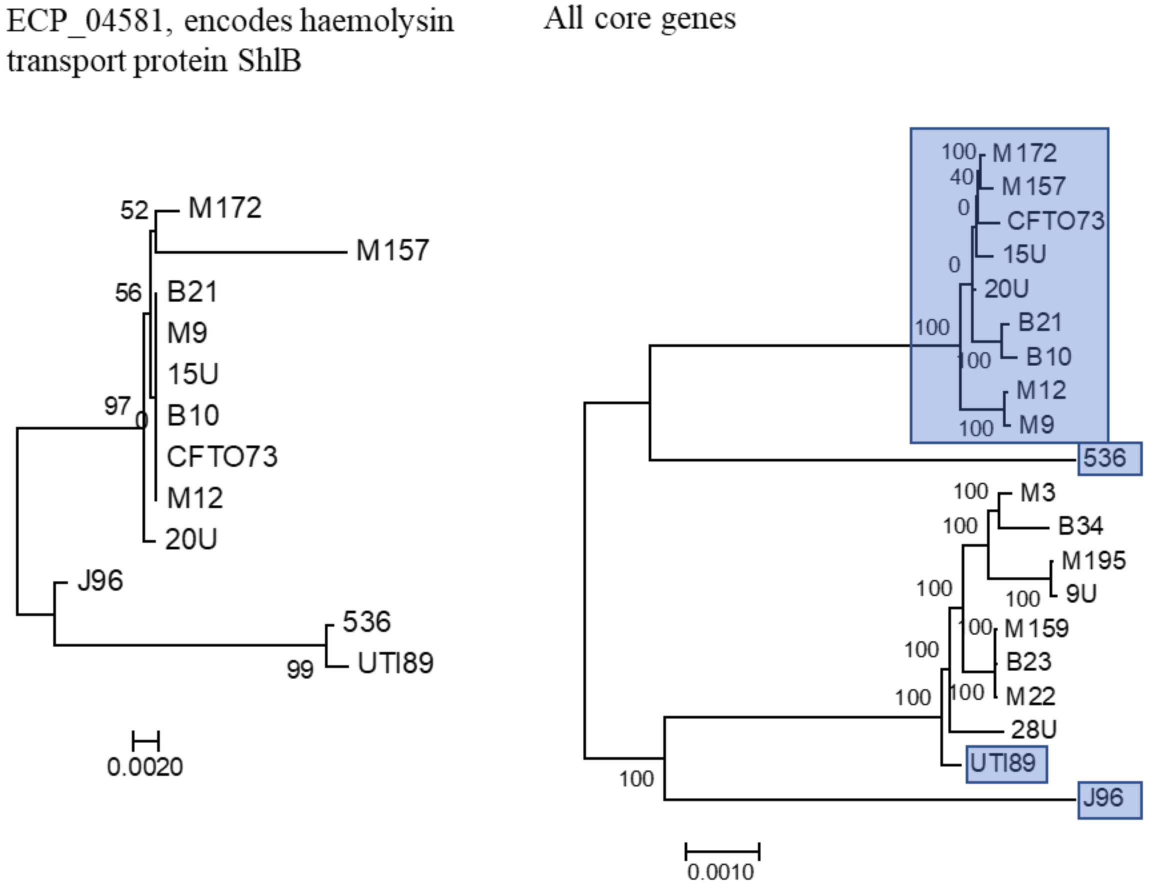


Figure 5.5. Maximum likelihood phylogeny constructed using an alignment of 1,773 bp of gene ECP_04581 (lous ID str. 536) encoding haemolysin transport protein ShlB from 12 UPEC strains (left). For topological comparison the core genome phylogeny of all 20 UPEC strains shown first in Figure 5.2a is shown (right). Strains common to the gene phylogeny are highlighted in blue in the core gene phylogeny. Percentage bootstrap support values are shown on internal branches. The scale bar indicates the number of substitutions per site represented by the branch length shown.

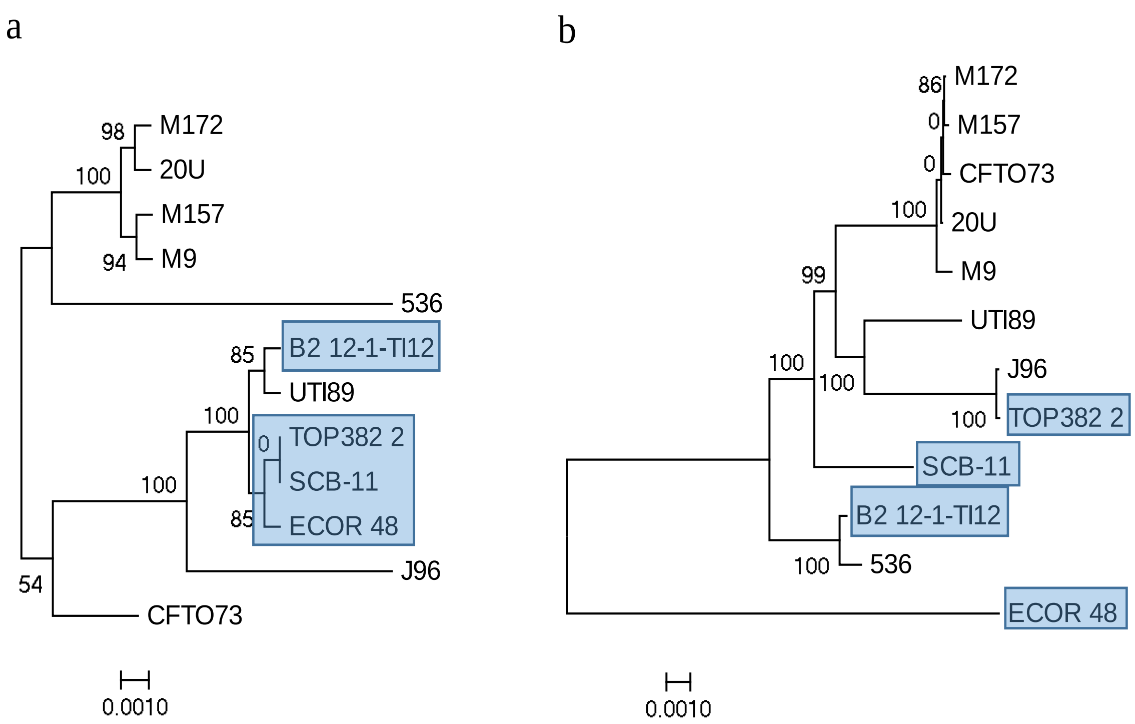
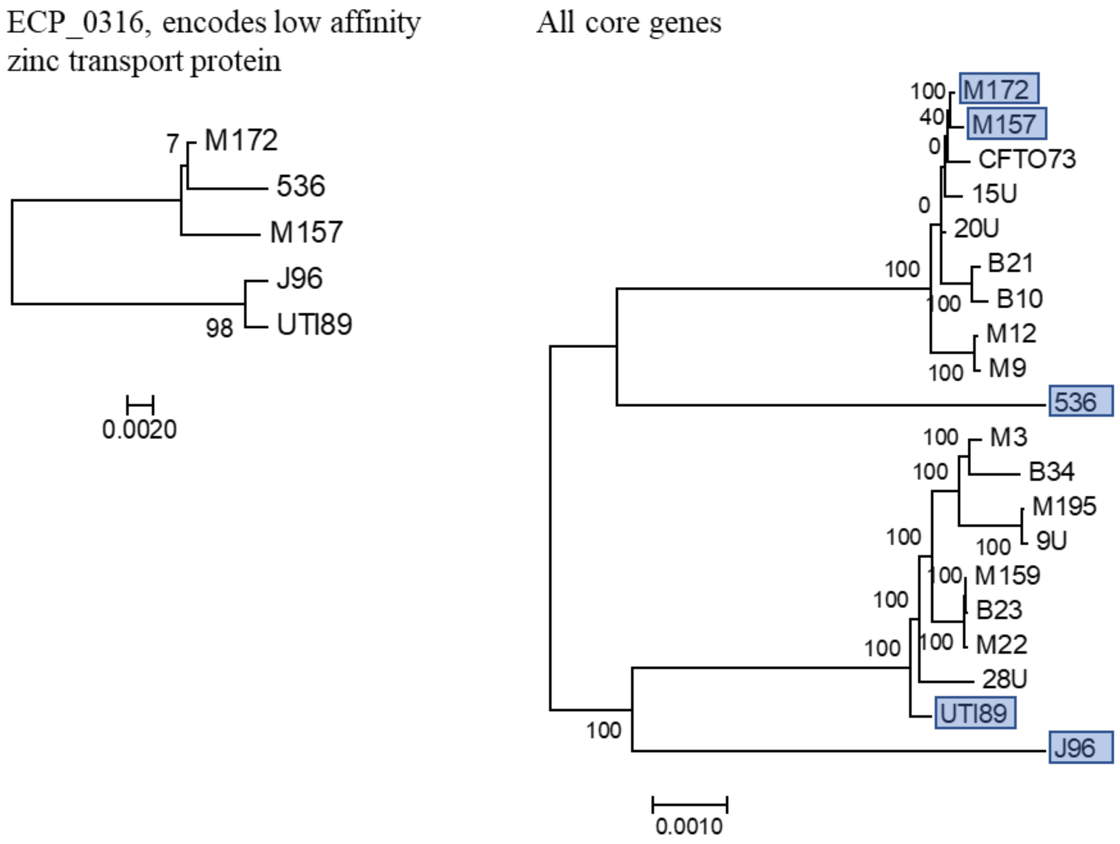


Figure 5.6. Maximum likelihood phylogenies constructed using (a) 12,318 bp of the 9 gene cluster 1A1 found in 8 of the 21 analysed strains and significantly associated with phenotype 1A and (b) a core gene phylogeny constructed using a 3.3 Mb core genome alignment from the same strains. Other than the 8 strains, this cluster is found in 3 group B2 strains (B2 12-1-TI12 (AIEC), SCB-11 (NMEC), and TOP382 2 (commensal human) and 1 group D1 strain: ECOR 48 (commensal human) (highlighted blue in both phylogenies). Percentage bootstrap support values are shown on internal branches. The scale bar indicates the number of substitutions per site represented by the branch length shown.

Phenotype group 1B was assigned to strains exhibiting 20%-100% contractility decrease after 5 hours (Table 5.4) and three genes were significantly associated with the phenotype. Each of these genes were present in a different set of strains. Two were annotated as hypothetical proteins, one of which was found to be a homologue of a PFG (locus ID EcF11_3640 in str. F11) identified by Wiles et al. (2013) as significantly associated with increased survival in a pericardial model of zebrafish infection. The single annotated gene (locus ID ECP_0316 in str. 536, present in 536, J96, M157, M172, and UTI89) encoded a low-affinity zinc transport protein which was present with a gene encoding a hypothetical protein in 5 of 9 phenotype group 1B strains and absent in 4 group 1B strains. This gene was also absent in 1B strains, and absent in 12 non-1B group strains (p=0.0062, Fisher's exact test). In a phylogeny constructed using orthologues of this gene a topology is displayed which is inconsistent with the core genome phylogeny (Figure 5.7). Strain 536 clustered closer to strain M172 relative to M157, and both 536 and J96 are clustered significantly closer to the other strains in the gene phylogeny compared to in the core genome phylogeny.

Table 5.4. Genes significantly associated with phenotype 1B: mean ureter contractility decrease of 20% - 100% exhibited after 5 hours.

Reference strain	Gene locus in reference strain	Grouped gene cluster	Group strains present in	Group strains absent in	Non-group strains	Non-group strains	Naïve p value	Inferred annotation based on identity to genes of known function	Gene and protein name based on identity to genes of known
536	ECP_3019	-	5	4	0	12	0.0062	Conserved hypothetical protein	-
	ECP_0316	-	5	4	0	12	0.0062	Low-affinity zinc transport protein	-
	ECP_2043	-	6	3	1	11	0.0158	DUF932 domain protein; unknown function	-



ECP_0316, encodes low affinity zinc transport protein

All core genes

Figure 5.7. Maximum likelihood phylogeny constructed using 300 bp of gene ECP_0316 (locus ID str. 536) which encodes a low-affinity zinc transport protein. For topological comparison the core genome phylogeny of all 20 UPEC strains is also shown. Strains common to the gene phylogeny are highlighted in blue in the core gene phylogeny. Percentage bootstrap support values are shown on internal branches. The scale bar indicates the number of substitutions per site represented by the branch length shown.

Phenotype group 1C was assigned to strains exhibiting a 60%-100% contractility decrease after 5 hours (Table 5.5) and 4 genes were significantly associated with the phenotype. These included 2 genes clustered together (cluster 1C1 in Table 5.5) which encode the 16S methyltransferase protein KsgA and a hydrolase enzyme (locus ID M157_00002 and M157_00003). This cluster was present only in strains M157, M159, and M195 out of the 20 UPEC strains analysed and the *E. coli* K012 TG2 control strain. Another gene associated with phenotype 1C was annotated as an uncharacterised protein (locus ID M157_03942 in str. M157), and the fourth gene (locus ID ECP_1146 in str. 536) encodes a transcriptional regulator and is present only in strains 536, J96, M157. All 4 genes are present in 3 of 5 phenotype 1C strains, absent from non-1C strains, and absent in 2 phenotype 1C strains (p = 0.0075 phenotype association, Fisher's exact test (Fisher 1922, Agresti 1992)). The 1C1 gene with locus ID M157_00003 (inferred to encode a hydrolase protein) was found to be a homologue of a PFG (locus ID EcF11_2058 in str. F11) associated with increased survival in a bacteraemic model of zebrafish infection (Wiles et al. 2013).

Table 5.5. Genes significantly associated with phenotype 1C: mean ureter contractility decrease of 60% - 100% exhibited after 5 hours.

Reference strain	Gene locus in reference strain	Grouped gene cluster	Group strains present in	Group strains absent in	Non-group strains	Non-group strains	Naïve p value	Inferred annotation based on identity to genes of known function	Gene and protein name based on identity to genes of known
536	ECP_1146	-	3	2	0	16	0.0075	Helix-turn-helix transcriptional regulator	-
M157	M157_00003	1C1	3	2	0	16	0.0075	Hydrolase	-
	M157_00002		3	2	0	16	0.0075	16S ribosomal RNA methyltransferase	*ksgA*, KsgA
	M157_03942	-	3	2	0	16	0.0075	Uncharacterised protein	-

Phenotype group 1D was assigned to strains exhibiting an 80%-100% contractility decrease after 5 hours (Table 5.6) and 13 genes were significantly associated with the phenotype, all of which were present in group strains 536 and J96 only. The genes included 2 clusters of 8 and 3 genes (labelled as clusters 1D1 and 1D2 in Table 5.6). Cluster 1D1, as it occurs in the strain 536 chromosome, is shown in Figure 5.8. It consists of genes encoding a second copy of the *hlyCABD* operon distinct from that detailed under the phenotype 1A section (*hly*I), and denoted *hly*II (genes with locus IDs ECP_3826 – ECP_3829 in str. 536). The cluster also contains genes encoding two ABC transporter ATP-binding proteins, a putative periplasmic solute binding protein, and a hypothetical protein (locus ID ECP_3822 – ECP_3825 in str. 536). Gene cluster 1D2 included 2 genes encoding proteins not yet characterised, one of which (locus ID 2812 in str. 536) was found to be a homologue of a PFG gene (locus ID 1892 in str. CFT073) associated with survival in a bacteraemia model of infection (Subashchandrabose et al. 2013). The third ID2 gene encodes a putative membrane protein (locus ID ECP_2813 in str. 536). Two additional genes associated with phenotype 1D encoded an uncharacterised protein (locus ID ECP_03514) and a 50S ribosome-binding GTPase protein (locus ID ECP_3008 in str. 536) which was found to be a homologue of a gene associated with survival in a zebrafish embryo model of infection using UPEC strain F11 (Wiles et al. 2013).

Table 5.6. Genes significantly associated with phenotype 1D: mean ureter contractility decrease of 80% - 100% exhibited after 5 hours.

Reference strain	Gene locus in reference strain	Grouped gene cluster	Group strains present in	Group strains absent in	Non-group strains	Non-group strains	Naïve p value	Inferred annotation based on identity to genes of known function	Gene and protein name based on identity to genes of known
536	ECP_3822	1D1	2	1	0	18	0.0143	Putative ABC transporter ATP-binding protein	-
	ECP_3823		2	1	0	18	0.0143	Putative ABC transporter ATP-binding protein	-
	ECP_3824		2	1	0	18	0.0143	Putative periplasmic solute binding protein	-
	ECP_3825		2	1	0	18	0.0143	Hypothetical protein	
	ECP_3826		2	1	0	18	0.0143	Hemolysin C, Hemolysin-activating lysine-acyltransferase HlyC	*hlyC*, HlyC
	ECP_3827		2	1	0	18	0.0143	Hemolysin A	*hlyA*, HlyA
	ECP_3828		2	1	0	18	0.0143	Hemolysin B, Alpha-hemolysin translocation ATP-binding protein HlyB	*hlyB*, HlyB
	ECP_3829		2	1	0	18	0.0143	Hemolysin D, Hemolysin secretion protein D	*hlyD*, HlyD
	ECP_2811	1D2	2	1	0	18	0.0143	Uncharacterised protein	-
	ECP_2812		2	1	0	18	0.0143	Uncharacterised protein	-
	ECP_2813		2	1	0	18	0.0143	Putative membrane protein	-
	ECP_3008	-	2	1	0	18	0.0143	50S ribosome-binding GTPase	-
	ECP_3514	-	2	1	0	18	0.0143	Uncharacterised protein	-

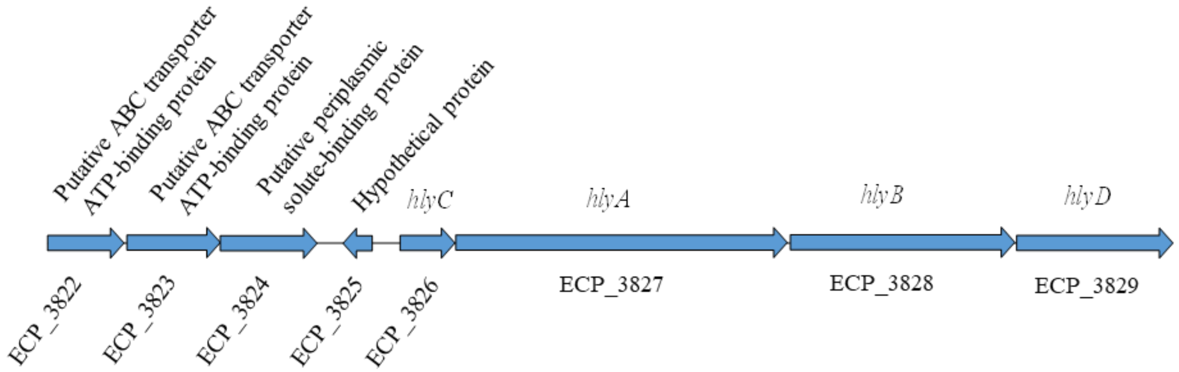

Putative ABC transporter ATP-binding protein
Putative ABC transporter ATP-binding protein
Putative periplasmic solute-binding protein
Hypothetical protein
hlyC
hlyA
hlyB
hlyD
ECP_3822
ECP_3823
ECP_3824
ECP_3825
ECP_3826
ECP_3827
ECP_3828
ECP_3829

Figure 5.8. The 8 gene cluster 1D1, identified as significantly associated with strains exhibiting phenotype 1D and unique to strains 536 and J96. Gene locus IDs as in strain 536 (bottom). Gene names or descriptions of protein functions where a gene name is unavailable were provided based on protein sequence identity to known proteins (top).

Genes associated with phenotypes observed after 9 hours

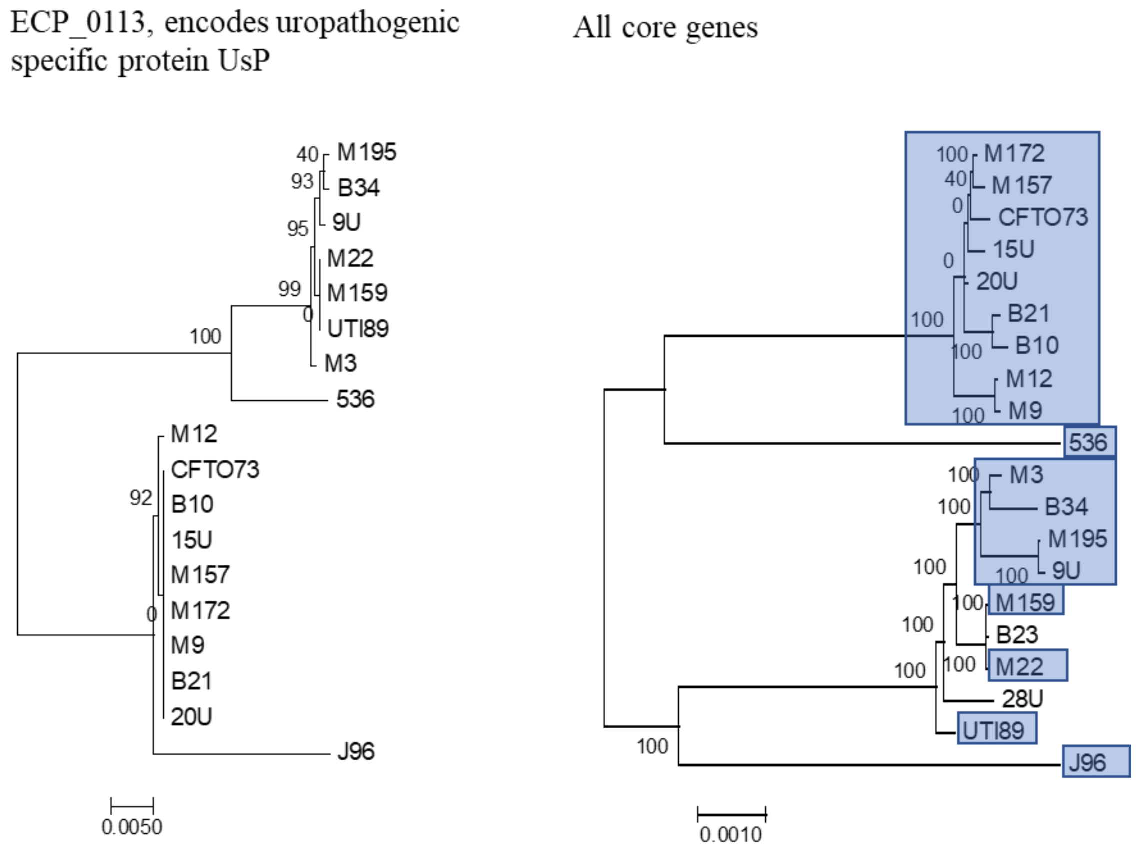
The gene enrichment analysis revealed 8 genes to be significantly associated phenotype groups observed after 9 hours into the phenotype experiment. For group 2A, a cluster of 2 genes (labelled as 2A1 in Table 5.7) were found to be significantly associated with strains in group 2A (8%-100% contractility decrease after 9 hours (control level 7%)). The genes were both predicted to encode proteins of uncharacterised function, which were present in all 19 group 2A1 strains and absent from all non-2A1 strains.

For group 2B (strains exhibiting 20%-100% contractility decrease after 9 hours), 2 genes were reported as present in all 16, present in 2 non-2B strains, and absent in 3 non-2B strains (phenotype association p=0.0075, Fisher's exact test (Fisher 1922, Agresti 1992)). One of these genes encodes an uncharacterised protein, and the other (locus ID ECP_1139 in str. 536) was found to be a homologue of the ExPEC VAF-encoding gene *usp* which encodes the uropathogenic specific protein Usp. Usp is a bacteriocin-like genotoxin which provokes DNA damage to mammalian cells (Nipic et al. 2013). It is also a homologue of a gene (locus ID 04304 in str. F11) significantly associated with survival in a zebrafish bacteraemia model of infection in strain F11 (Wiles et al. 2013). A phylogeny of gene ECP_0113 shows a topology which is broadly consistent with that of the core genome phylogeny (Figure 5.9). In both phylogenies strains 536 and J9 are most closely related to the same cluster of 9 strains but are separated from the by a long branch relative to other branches in the respective trees.

For group 2C (strains exhibiting 40%-100% contractility decrease after 9 hours) a gene present in 8 of 15 phenotype strains was identified which encodes a cell membrane glycotransferase (locus ID RG58_00590 in str. M9, Table 5.7). No genes were found to be significantly associated with phenotype 2D (strains exhibiting a ureter contractility decrease of 60%-100% contractility after 9 hours). For phenotype 2E (strains exhibiting a ureter contractility decrease of 80%-100% contractility after 9 hours) 3 genes were identified as significantly associated, two of which encode uncharacterised proteins. The other gene encodes GTPase binding protein Der (locus ID ECP_3850 str. 536) (Table 5.7).

Table 5.7. Genes significantly associated with phenotypes of group 2: mean contractility decrease of 8% - 100% after 9 hours (group 2A), 20% - 100% after 9 hours (group 2B), 40% - 100% after 9 hours (group 2C), and 80% - 100% after 9 hours (group 2E). No genes were found to be significantly associated with phenotype 2D.

Group	Reference strain	Gene locus in reference strain	Grouped gene cluster	Group strains present in	Group strains absent in	Non-group strains	Non-group strains	Naïve p value	Inferred annotation based on identity to genes of known function	Gene and protein name based on identity to genes of known
2A	536	ECP_1137	2A1	19	0	0	2	0.0048	DUF4222 domain protein; unknown function	-
		ECP_1139		19	0	0	2	0.0048	DUF1317 domain protein; unknown function	-
2B	536	ECP_3864	-	16	0	2	3	0.0075	Uncharacterised protein	-
		ECP_0113	-	16	0	2	3	0.0075	Uropathogenic specific protein	*usP* , UsP
2C	M9	RG58_00590	-	8	6	0	7	0.0180	Cell membrane glycotransferase	-
2E	536	ECP_2042	2E1	11	0	4	6	0.0039	Uncharacterised protein	-
		ECP_2041		11	0	4	6	0.0039	Uncharacterised protein	-
		ECP_3850	-	6	5	0	10	0.0124	GTPase binding protein	*deR* , DeR


ECP_0113, encodes uropathogenic specific protein UsP

All core genes

Figure 5.9. Maximum likelihood phylogenies constructed using an alignment of 1,782 bp of gene ECP_0113 (locus ID str. 536) from 18 strains, which encodes the uropathogenic specific protein Usp. For topological comparison the core genome phylogeny of all 20 UPEC strains is also shown. Percentage bootstrap support values are shown on internal branches. The scale bar indicates the number of substitutions per site represented by the branch length shown.

5.4. Discussion

The purpose of the work in this chapter was to use ureter contractility inhibition phenotype and genome sequence data provided by Dr Rachel Floyd and Professor Craig Winstanley for the 20 tested UPEC strains to provide insights into the genetic and evolutionary basis of *E. coli* ureter contractility inhibition phenotypes. This was approached by conducting analysis to address a hypothesis, an overall aim, and 3 objectives using an up-to-date computer capable of high-throughput genomic analysis. The hypothesis was effectively addressed by analysing strain genome sequence data together with each strain's phenotype at 5-hour and 9-hour time points to determine genes which were significantly associated with phenotypes. By addressing the hypothesis, the first part of the overall aim: to determine genetic differences which are significantly associated with observed differences in ureter contractility inhibition phenotypes across the 20 UPEC strains, was addressed. The results of the analysis are discussed in the following sections, by 5 and 9-hour time point and by devised phenotype group.

Genes associated with mean decrease in ureter contractility by UPEC strains of 8% - 100% after 9 hours (control level 7%) (phenotype group 1A) can be considered those associated with the most severe virulence phenotypes as they contributed to a relatively rapid decrease in ureter contractility. Perhaps the most noteworthy group 1A genes identified are those encoding a haemolysin operon *hlyCABD* (denoted *hly*I) which are fully present in 8 strains and absent in the remaining 13 strains analysed. The 4 genes encode and facilitate the secretion of the pore-forming alpha-haemolysin, a well-known ExPEC VAF (Velasco et al. 2018), that has previously been clinically associated with an increased severity of human urinary tract infections (UTIs) (Marrs et al. 2005). Alpha-haemolysin has been found to initiate the exfoliation of human bladder urothelial cells within the first stages of infection (Smith et al. 2006). Floyd et al. (2010) found this to be a crucial part of UPEC infection. The outer ureter urothelial facet cells were experimentally observed to be exfoliated by strain J96 after 5 hours, which was thought to allow colonisation of the underlying urothelial cells. Furthermore, the transmembrane pores which alpha haemolysin generates have been found to alter calcium oscillations in renal cells which can cause messenger signalling responses in the host cell (Uhlen et al. 1997). Urothelial cells modulate contraction of the underlying smooth muscle (Mastrangelo et al. 2007). Due to this, the host cell modulation of cell signalling by alpha-haemolysin to reduce muscle contractility was inferred by Floyd et al. (2010) to be a potentially important factor in infection of urothelial cells by UPEC. Specifically, secreted alpha haemolysin was proposed to play a role in relaxation of the underlying smooth muscle of urothelial cells via altering signalling in urothelial cells. Alpha-haemolysing pore formation and subsequent urothelial cell lysis or underlying smooth muscle relaxation mediated by cell signalling modulation might therefore be an explanation for the phenotype. The phylogeny of the 9-gene cluster (referred to as 1A1, Figure 5.6) which includes this operon revealed a topology for strains different to that of the core genome phylogeny (Figure 5.3a). The close clustering of strains M17, 20U, M157, and M9 with strain 536 suggests that the gene cluster may have transferred to the lineage leading to the 4 strains from an ancestor of UPEC strain 536 and subsequently evolved in a clonal manner. Also, the close clustering of strain CFT073 to strain relative to others indicates the HT of the cluster between the lineages leading to CFT073 and the cluster of strains including 536 and J96. The close clustering of non-UPEC strains B2 12-1-TI12, TOP382 2, SCB-11, and ECOR 48 with strain UTI89 relative to others compared to the core genome phylogeny indicates the former 4 strains potentially acquired the cluster from an ancestor of UTI89, as they have not been associated with a UPEC phenotype. However, this is speculative as the direction of HT is difficult to infer. These data suggest that the virulence-associated cluster has been horizontally acquired by strains in multiple events and that it might provide the recipient strains with an at least mild ureter-contractility decrease phenotype.

A gene encoding the haemolysin transporter protein ShlB (locus ID ECP_4581 in str. 536) was also present in 11 of 15 phenotype 1A strains. ShlB interacts with cell-bound haemolysin and is crucial for haemolysin secretion (Braun et al. 1991) so its presence adds further support for the role of *hly*I operon genes *hly*CABD in causing phenotype 1A. The phylogeny constructed using ShlB nucleotide sequence (Figure 5.5) shows that strains J96, 536, and UTI89 cluster closer than in the core genome phylogeny, suggesting that at least 2 of the 3 strains may have acquired the gene horizontally. Other strains with the gene appear to have inherited it in a clonal manner as their phylogenetic relationships are consistent to those in the core genome phylogeny.

Inference of haemolysin-mediated damage of urothelial facet cells by UPEC might also be used to directly explain phenotype 1A when urothelial host response is considered. Urothelial cells have been found to show heightened nitric oxide (NO) production as a bactericidal response to infection with varied effectiveness (Poljakic and Pearson 2003). Floyd et al. (2010) suggested that exfoliation of ureter urothelial cells may cause NO to be released and come into contact with the ureter smooth muscle cells. As NO is a muscle relaxant (Hosoki et al. 1997) it was hypothesised that increased alpha-haemolysin expression would result in a gradual NO-mediated decrease in ureter contractility over time. Another gene encoding an oxidoreductase (locus ID c3567 in str. CFT073) associated with phenotype 1A might also fit in with this hypothesis. Nitric oxide is toxic to *E. coli* in high quantities and is typically detoxified through the use of a reductase enzyme (Gomes et al. 2002). As the gene in question has not yet been ascribed a specific oxidoreductase function it can be speculated that the gene may work to detoxify NO the cell is exposed to as a result of the alpha-haemolysin activity.

Another notable gene (locus ID c3564 in str. CFT073) significantly associated with phenotype 1A encodes the KdpD sensor protein that regulates the *kdp*ABC operon (Epstein. 2015). This operon expresses a transport ATPase with a high affinity for potassium which will transport potassium into cells when cell growth is limited through a lack of potassium (Epstein. 2015). The gene may be being used to transport potassium into the cell to promote fast growth and enhance colonisation speed. However, it might also be proposed that the Kdp operon could be used to directly reduce ureter contractility. Floyd et al. (2010) reported that strain J96 infection was associated with potassium leaving ureter muscle cells, resulting in a state of prolonged muscular contraction where depolarisation does not reoccur, and contraction occurs increasingly weakly compared to uninfected tissue. The *kdp*ABC operon could be acting to uptake potassium into the cell with high affinity, hence depriving surrounding muscle cells of potassium. This would prevent effective depolarisation of the muscle cells and gradually reduce contractility over time with increased bacterial growth. The gene was also identified by Wiles et al. (2013) as being required for survival within multiple niches during zebrafish ExPEC infection, which emphasises its potential virulence association.

For phenotype 1B (a mean ureter contractility decrease of 20% - 100% after 5 hours) the only gene with a functional annotation determined as significantly associated with the phenotype encodes a low-affinity zinc transport protein. Zinc is an important metal for *E. coli* growth and is used in enzymatic function, protein synthesis, and replication (Palmer and Skaar 2016). Zinc is also often sequestered in mammalian cells (Jesse et al. 2014) so is likely to have low availability in the ureter and hence it is plausible that a zinc-uptake transport protein would contribute significantly to rates of growth, proliferation, and colonisation of host cells. In a phylogeny of this gene (Figure 5.7), the close clustering of strains M172, 536, and M157 indicate HT of the gene from one lineage to the others. HT can similarly be inferred between the lineages of strains J96 and UTI89. This suggests that the acquisition of this zinc transport gene has potentially contributed to a more significant ureter contractility phenotype.

Phenotype 1C was found exhibited by the 5 strains 536, J96, M157, M159, and M195, which exhibited a mean ureter contractility decrease of 60%-100% after 5 hours in the experiment. Genes significantly associated with this phenotype are potentially more likely to have a virulence association than those associated with phenotypes 1A and 1B, as this is a more severe phenotype. A gene from this group of note was found in strains M157, M159, M195. The gene (locus ID M157_00002 in str. M157) encodes the 16S ribosomal RNA methyltransferase protein KsgA. This has previously been found to have a role in protecting against oxidative stress damage to DNA, which helps to prevent mutations in *E. coli* (Zhang-Akiyama et al. 2009). This might explain the gene's significant association with the phenotype. Zhang-Akiyama et al. (2009) found that the KsgA protein was important for preventing DNA mutation and that deactivation of the gene encoding KsgA increased the rate of spontaneous mutation in *E. coli* strain KSR7. It can be speculated that if strains are encountering oxidative stress such as the NO possibly provided from urothelial cells (Floyd et al. 2010), KsgA may help to maintain growth rates despite NO presence upon urothelial cell invasion. Interestingly strains M159 and M195 lack the oxidoreductase-encoding gene associated with phenotype 1A so this gene might exist to serve a similar function in virulence conditions. The occurrence of the gene in only these strains indicates it likely was acquired via HT from one lineage to the other two, thereby providing the recipient strains with a gene which contributed to phenotype 1C.

The 3 strains 536, J96, and M159 grouped into group 1D are those which exhibited a mean ureter contractility decrease of 80% - 100% after 5 hours. Genes significantly associated with this phenotype can thus be inferred to contribute to the most rapid and significant ureter contractility decrease phenotype. Of note is that a second *hlyCABD* operon was among the highlighted genes (operon denoted *hly*II), present in only strains 536 and J96 (Nagy 2006, Velasco et al. 2018). A more severe phenotype in these two strains is consistent with both carrying a second *hlyCABD* operon. In a study of strain J96, Velasco et al. (2018) determined that *hly*II is regulated by the regulator Zur that is encoded by the gene *zur*, but the first *hlyCABD* operon (*hly*I) is not regulated by it. Zur was also found to only regulate *hly*II in the presence of low levels of zinc. It was hypothesised that the system has evolved to encourage a cycle of rapid haemolysin-mediated cell lysis in the presence of low zinc and subsequent growth from the newly released zinc from the lysed host cells. Low levels of environmental zinc, due to host cell sequestration, are thought to cause *hly*II expression which results in increased cell lysis when combined with the expression of *hly*I. The released zinc is then taken up from lysed host cells, and a subsequent burst in cell growth occurs until environmental zinc is again depleted, after-which the cycle repeats. It can be speculated that in this experiment the system could be acting where urothelial cells are being lysed and zinc is being taken up by the zinc transporter found to be associated with phenotype 1B (which included strain J96). This could explain the significantly higher rate of ureter contractility decrease observed for strains J96 and 536. Similarly, membrane bound genes including str. 536 locus ID ECP_3822 – ECP_3824, encoding putative ABC transporter ATP-binding proteins and a putative periplasmic solute binding protein, were also found only in strains 536 and J96. These three transport-related genes could also be involved in uptake of zinc, but the transporter substrate has not been characterised so it could be any extracellular ion or lysed urothelial cell component which promotes increased growth rate.

Group 2 was comprised of strains which exhibited ureter contractility decrease phenotypes after 9 hours. Eight genes were identified as significantly associated with the 4 phenotypes 2A (ureter contractility decrease of 8% - 100% after hour 9 (control level 7%)), 2B (ureter contractility decrease of 20% - 100% after hour 9), 2C (ureter contractility decrease of 40% - 100% after hour 9, and 2E (ureter contractility decrease of 80% - 100% after hour 9). No genes were found to be associated with group 2D. Only 3 group 2 genes could be functionally annotated. One gene of note (locus ID 00108 in str. 536) was significantly associated with contractility of less than 80% after 9 hours. It encoded the uropathogenic specific protein Usp which is a known ExPEC VAF and was a homologue of a gene significantly associated with UPEC strain F11 survival (locus ID EcF11_4304) in a zebrafish bacteraemia model of infection (Wiles et al. 2013). Usp has been functionally characterised as a bacteriocin-like genotoxin (Nipic et al. 2013), a group of toxins which cause DNA damage. As all 16 phenotype 2A strains possessed the gene and only 2 of 5 non-2A strains have it, it is possible the toxin has a broad effect in contributing to ureter contractility phenotypes through causing urothelial cell death. A phylogeny of the gene (Figure 5.9) indicates that the gene is unlikely to have been horizontally transferred between the 20 UPEC strains. This is evident because the strain relationships in the phylogeny constructed using the gene sequence are consistent with those in the core genome phylogeny. This suggests that the gene was clonally inherited and was deleted in strains 28U and B23.

Another noteworthy gene (locus ID RG58_00590 in str. M9) encodes a cell membrane glycotransferase significantly associated with strains exhibiting a ureter contractility decrease of 40%-100%. Glycotransferases are known to contribute to membrane and cell wall formation (Ha et al. 2000) so it can be speculated that the enzyme could potentially be involved in providing additional membrane structure which aids in the avoidance of host cell defences, but this would require confirmation through experimentation.

The phenotype-associated genes identified in this work indicate that a decrease in the amplitude of ureter contractility is necessary for UPEC colonisation of and proliferation within ureters. However, the phenotypic differences observed in the experiment do not appear to be the result of a single set of genes. The results of this investigation indicated that the range of phenotypes observed in the experiment after 5 hours are the result of the expression of a range of genes shared across different subsets of strains which have been acquired through HT in many cases. The phylogeny constructed based on gene content (Figure 5.3b) indicated that HT is a common phenomenon between strains 536 and J96 but not between the other strains. This is as the other non-536 and J96 strains were consistently positioned in both the gene content and the core genome phylogeny.

It can be inferred that a phenotype is associated with the presence of the 9-gene cluster 1A1, containing genes located adjacently on the chromosome which include the haemolysin operon, an oxidoreductase which potentially reduces oxidative stress, and the *kdpD* operon associated with potassium uptake, an effect which may directly reduce ureter muscle contractility. This was the case for strains 20U, 536, CFT073, J96, M157, M172, M9, UTI89. However, 7 strains with phenotype 1A did not possess the cluster (15U, B10, B21, M159, M22, M3, and M195). Strains 15U, B10, and B21 possessed genes encoding the putative DNA binding protein (locus ID ECP_4584 in str. 536), a gene encodes a haemolysin transporter protein: ShlB (locus ID ECP_4581 in str. 536), and a gene encoding RNA polymerase-binding transcription factor DksA (locus ID ECP_1138 in str. 536) which they shared with strains that possess cluster 1A1. However, of strains with cluster 1A1 and which exhibited phenotype 1B, UTI89 and J96 did not possess the phenotype 1B genes ECP_3019 (locus ID str. 536) and ECP_2043 (locus ID str. 536) but B21 did, and strain CFT073 did not possess ECP_0316 (encoding a low-affinity zinc transport protein). This indicates that a differential presence of genes were responsible for phenotypes 1B. This is similarly the case with phenotype 1C where strains M159 and M195 did not possess phenotype 1C gene encoding putative transcriptional regulator ECP_1146 (locus ID str. 536) present in strains 536, J96, and M157 and strains 536 and J96 did not possess phenotype 1C genes M157_00002 (locus ID in str. M157, encoding the protein KsgA), M157_00003 (locus ID in str. M157, encoding a hydrolase protein), and M157_03942 (locus ID str. M157) but strains M157, M159, M195 do.

On the basis of these results, for strain 20U, phenotype 1A can be inferred to be caused at least in part through use of the *hly*I operon, an oxidoredictase, KdpD encoded by the 9-gene cluster denoted 1A1 and with genes ECP_4584, ECP_4581, and ECP_1138 (locus ID str. 536) which encode a putative DNA-binding protein, haemolysin transporter protein ShlB, and RNA polymerase-binding transcription factor DksA respectively. For strain 15U, phenotype 1A can be inferred to be caused in part by the latter 3 genes only. For strain B21 phenotype 1B can be inferred to have been caused in part by genes ECP_4584, ECP_1138, ECP_3019, and ECP_2043 (locus ID str. 536) encoding the DNA binding protein, DksA, and two proteins with unknown functions. For strain CFT073 the same genes are inferred to result in phenotype 1B, but with the addition of the 9-gene cluster. For strains M172 and UTI89 it is all the same genes as for CFT073 with ECP_0316 (locus ID in str. 536) which encodes a low-affinity zinc transport protein in addition, but without ECP_3019 for strain UTI89. For strains 536 and J96 phenotype 1D can be inferred to be caused in part by the 9-gene cluster denoted 1A1 which included the *hly*I operon, ECP_4584, ECP_4581, ECP_1138, ECP_0316, ECP_1146 (encoding a transcriptional regulator) (locus IDs in str. 536), the 8-gene cluster 1D1 which included the *hly*II operon, the 3-gene cluster 1D2, ECP_3008 (locus ID str. 536, encoding a 50S ribosome-binding GTPase), and ECP_3514 (locus ID str. 536, uncharacterised protein). In addition to these genes, all UPEC strains but 16 phenotype 2B strains may have employed gene ECP_0113 (locus ID str. 536); encoding the uropathogenic specific protein and 8 phenotype 2C strains may have employed gene RG58_00590 (locus ID str. M9) encoding a cell-membrane glycotransferase to carry out decreases in ureter contractility.

To corroborate the inferred link the highlighted genes have in contributing to different ureter contractility decrease phenotypes, direct experimental investigation of the impact of each gene within the rat ureter model would be desirable. This could be carried out through knocking out the genes to cause loss of function and comparing the ureter contractility decrease phenotypes of the strains to the wild type strains without the change. This would provide evidence as to the relative contribution of each gene in causing the range of phenotypes seen. Many of the genes identified have an uncharacterised function. Further work to characterise the molecular function of these proteins would therefore also be an important part of detailing the specific contribution of each gene identified in the investigation in causing ureter contractility decrease phenotypes.

To reflect on the representativeness of the phenotypes exhibited by the 20 UPEC strains employed in this study relative to those in existence in UPEC strains, it can be inferred that the strains were suitably representative for the purposes of addressing the hypothesis, overall aim, and objectives of this study. This is as 33 genes were identified to have a significant association in contributing to a multiple number of independent and overlapping ureter contractility inhibition phenotypes. This number of phenotype-associated genes and phenotypes were obtained as information from strains isolated from a UPEC infections ranging from mild to severe severity across three decades, and this information indicates that there is a case to be made for their representativeness to the range of UPEC strains in existence. However, to quantify their specific representativeness to UPEC in existence, it would be of benefit to repeat the investigation using a greater number of perhaps 100 representative UPEC strains isolated from UPEC infections of ranging severity, across a broader range of locations and time points and compare the phenotype-associated genes and phenotypes in the study to those highlighted in this study.

The results indicate that each of the ureter contractility inhibition phenotypic groups are caused by the specific highlighted genes highlighted. This finding supports the hypothesis. The first objective, to determine if HT has contributed to the phenotypic differences observed across strains was also addressed by the analysis. This was as the results from the phenotype-associated gene phylogenies indicated that horizontal gene transfer of specific genes can be inferred to have contributed to the observed phenotypes across the 20 UPEC strains. Lastly, the second objective, to use phenotype-associated gene information to infer the mechanism of action underlying each observed phenotype pattern was addressed also. This was as in the discussion the proposed mechanisms of action underlying each of the devised phenotype groups were speculated based on the information obtained in analysis about phenotype-associated genes. Addressing the hypothesis and both objectives meant that the overall aim was addressed.

In summary, the hypothesis, overall aim, and objectives were effectively carried out through the conducting of an analysis into the genetic basis to UPEC ureter contractility inhibition phenotypes. The analysis compared the genomic contents of strains across the groups and this highlighted genes which are significantly associated with a range of phenotypes expressed over 9 hours. It can be concluded that multiple phenotypes are the result of the expression of a range of genes shared across different subsets of strains that have been acquired through HT in many cases. Out of the genes identified, of note is a 9-gene island, denoted '1A1' which can be proposed to contribute to the phenotype in 8 strains through the use of the *hly*I operon, a KdpD operon regulator present in 8 strains which allows the uptake of potassium and may directly inhibit ureter contractility, and an oxidoreductase enzyme which may be involved in reducing the impact of oxidative stress defences initiated by host cells. Also of note was the identification of the two previously reported haemolysin operons (denoted *hly*I and *hly*II, Velasco et al. 2018) as part of an 8 gene cluster in UPEC strains 536 and J96, which exhibited the earliest decreases in ureter contractility. The *hly*I and *hly*II operons were inferred to have played a significant role in strains 536 and J96 exhibiting a more severe phenotype than other strains by hour 5. To continue this research, experimental investigation into the impact that each gene highlighted in this work has within the rat ureter model, using mutant knock-out experimentation would also be preferable. It would also be of value to confirm the representativeness of the strain set used in this study by repeating the study using a greater number of perhaps 100 UPEC strains.

Chapter 6: Insights into the evolutionary history of the second *E. coli* type three secretion system (ETT2) and *eip* gene clusters

6.1. Introduction

In this final research chapter, work is presented regarding the evolutionary history of two associated *E. coli* genetic clusters which have potential clinical relevance, ETT2 and *eip*. Previous studies have hinted at them having a complex evolutionary history involving gene loss-of function and deletion mutations within multiple independent lineages since their acquisition in the ancestor of *E. coli* and *E. alberii*. The purpose of the work in this chapter was to revisit the story and present an up-to-date account of the cluster's evolutionary history revealing the extent of its complex evolution and genotypic diversity using the set of 120 phylogenetically diverse *E. coli* strains described in Chapter 3 and 200 other species representatives of the *Enterobacteriaceae* family, and through addressing a hypothesis, an overall aim, and objectives. The result was an account of the manner in which the clusters evolved, their distribution, genotypic variants of each, observed within the major *E. coli* evolutionary lineages, and the types of potentially active VAF–encoding genes present in strains of each of the major *E. coli* phylogenetic lineages.

6.1.1. Type three secretion systems

A well-studied bacterial cell-surface structure which is encoded by highly conserved genes contained in genomic islands (GI), is the type III secretion system (T3SS) (Blocker et al. 2003, Cornells 2000, Hueck 1998). Type III protein secretion is one of eight types of secretion used by Gram-negative bacteria, each associated with a protein structure. (Cornells 2000, Pallen et al. 2003). T3SSs function as a multiprotein molecular syringe termed the "needle complex", which harnesses the hydrolysis of ATP to export effector proteins from the *E. coli* cytoplasm across the inner membrane, periplasmic space, and outer membrane barriers through to targeted eukaryotic cell membrane and into its cytoplasm. (Kenny 2001, Pallen et al. 2003). The structure and genes share some similarity to those of the hook-basal body complex of the bacterial flagellum used for motility, including a polymeric hollow fibre secured to the outer surface, a similarity which allowed early researchers to attribute some function to the T3SS (Aizawa 1996) (Figure 6.1).

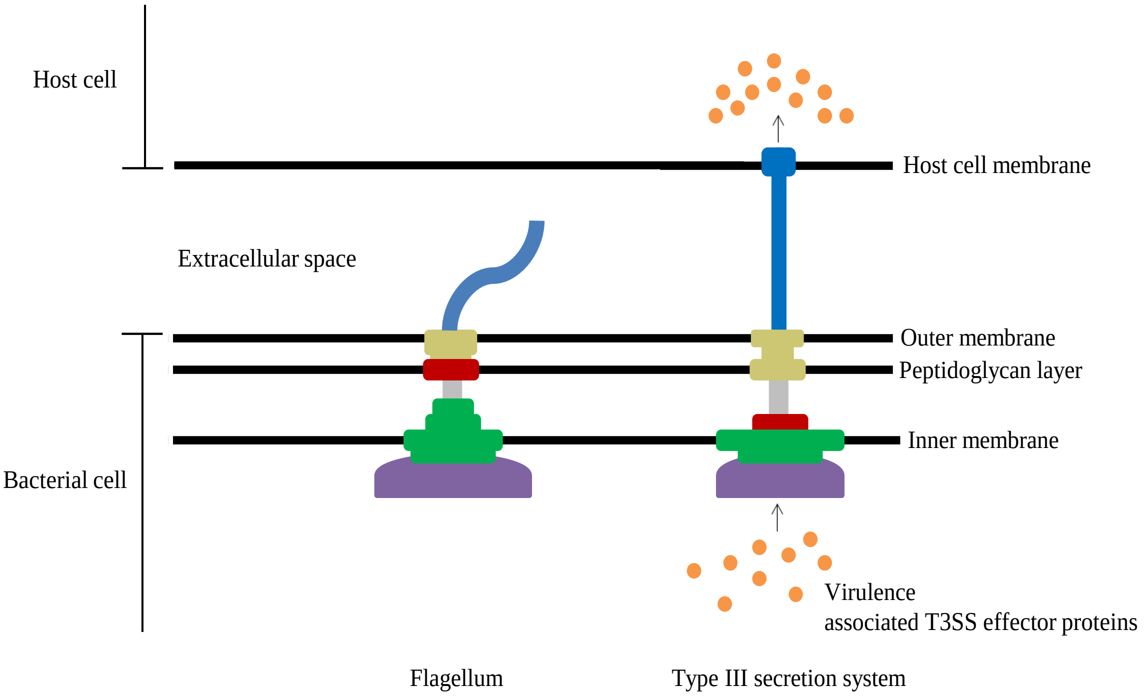


Figure 6.1. Diagram showing homologous components of a flagellar system and type III secretion system structure. Separate structural components are in separate colours, and components which share sequence or functional homology between the flagellar and the T3SS structure are in the same colour. The T3SS is shown transporting virulence associated T3SS effector proteins from within the bacterial cell, across the inner membrane, peptidoglycan layer, the outer membrane, and through the extracellular space and across the host membrane into the host cell (adapted from Figure 1 of Blocker et al. 2003).

Transcription and translation of GI containing genes which encode a T3SS involves hierarchical gene regulation to initiate complex protein interactions and control effector protein secretion (Gerlach et al. 2007). Effector proteins transported into eukaryote cells by T3SS are recognised not just as VAFs but comprise a wide range of functional roles including a contribution towards a symbiotic relationship between bacterium and host, although not within *E. coli* (Blocker et al. 2003). T3SSs within some pathogenic bacteria have been found to be involved with the delivery of VAFs that initiate the onset of disease symptoms (Muller et al. 2001). This has underlined the important role that the T3SS has in some major bacterial diseases (Blocker et al. 2003, Keyser et al. 2008, Muller et al. 2001).

Examples of well-characterised bacterial T3SSs include: the *Mxi-Spa* system from *Shigella*, the YscYop complex of the genus *Yersinia*, the system encoded by the locus for enterocyte effacement (LEE) in attaching and effacing strains of *Citrobacter rodentium*, EPEC and EHEC, and two T3SSs of interest found in *Salmonella enterica* and encoded by PIs named *Salmonella* pathogenicity island 1 and 2 (SPI-1 and SPI-2) (McNamara et al. 1998, McDaniel and Kaper 1997, Jerse et al. 1990, Leimbach et al. 2013). Two medically important *E. coli* pathovars to humans, EHEC and EPEC, possess a T3SS which transports proteins regarded as VAFs (Kenny 2001, McDaniel et al. 1995). Such pathogenic *E. coli* inject these T3SS effector proteins into host intestinal cells which elicit a histopathological effect on them termed attaching-effacing (A/E) lesions (Jarvis et al. 1995, Kenny 2001). These are characterised by functionally damaged microvilli, and pedestals which protrude from the host cell apical membrane which cup bacteria individually, facilitating the attachment of bacteria to the host apical cell surface and allowing the bacterium to grow and proliferate (McDaniel and Kaper 1997, McNamara et al. 1998).

6.1.2. The Second *E. coli* Type Three Secretion System (ETT2)

All genes necessary for EPEC and EHEC A/E lesion formation, including those encoding T3SS structural and secreted proteins, are contained within the LEE, a chromosomally-encoded 35 kb PI (McDaniel and Kaper et al. 1997, Zhang et al. 2004). However, genome sequencing of two strains of EHEC O157:H7 strain EDL933 (Perna et al. 2001) and Sakai (Hayashi et al. 2001) revealed the existence of a GI potentially encoding components of a second T3SS, termed *E. coli* T3SS 2 (ETT2) (with the LEE-encoded system defined as *E. coli* T3SS 1). ETT2 is a 29.9 kb gene cluster integrated at the *yqeG-glyU* tRNA locus, and containing 35 genes, some of which are homologous to the *S. enterica* SPI1 T3SS (Ren et al. 2004) (Figure 6.2).

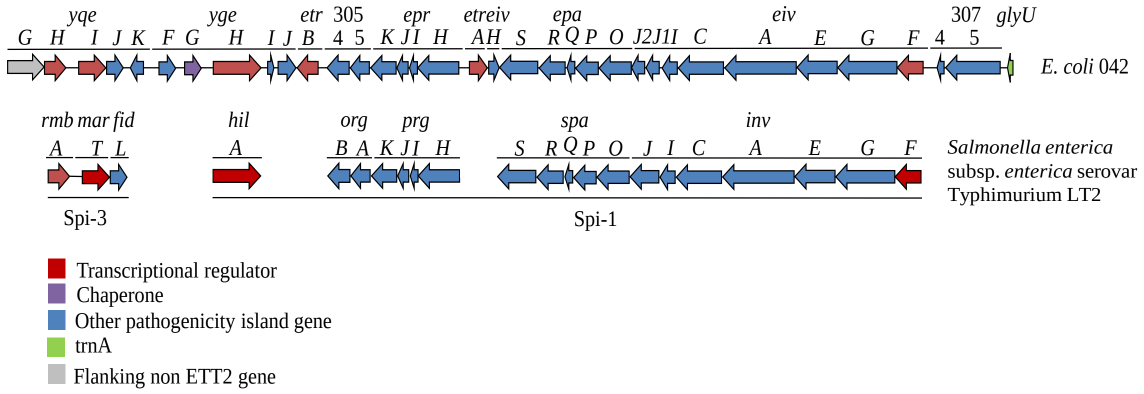


Figure 6.2. The 33 gene ETT2 genomic island structure, as present in *E. coli* strain 042, with structural homologues to the SpI-1 and SpI-3 genomic islands in *Salmonella enterica* subsp. *enterica* serovar Typhimurium LT2. Genes are drawn are arrows pointing downstream or upstream depending on whether presence is on the forward or reverse strand respectively. Genes are coloured by their putative gene products (adapted from Ren et al. 2004, used with permission from the American Society for Microbiology).

ETT2 was first reported as a 14.6 kb insertion relative to the laboratory strain K-12 MG1655 (Perna et al. 2001, Hayashi et al. 2001). However, it was later hypothesised to be a 14.6 kb deletion in K-12 with a reassessment of the cluster length to 17 kb (ECs3714 Sakai nomenclature) by Ren et al. (2004). Fragments of the cluster were found in EHEC and STEC strains with complete absence in non-pathogenic *E. coli* (Makino et al. 2003), but this length was disputed by Hartleib et al. (2003) who reported the boundary to be further upstream (ECs3703 (*rmbA/yqeH*): Sakai genome nomenclature), changing the length to 29.9kb. The position was resolved as here as Hartleib et al. (2003) reported a conserved cluster structure from this boundary point across genomes which exhibited ETT2 genes in an analysis of 245 strains (Hartleib et al. 2003). Ren et al. (2004) studied G+C content, and Sakai ETT2 gene homologues, through using tiling path PCR (TP-PCR) in non-genome sequenced strains to generate a more representative account of the cluster length and evolution. This involved amplifying and sequencing the complete ETT2 chromosomal region and by using primers designed to amplify overlapping ~5 kb fragments of ETT2 spanning a certain region. After this, a long PCR was used on the resulting sample to produce an amplicon of a given ETT2 cluster genotype type. Next, a long PCR for each genotype type using the genotype amplicon was used to detect each genotype in other strains. They included primers flanking absent regions for that genotype type. A short PCR which spanned the ETT2 PI flanking sequences was also used alongside this to screen for complete ETT2 cluster absence. The 29.9 kb length was supported and an accurate account of the cluster in O157:H7 (complete), K-12 (14.6 kb deletion), and CFT073 (complete absence) was defined (Figure 6.3).

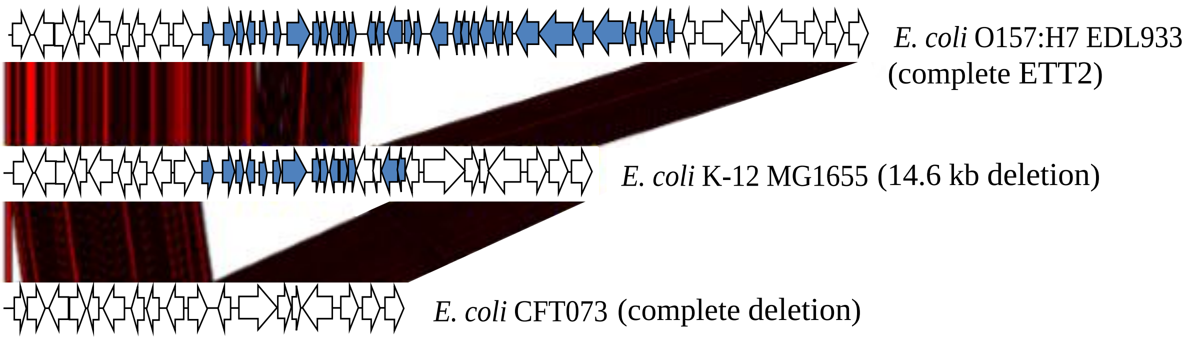

E. coli O157:H7 EDL933 (complete ETT2)
E. coli K-12 MG1655 (14.6 kb deletion)
E. coli CFT073 (complete deletion)

Figure 6.3. A schematic representation comparing the ETT2 gene cluster and flanking genes in the first three ETT2 genotypes discovered (O157:H7 EDL933, K12 MG655, and CFTO73). Genes are drawn are arrows pointing downstream or upstream depending on whether presence is on the forward or reverse strand respectively. ETT2 genes are coloured blue and flanking or non-ETT2 genes are coloured white. K-12 MG655 can be seen to possess around half the cluster genes relative to O157:H7, and CFT073 possesses none relative to O157:H7 (Adapted from Chaudhuri and Pallen 2006).

ETT2 clusters were found to be more widespread in *E. coli* than the LEE and interestingly present in whole or in part in a majority of the 163 representative *E. coli* strains investigated (Ren et al. 2004). A complete ETT2 cluster and those with a characteristic 8.7 kb deletion were the most common genotypes (seen only in the major phylogenetic groups A and B1 (the 8.7 kb deletion was included within the boundaries of the 14.6 kb deletion observed in group A strain K-12 MG1655)). This 8.7 kb genotype was thought to result from deletion of the region during homologous recombination, as 7 bp repeats are found flanking the deleted region in O157:H7. IS1, IS2, and IS3 elements were recorded at the sites of deletions in many strains carrying ETT2 also, suggesting homologous recombination between IS elements after their insertion might be the primary mechanism of ETT2 gene cluster attrition (through deletion) in strains with partial to near complete clusters (Figure 6.4, Ren et al. 2004)

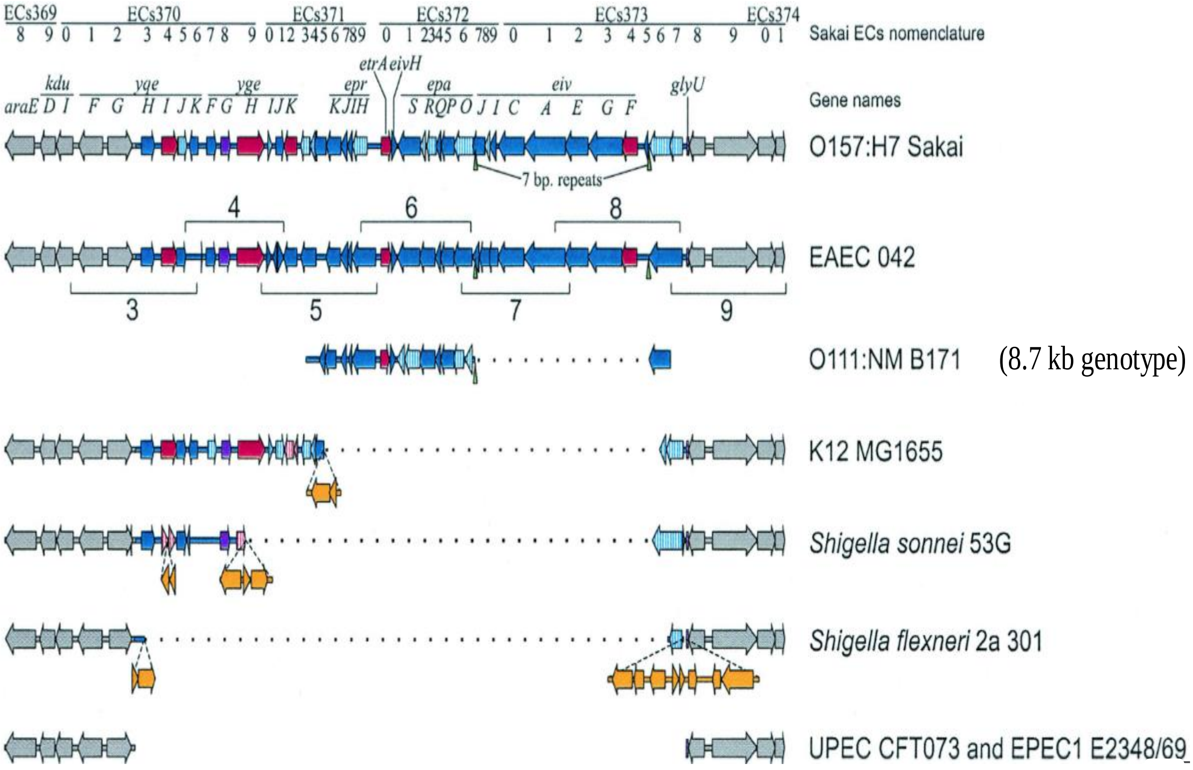


Figure 6.4. Genotypic structural diversity in the ETT2 gene cluster in selected *E. coli* and Shigella strains. ETT2 cluster genes are all coloured non-grey, with homologous genes aligned vertically and deletions indicated by dots. The complete sequence is seen in both strains O157:H7 strain Sakai and 042, strain O111:NM strain B171 represents *E. coli* with the 8.7 kb deletion, and K12 strain MG1655, *S. sonnei* strain 53 G, and *S. flexneri* strain 2a 301 represent varying degrees of sequence deletion. UPEC CFT073 represents complete cluster absence common in UPEC. The 7 bp repeats are highlighted in complete strains, and insertions associated with IS elements relative to the complete cluster are highlighted in yellow with dashed lines (Figure 2 in Ren et al. 2004, used with permission from the American Society for Microbiology).

Ren et al. (2004) rejected the hypothesis that ETT2-associated genes might be a marker of virulence, as complete ETT2 clusters were found in some commensal strains. Multiple deactivating frameshift mutations resulting from indels were also found in many strains in different lineages including EHEC O157:H7, but not EAEC strain 042 (Ren et al. 2004). ETT2 sequence deletions appeared alternatively to be a marker of *E. coli* phylogenetic ancestry, because the cluster is absent from the early diverging B2 group and exhibits varying degrees of mutational attrition and gene loss in all other major phylogenetic lineages including of the same genes in separate events (Ren et al. 2004). This was observed by superimposing the TP-PCR ETT2 genotype data onto a whole-genome phylogeny reconstruction, constructed using neighbor-joining analysis of multilocus enzyme electrophoresis (MLEE) data (Figure 6.5, Ren et al. 2004).

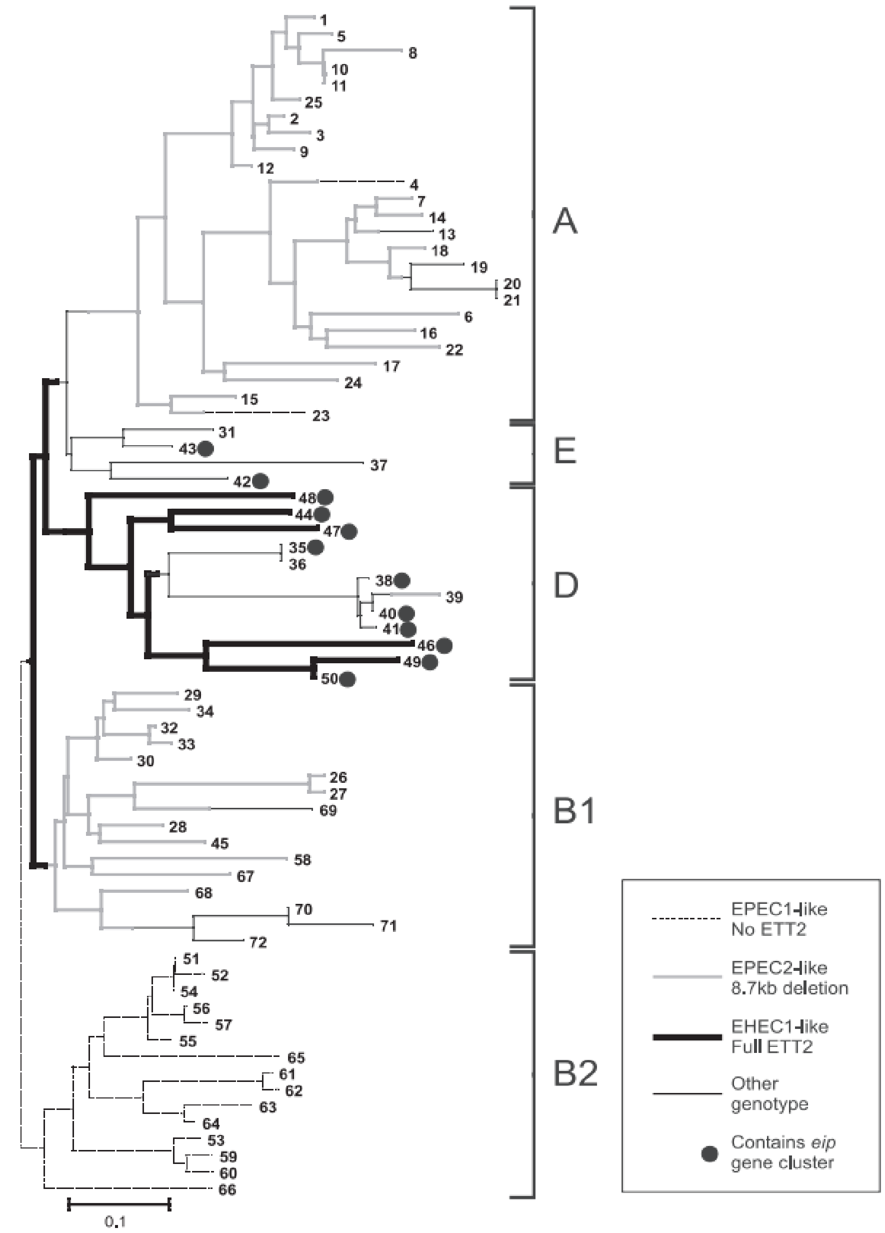


Figure 6.5. TP-PCR ETT2 cluster results of the three main genotypes; absence of any cluster genes (dashes), a characteristic 8.7 kb deletion (grey), or full ETT2 cluster (bold) superimposed phylogenetically onto the whole genome MLEE phylogeny. The major phylogenetic groups are labelled, and circles indicate presence of the *eip* cluster (Figure 5 in Ren et al. 2004, used with permission from the American Society for Microbiology). The scale bar indicates the number of substitutions per site represented by the branch length shown.

Consistent with Ren et al. (2014), a recent study by Wang et al. (2016b) found ETT2 to have undergone widespread attrition when ETT2 genotypes were characterised in 245 APEC strains using eight tiling-path PCR (TP-PCR). Five different genotypic isoforms were identified in 58% of strains including an 042-like ETT2 DNA sequence. Genotypic isoforms included isoform A: 042-like, isoform B: a 4.99 kb deletion, isoform C: a 4.99 kb deletion and a 1.33 kb IS transposase insertion, isoform D: a 5.68 kb deletion, and isoform E: an 8.47 kb deletion. Between phylogenetic groups, four genotypes were found in groups A and B1, and five genotypes were found in groups D and B2 when phylogenetic group assignment was carried out by the triplex Clermont PCR method (Clermont et al., 2000). Nine D1 and 5 B2 strains had 042-like ETT2 sequences, type B and C isoforms were most common in group A, and type E isoforms most common in group B1. Other *E. coli* studies which have reported incomplete ETT2 genotypes include that by Cheng et al. (2012), Huja et al. (2015), Prager et al. (2004). Huja et al. (2015) reported genotypes in five avian pathogenic *E. coli* which included a proposed deletion of the same six-gene region from *eivA* to the end of the island in four strains, when ETT2 gene presence was determined through the creation and use of *E. coli* strain O157 H7 Sakai PCR primers for each gene. Prager et al. (2004) described four new ETT2 genotypes identified from *E. coli* O138:H−, O139:H1, and O147:H6 strains, which were isolated from oedema cases in humans, pigs, and goats. Significant gene deletions were inferred to have occurred since four of five of the islands exhibited deletions of at least five genes. The presence of ETT2 gene homologues was inferred through using PCR primers targeted at the start, middle, and end of the island designed from *yqeH* to tRNA *glyU* using *E. coli* O157:H7 strain Sakai. Cheng et al. (2012) more recently described ten incomplete ETT2 genotypes in an analysis of 168 pathogenic *E. coli* isolated from pigs with colibacillus or cows with mastitis. ETT2 gene presence was carried out through the design and use of 33 PCR primers designed using the genome sequence of *E. coli* strain O157:H7 strain Sakai as well. A notable finding was that nine of the ten genotypes exhibited the same proposed six-gene deletion of genes *eivC* to 3074. Perhaps the most interesting ETT2 genotype finding is that an almost complete 042-

like genotype was identified in 13 of 31 *Escherichia albertii* strains located adjacent to the same trnA *glyU* locus as in *E. coli* strain 042 (Ooka et al. 2015). The presence of ETT2 in *E. albertii* was taken to indicate that the island was acquired by the ancestral lineage of *E. coli* and *E. albertii* prior to the divergence of the two species (Ooka et al. 2015). Table 6.1 shows the ETT2 genotypes reported for each of these studies.

Table 6.1. All published genotypes of ETT2 island genes present in one or more strains, reported in six separate studies. ETT2 gene names and strain groups for each genome are labelled as they occur in *E. coli* strain 042. Dark blue and light blue indicates complete and partial presence of genes respectively.

Reference	Genotype name																																	
Cheng et al. 2012	B	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■							■
	C	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■								■
	D	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■							■
	E	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■									■
	F	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■									
	G	■									■	■	■	■	■	■	■	■	■	■	■	■	■	■	■									■
	H	■	■	■	■	■			■	■	■	■	■	■	■									■	■									■
	I	■	■	■	■	■			■	■					■									■	■	■	■	■	■	■	■	■	■	■
	J	■	■	■	■	■	■		■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■		■							
	K	■	■		■																													
Huja et al. 2015	APEC O2	■	■	■		■	■	■		■	■	■	■	■	■	■	■	■	■	▒	■	■	■	■	■	■	▒	■	■	■	■	■	■	■
	APEC O78 789	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	▒	■	■	■	■	■	▒	▒	▒						
	APEC O78 NC20163	■	■	■	■	■	■	▒		■	▒	■		■	■	■	■	■	▒	▒	■	■	■	■	■	■	▒							
	APEC O78 IMT2125	■	■	■	■	▒	■	▒		■	▒	■	■	■	■	■	■	■	▒	▒	■	■	■	■	■	■	▒							
	APEC O78 Chi7122	■	■	■	■	▒	■	▒		■	▒	▒	■	■	■	■	■	■	▒	▒	■	■	■	■	■	■	▒							
Ooka et al. 2015	*E. albertii* CB9786	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Prager et al. 2004	III	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▒						▒
	VI	■	■	■	■	■	■	■	■	■	■																							▒
	VIII	■	■	■	■	■	■	■																										
	IX	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	▒						
Ren et al. 2004	O15:H7 Sakai	■	■	▒	■	■	■	■	▒	■	■	▒	▒	■	■	■	▒	■	■	■	▒	■	■	▒	■	▒	■	■	■	■	■	■	■	▒
	8.7 Kb	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■									■	■
	O111:NM B171												■	■	■	■	■	■	■	▒	■	■	■	▒										■
	K-12 MG1655	■	■	■	■	▒	■	■	▒	▒	▒	▒	■																					▒
	Shigella sonnei 536	■	▒	■	▒		■	▒																										▒
	S. flexneri 2a 301																																	▒
	CFT073																																	
Wang et al. 2016	B: APDE01	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▒			▒	■	■
	C: APDE099	■	■	■	■	■	■	▒	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▒			▒	■	■
	D: APCE104	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▒				■	■
	E: ADP081	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■								■	■

6.1.3. ETT2 genes and virulence

Several studies have found evidence indicating that some ETT2 genes are implicated in virulence, which is in contrast to the suggestion of Ren et al. (2004) that ETT2 is a disused pathogenicity island. Zhang et al. (2004) found that mutational inactivation of two ETT2 cluster regulatory genes *etrA* and *eivF* each significantly increased secretion of proteins encoded by the LEE (Zhang et al. 2004). This resulted in an increase in adhesive A/E towards human intestinal cells. Studies where microarrays and transcriptional fusions were used show that these two genes negatively affect transcription of the genes within the O157:H7 LEE (Zhang et al. 2004). In a separate EHEC strain O26:H, expression of the *etrA* and *eivF* genes was found to suppress protein secretion under LEE-inducing conditions. These findings suggested that the ETT2 cluster has a regulatory influence on gene expression in the LEE, and provided the first primary evidence of the existence of cross-regulation between T3SSs (Zhang et al. 2004). Also, a recent study by Wang et al. (2017) found that creation of an *etrA* mutant in avian pathogenic *E. coli* strain APCE94 was associated with significantly reduced rates of *etrA* mutant colony population growth and virulence when the *etrA* mutants were injected as a culture into in avian hosts compared to the *etrA* mutant free wild type. Disruption of *etrA* also reduced expression levels for fimbriae-associated genes, which slowed motility and resulted in an increased expression of pro-inflammatory cytokine immune genes when in macrophages, compared to wild type strains. Another ETT2 regulator, *etrB* (previously known as *ygeK*) has also been found to be potentially highly important for enterohaemorrhagic *E. coli* virulence (Luzander et al. 2016). In a study, the product of *etrB* directly interacted with the *ler* regulatory region which activates LEE expression to facilitate A/E lesion formation. Furthermore, *etrB* was found to be regulated by the transcription factor QseA encoded by the gene *qseA*, indicating that *qseA*, *ler*, and *etrB* are part of a regulatory circuit implicated in colonization of host intestinal region (Luzander et al. 2016).

Non-regulator ETT2 genes have also been found to be potentially implicated in pathogenicity: *eivC,* a gene encoding a putative invasion protein homologous to a group of ATPases, has been shown to demonstrate ATPase activity, which is important for T3SS function (Wang et al. 2016). *eivC* mutants were prepared in avian pathogenic *E. coli* strain APEC94. The disruption of *eivC* led to reduced flagella expression and production and increased fimbriae expression and production on the bacteria surface, decreasing overall cell motility (Wang et al. 2016). Also, *eivC* disruption resulted in attenuated virulence of *E. coli* strains and reduced resistance to immune system factors present in avian host blood. All effects of *eivC* disruption were restored once a complete *eivC* gene was complemented into *eivC* mutant strains using transformation. Similarly, a disruption of ETT2 genes *eprHIJK* which encode a putative T3SS inner membrane ring, a lipoprotein precursor, in 11 extraintestinal pathogenic *E. coli* strains isolated from septicemia and meningitis patients, led to significantly reduced virulence when injected into 1-day old chickens compared to *E. coli* with non-disrupted *eprHIJK* genes, which caused 75% mortality (n = 8) after eight days (Ideses et al. 2005). The virulence phenotype was restored after complementation with intact *eprHIJK* genes into the mutants, and all ETT2 genotypes included the same ~5 kb deletion between *eivA* and *eivF* and premature stop codons in several genes (Ideses et al. 2005). A later study also investigated the effect of disrupting the *eprHIJK* gene region in four strains of avian pathogenic *E. coli* of different origins. Huja et al. (2015) found the region was essential for strain survival in host blood and the region was found in all four strains of differing origin. Strains with a disrupted *eprHIJK* gene region grew significantly slower in host blood serum than those without the disruption. Based on this finding, it was suggested that the ETT2 island most likely is not involved in secretion in the strains as premature stop codons had disrupted several genes in each genotype including *ygeF*, *ygeH*, *ygeI*, 3074, *eivH*, *epaS* and deletions occurred in *eivE*, *eivG*, and *eivF*. As *eprHIJK* putatively encode an inner membrane protein ring, it was proposed that intact *eprHIJK* genes at least were involved in enhancing structural properties of the bacterial outer surface which aided in survival within host blood.

6.1.4. The *eip* gene cluster

Ren et al. (2004) identified a second 20.9 kb six-gene cluster referred to as the *eip* gene cluster (Figure 6.6), which was at a separate locus from ETT2, but which also contained homologues of *Salmonella* T3SS SPI-1 genes (Figure 6.6). Short PCRs targeting fragments were spaced throughout the *eip* cluster and applied to the ECOR strains, and presence was mapped onto the MLEE phylogeny (Figure 6.5). The *eip* cluster was only present in strains with complete ETT2 clusters and it was unique to phylogenetic groups D and E. The *eilA* gene encodes a regulator which Sheikh et al. (2006) found increased expression of seven genes: *eip* locus genes *eipB, eipC, eipD, eicA,* and *air*, and ETT2 genes *eivF,* and *eivA* in *E. coli* strain 042. *eilA* mutants were overall less adherent to epithelial cells, and this association led to the hypothesis that the cluster was acquired in association with ETT2 in a duplication event after the ETT2-*eip* system diverged from the T3SS *S. enterica* SPI-1 island after *E. coli* group B2 diverged (Sheikh et al. 2006, Ren et al. 2004).

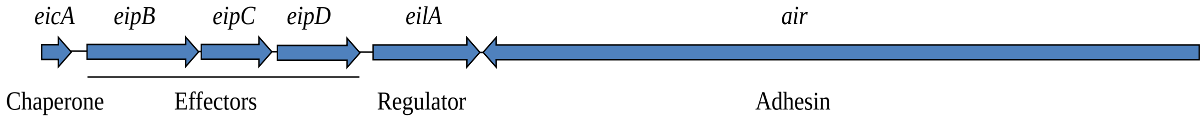


Figure 6.6. The 20.9 kb six-gene *eip* locus, with genes coloured as blue arrows as they appear in *E. coli* strain 042. Arrows pointing right and left indicate genes on the forward and reverse strands, respectively. Functional annotations are provided under each gene (adapted from Sheikh et al. 2006).

6.1.5. Investigation of the ETT2 and *eip* gene cluster

In the interest of confirming results of previous research and further developing a deeper understanding of the evolution of ETT2 and *eip* cluster genotypes and likely phenotypes across *E. coli*, a new investigation of the gene clusters in light of new genome sequence data was necessary. In this chapter, an aim and its hypothesis and three research questions were addressed using three bacterial strain sets in order to carry out the investigation. These included the two previously used phylogenetically diverse strain sets of 100 *E. coli* strains and 20 cryptic clade strains introduced in Chapter 4 and a new set consisting of 200 strain representatives of species from across the *Enterobacteriaceae*. Carrying out this work using these strains was important from an evolutionary biology perspective. Firstly, in terms of understanding when the clusters first appeared and how structure changed through gene truncations, deletions, and how inheritance occurred between *E. coli* phylogenetic group lineages. Any new findings can be added to and used to support or disprove previously hypotheses of ETT2 and *eip* cluster evolution. Secondly, in terms of understanding if core presence of intact functional ETT2 and *eip* cluster genes can be linked with phylogenetic group, so that it may be possible to infer a likely cluster-related pathogenic or non-pathogenic phenotype in a strain based on its phylogenetic group assignment.

6.1.6. Hypothesis

The hypothesis of this chapter was designed to confirm the hypothesis that ETT2 genotypes can be used as a phylogenetic marker as proposed by Ren et al. (2004):

ETT2 cluster genotypes can be used as markers of *E. coli* phylogenetic ancestry.

6.1.7. Aims and objectives

The overall aim was to present an up-to-date account of ETT2 and *eip* cluster evolutionary history and genetic diversity. Principally it was to test the hypothesis ETT2 genotypes can be used as a phylogenetic marker as proposed by Ren et al. (2004) using a diverse set of *E. coli* strains and members of *Enterobacteriaceae*. It was also to test the following objectives designed to confirm the results of previous research, provide a deeper understanding of how the ETT2 and *eip* cluster genotypes evolved, infer any likely ETT2 or *eip* cluster phenotypes based on genotype, and determine if such phenotypes are phylogenetically correlated.

1. Determine if the number of separate ETT2 and *eip* cluster genotypes present across *E. coli* evolutionary lineages is greater than that reported in previous studies.
2. Identify if the evolutionary point of origin for the ETT2 and *eip* clusters is as inferred in previous studies (in the ancestor of *E. coli* and *E. albertii* for the ETT2 cluster and for the *eip* cluster in a duplication event at the same time).
3. Determine likely pathogenic or non-pathogenic ETT2 and *eip* cluster phenotypes for strains based on genotypes and infer if such likely phenotypes can be inferred for strains based on their phylogenetic group assignment.

6.2. Methods

There are no methods specific to this chapter.

6.3. Results

6.3.1. Selection of a non-coli *Escherichia* strain set

To select strains for investigating the presence of ETT2 and *eip* locus genes, the data set of 120 phylogenetically diverse *E. coli* representing groups A-G C-I to C-V which were collated in Chapter 3 were employed. A selection of 200 other bacteria from *Enterobacteriaceae*, was also collated to determine if ETT2 and *eip* locus genes existed in species members of the family other than *E. coli* and *E. albertii* which have been previously not recorded. Genomes from members of *Enterobacteriaceae* were obtained from GenBank and filtered for genome assembly quality. A core gene alignment was obtained using Roary, which was then constructed into a phylogeny using RAxML (representative shown in Figure 6.7). The resulting phylogeny was then filtered for phylogenetic diversity (as for the *E. coli* 100-strain set, Chapter 3) to obtain the 200 strain *Enterobacteriaceae* genome set, which included 74 species from 45 genera (Table 6.2). In a core genome phylogeny of strains from this set sharing a core genome of 1,031,741 bp, the groups *Citrobacter*, *E. albertii*, and *E. fergusonii* were shown to be most closely related to *E. coli* relative to other groups (Figure 6.7). Due to this, 25, 7, and 3 strains from these groups respectively were chosen to be in the 200 *Enterobacteriaceae* strain set (a proportionally larger number than other groups) (Figure 6.7, Figure 6.8, Figure 6.9, Table 6.2).

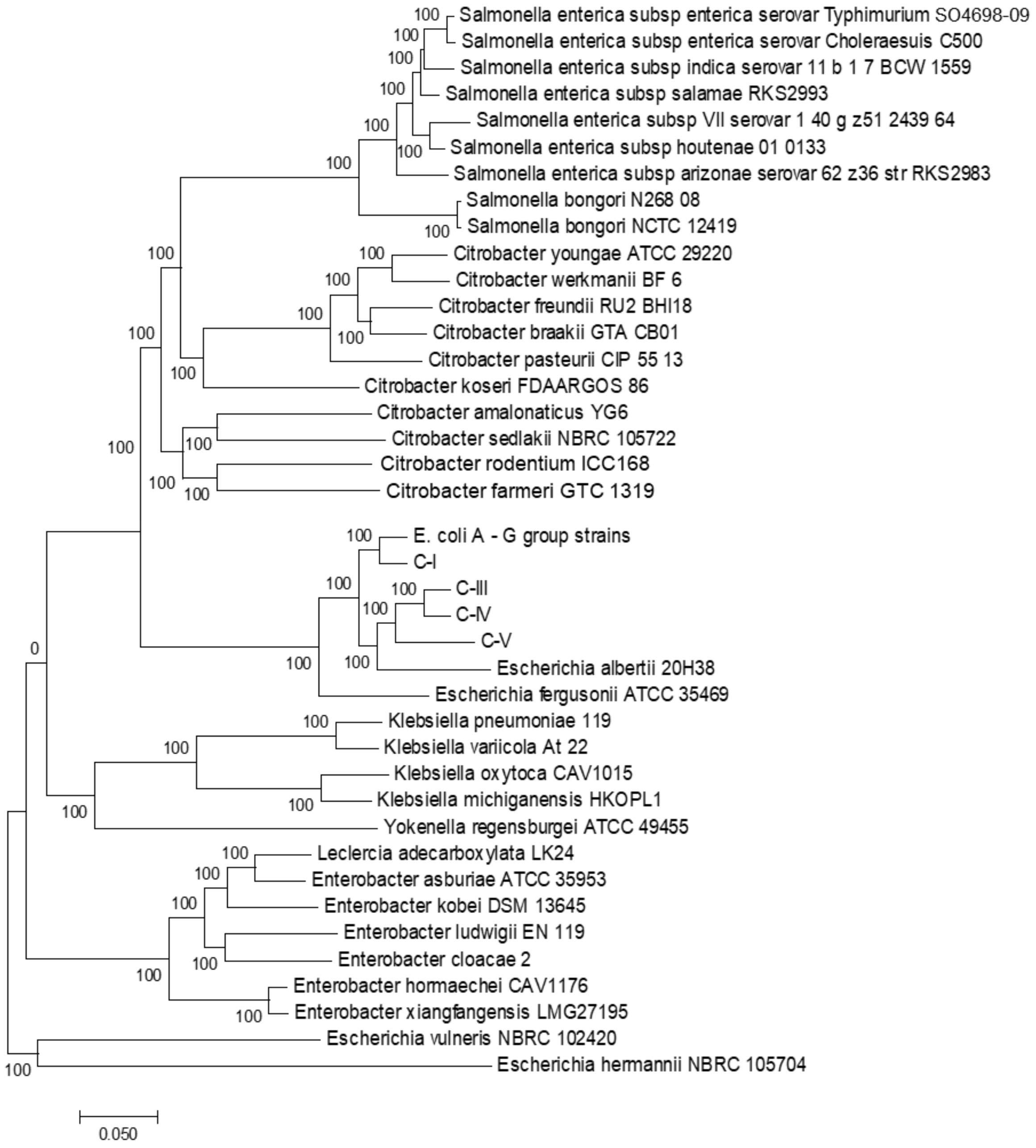

100
Salmonella enterica subsp enterica serovar Typhimurium SO4698-09
100
Salmonella enterica subsp enterica serovar Choleraesuis C500
100
Salmonella enterica subsp indica serovar 11 b 1 7 BCW 1559
100
Salmonella enterica subsp salamae RKS2993
100
Salmonella enterica subsp VII serovar 1 40 g z51 2439 64
100
Salmonella enterica subsp houtenae 01 0133
100
Salmonella enterica subsp arizonae serovar 62 z36 str RKS2983
Salmonella bongori N268 08
100
Salmonella bongori NCTC 12419
100
100
Citrobacter youngae ATCC 29220
100
Citrobacter werkmanii BF 6
100
Citrobacter freundii RU2 BHI18
100
100
Citrobacter braakii GTA CB01
Citrobacter pasteurii CIP 55 13
100
Citrobacter koseri FDAARGOS 86
Citrobacter amalonaticus YG6
100
100
Citrobacter sedlakii NBRC 105722
Citrobacter rodentium ICC168
100
100
Citrobacter farmeri GTC 1319
100
E. coli A - G group strains
C-I
100
100
C-III
100
C-IV
C-V
0
100
100
Escherichia albertii 20H38
Escherichia fergusonii ATCC 35469
100
Klebsiella pneumoniae 119
100
Klebsiella variicola At 22
Klebsiella oxytoca CAV1015
100
Klebsiella michiganensis HKOPL1
100
Yokenella regensburgei ATCC 49455
100
Leclercia adecarboxylata LK24
100
Enterobacter asburiae ATCC 35953
100
Enterobacter kobei DSM 13645
Enterobacter ludwigii EN 119
100
Enterobacter cloacae 2
100
Enterobacter hormaechei CAV1176
100
Enterobacter xiangfangensis LMG27195
Escherichia vulneris NBRC 102420
100
Escherichia hermannii NBRC 105704
0.050

Figure 6.7. Midpoint rooted phylogeny of bacterial genomes in the family *Enterobacteriaceae* most closely related to *E. coli*. The tree was created by RAxML maximum likelihood analysis of a 1,031,741 bp core genome alignment generated using Roary pan genome analysis. The phylogeny shows that *E. albertii, E. fergusonii,* and *Citrobacter* strains are the closest relatives of *E. coli* other than *Salmonella* strains. Percentage bootstrap support values are shown on internal branches. The scale bar indicates the number of substitutions per site represented by the branch length shown.

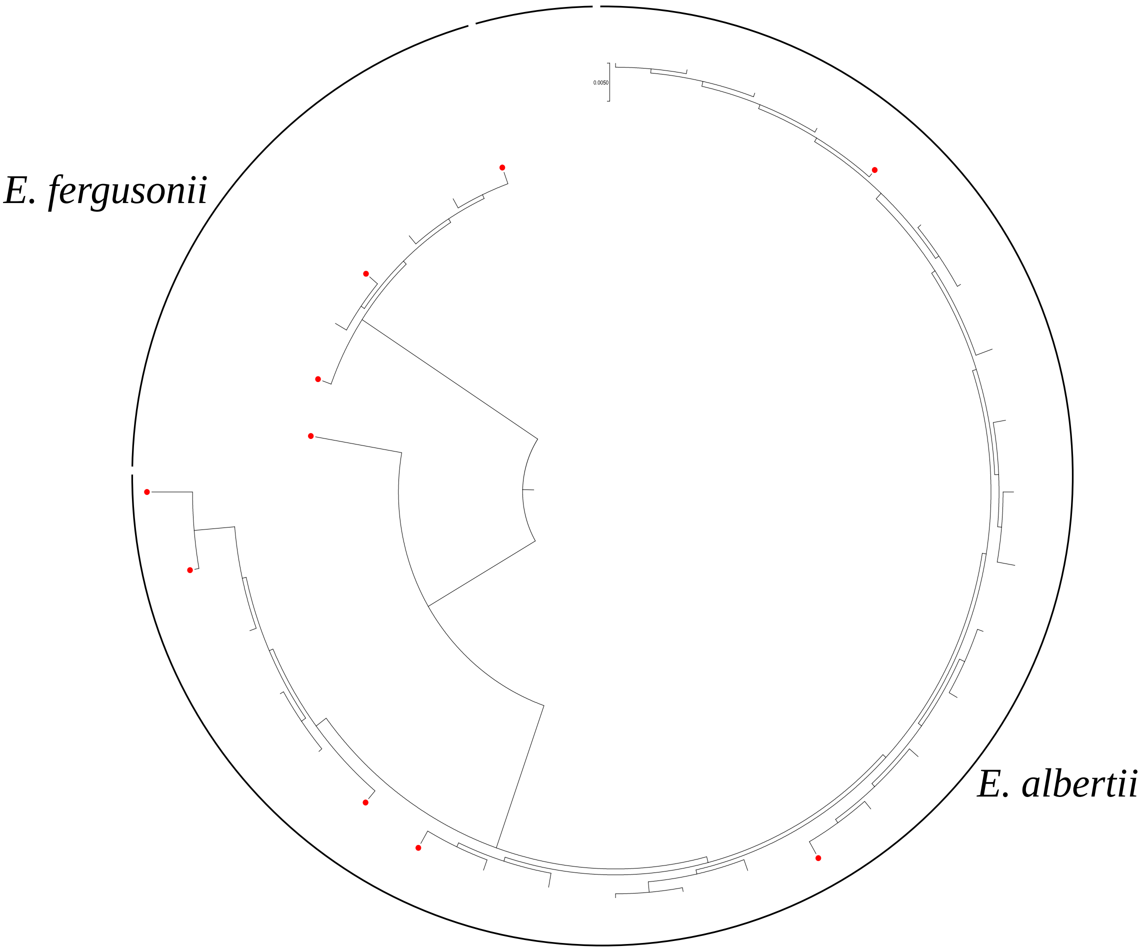


Figure 6.8. Midpoint rooted RAxML maximum likelihood phylogeny of 120 randomly selected core genes present across 35 strains. *E. albertii* and *E. fergusonii* phylogenetic species groups are labelled in the outer ring, with gaps in the ring indicating group borders. 10 strains were chosen from these 35 to be a set of phylogenetically representative non-*coli Escherichia* to check for ETT2 and *eip* locus gene presence and are labelled with a red dot. The scale bar indicates the number of substitutions per site represented by the branch length shown.

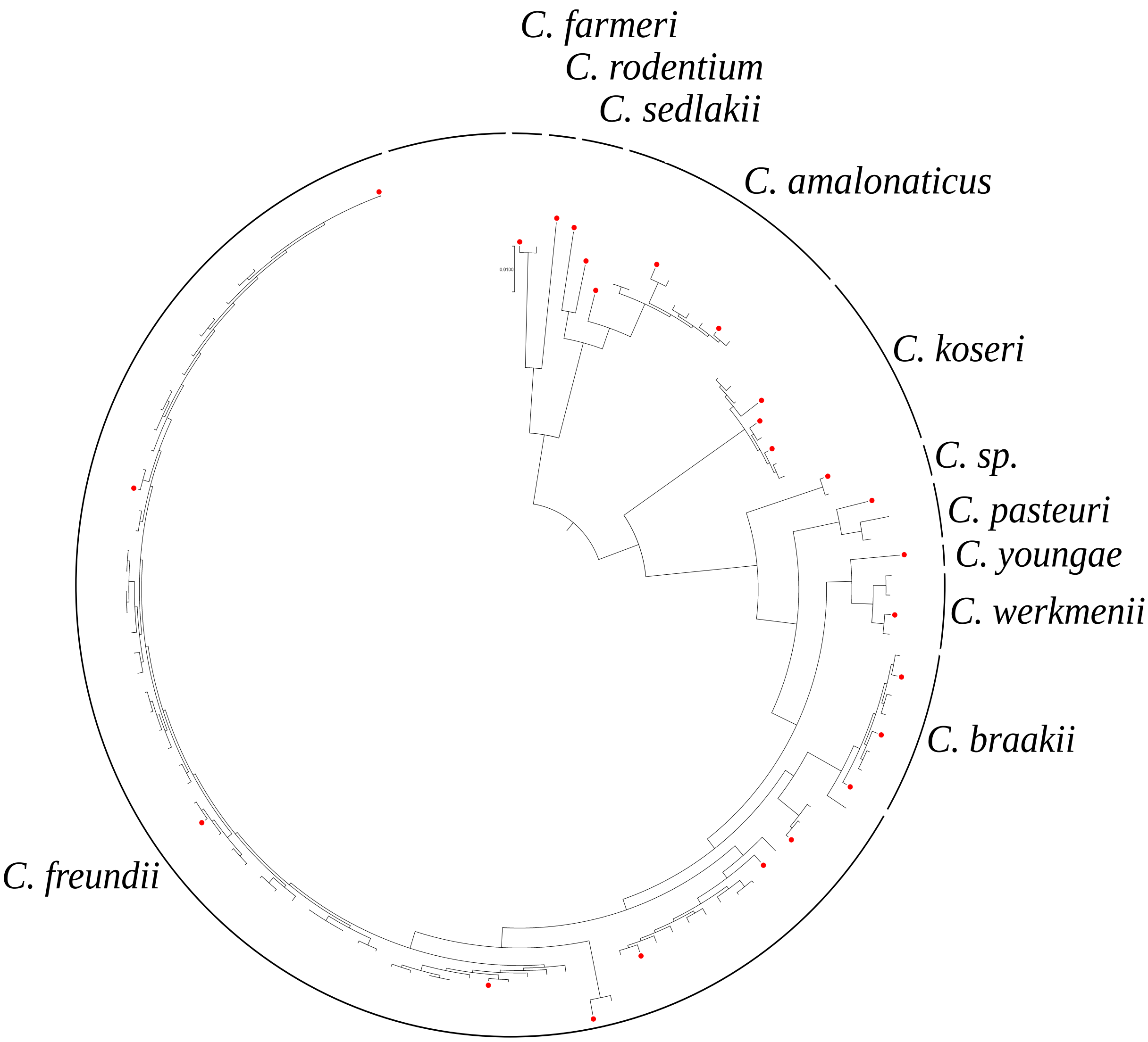


Figure 6.9. Midpoint rooted RAxML maximum likelihood phylogeny of 120 randomly selected core genes present across 117 *Citrobacter* strains obtained from GenBank. Phylogenetic species groups are labelled in the outer ring *C. amalonaticus, C. Braakii, C. farmerii, C. freundii, C. koseri, C. pasteuri, C. rodentium, C. sedlakii, C. werkmenii, C. youngae*, and *C. sp.* (no species name given), with gaps in the ring indicating group borders. 25 strains were chosen from these 117 to be a set of phylogenetically representative *Citrobacter* to check for ETT2 and *eip* locus gene presence and are labelled with a red dot. The scale bar indicates the number of substitutions per site represented by the branch length shown.

Table 6.2. Species names, strain names, and accession numbers for 200 selected genome representatives of the family *Enterobacteriaceae* comprising of 74 species from 45 genera chosen for investigating ETT2 gene presence, including 10 non-coli *Escherichia* and 25 *Citrobacter* previously selected strains. All genomes have an N50 greater than 100 kb.

Species and strain name	Accession
Arsenophonus sp. CB	CP013920
Brenneria goodwinii OBR1	CGIG00000000
Budvicia aquatica DSM 5075 ATCC 35567	ATYS00000000
Cedecea neteri M006	CP009458
Cedecea neteri ND14a	CP009459
Cedecea neteri SSMD04	CP009451
Citrobacter amalonaticus FDAARGOS 122	CP014015
Citrobacter amalonaticus FDAARGOS 166	LORU00000000
Citrobacter amalonaticus L8A	JMQQ00000000
Citrobacter amalonaticus Y19	CP011132
Citrobacter braakii 641 SENT	JUYY00000000
Citrobacter braakii GTA CB04	JRHL00000000
Citrobacter braakii SCC4	MTCP00000000
Citrobacter farmeri GTC 1319	BBMX00000000
Citrobacter freundii 4 7 47CFAA	JH414876
Citrobacter freundii B38	CP016762
Citrobacter freundii BD	CP018810
Citrobacter freundii CAV1321	CP011612
Citrobacter freundii CF04	BDFL00000000
Citrobacter freundii P10159	CP012554
Citrobacter freundii RU2 BHI16	JRTJ00000000
Citrobacter freundii UCI 31	KI929269
Citrobacter koseri 2	LK931336
Citrobacter koseri ATCC BAA 895	CP000822
Citrobacter koseri DNF00568	KQ959519
Citrobacter pasteurii CIP 55 13	CDHL00000000
Citrobacter rodentium ICC168	FN543502
Citrobacter sedlakii NBRC 105722	BBNB00000000
Citrobacter sp. MGH106	KQ089822
Citrobacter werkmanii NBRC 105721	BBMW00000000
Citrobacter youngae ATCC 29220	GG730308
Cronobacter condimenti 1330 LMG 26250	CP012264
Cronobacter dublinensis subsp dublinensis LMG 23823	CP012266
Cronobacter malonaticus CMCC45402	CP006731
Cronobacter sakazakii ES15	CP003312
Cronobacter turicensis 564	CALB00000000
Dickeya chrysanthemi Ech1591	CP001655
Dickeya dadantii 3937	CP002038
Dickeya dianthicola GBBC 2039	CM001838
Dickeya dianthicola RNS04 9	KQ046817
Dickeya paradisiaca Ech703	CP001654
Dickeya solani D s0432 1	AMWE00000000
Dickeya solani IPO 2222 2	CP015137
Dickeya zeae CSL RW192	CM001972
Dickeya zeae DZ2Q	APMV00000000
Edwardsiella anguillarum ET070829	JABY00000000
Edwardsiella anguillarum ET080813	CP006664
Edwardsiella hoshinae ATCC 35051	CP016043
Edwardsiella hoshinae NBRC 105699 ATCC 33379	BAUC00000000
Edwardsiella ictaluri 93 146	CP001600
Edwardsiella piscicida ACC35 1	MPNU00000000
Edwardsiella piscicida C07 087	CP004141
Edwardsiella tarda ASE201307	MBLV00000000
Edwardsiella tarda EIB202	CP001135
Enterobacter asburiae ATCC 35953	CP011863
Enterobacter cloacae 2	CP016906
Enterobacter cloacae LB2	LFLH00000000
Enterobacter hormaechei CAV1176	CP011662
Enterobacter kobei DSM 13645	CP017181
Enterobacter ludwigii EN 119	CP017279
Enterobacter ludwigii NCR3	MCGF00000000
Enterobacter xiangfangensis LMG27195	CP017183
Enterobacter xiangfangensis NS19	LDQK00000000
Erwinia amylovora 01SFR BO	HF560647
Erwinia amylovora ATCC 49946	FN666575
Erwinia billingiae Eb661	FP236843
Erwinia billingiae OSU19 1	LHXI00000000
Erwinia gerundensis EM595	LN907827
Erwinia iniecta B120	JRXE00000000
Erwinia mallotivora BT MARDI	JFHN00000000
Erwinia persicina NBRC 102418	BCTN00000000
Erwinia piriflorinigrans CFBP 5888	CAHS00000000
Erwinia pyrifoliae DSM 12163	FN392235
Erwinia tasmaniensis Et1 99	CU468135
Erwinia teleogrylli SCU B244	KQ947376
Erwinia toletana DAPP PG 735	AOCZ00000000
Erwinia tracheiphila BuffGH	JXNU00000000
Erwinia typographi M043b	JRUQ00000000
Escherichia albertii CB9791	BBVS00000000
Escherichia albertii EC06 170	AP014857
Escherichia albertii HIPH08472	BBVZ00000000
Escherichia albertii KF1	CP007025
Escherichia albertii NIAH Bird 5	BBVP00000000
Escherichia albertii TW07627	CH991901
Escherichia fergusonii ATCC 35469	CU928158
Escherichia fergusonii ECD227	CM001142
Escherichia fergusonii FDAARGOS 170	LORS00000000
Escherichia fergusonii GTA EF03	JZWN00000000
Escherichia hermannii NBRC 105704	BAFF00000000
Escherichia vulneris NBRC 102420	BBMZ00000000
Ewingella americana ATCC 33852	JMPJ00000000
Franconibacter helveticus LMG 23732	AWFX00000000
Hafnia alvei ATCC 13337	JMPK00000000
Hafnia alvei FB1	CP009706
Hafnia alvei FDAARGOS 158	CP014031
Klebsiella michiganensis HKOPL1	CP004887
Klebsiella oxytoca CAV1015	CP017928
Klebsiella pneumoniae 119	LT216436
Klebsiella variicola At 22	CP001891
Kluyvera cryocrescens L2	LGHZ00000000

Table 6.2 continued.

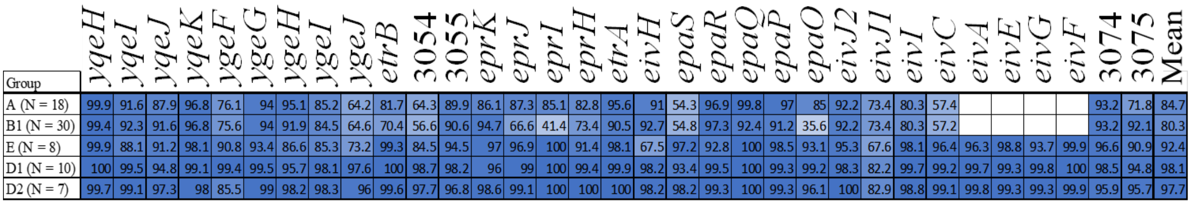
Species and strain name	Accession
Kluyvera intermedia CAV1151	CP011602
Kluyvera intermedia NBRC 102594 ATCC 33110	BCYS00000000
Kosakonia cowanii 888 76	CP019445
Kosakonia oryzae D4	LT799040
Kosakonia radicincitans GXGL 4A	CP015113
Kosakonia sacchari BO 1	CP016337
Leclercia adecarboxylata I1	MUFS00000000
Leclercia adecarboxylata LK24	LDWM00000000
Leminorella grimontii ATCC 33999 DSM 5078 2	JMPN00000000
Leminorella grimontii ATCC 33999 DSM 5078	AUUA00000000
Lonsdalea quercina subsp quercina ATCC 29281	JIBO00000000
Mangrovibacter phragmitis MP23	LYRP00000000
Moellerella wisconsensis ATCC 35017	LGAA00000000
Morganella morganii 340	JQGP00000000
Morganella psychrotolerans GCSL Mp20	LZEY00000000
Pantoea agglomerans 190	JNGC00000000
Pantoea agglomerans C410P1	CP016889
Pantoea ananatis AJ13355	AP012032
Pantoea ananatis AMG521	LMYG00000000
Pantoea anthophila 11 2	JXXL00000000
Pantoea conspicua IF5SW P1	MIZY00000000
Pantoea dispersa SA2	LDSD00000000
Pantoea eucrina Russ	MAYN00000000
Pantoea rwandensis ND04	CP009454
Pantoea septica FF5	CCAQ00000000
Pantoea sesami Si M154	FQWJ00000000
Pantoea stewartii subsp indologenes LMG 2632	JPKO00000000
Pantoea stewartii subsp stewartii DC283	AHIE00000000
Pantoea vagans C9 1	CP002206
Pectobacterium atrosepticum 21A	CP009125
Pectobacterium atrosepticum CFBP 6276	CM001850
Pectobacterium betavasculorum NCPPB 2793	JQHL00000000
Pectobacterium carotovorum subsp actinidiae ICMP 19971	MPUI00000000
Pectobacterium carotovorum subsp brasiliense BC1	CP009769
Pectobacterium carotovorum subsp brasiliense CFIA1001	JPSM00000000
Pectobacterium carotovorum subsp carotovorum BC D6	JUJT00000000
Pectobacterium carotovorum subsp carotovorum PCC21	CP003776
Pectobacterium carotovorum subsp odoriferum NCPPB 3839	JQOG00000000
Pectobacterium parmentieri CFIA1002	JENG00000000
Pectobacterium parmentieri RNS08 42 1A	CP015749
Pectobacterium wasabiae CFBP 3304 2	CP015750
Pectobacterium wasabiae CFBP 3304	AKVS00000000
Photorhabdus asymbiotica ATCC 43949	FM162591
Photorhabdus luminescens ATCC 29999	FMWJ00000000
Photorhabdus luminescens subsp laumondii TTO1	BX470251
Photorhabdus temperata subsp thracensis DSM 15199	CP011104
Pluralibacter gergoviae FB2	CP009450
Pragia fontium DSM 5563 ATCC 49100	FOLW00000000
Proteus hauseri ATCC 700826	LXEV00000000
Proteus mirabilis AOUC 001	CP015347

Species and strain name	Accession
Proteus penneri ATCC 35198	GG662004
Proteus vulgaris ATCC 49132	KN150745
Proteus vulgaris CYPV1	CP012675
Providencia alcalifaciens 205 92	JALD00000000
Providencia burhodogranariea DSM 19968	KB233222
Providencia heimbachae ATCC 35613	LXEW00000000
Providencia rettgeri 729 12	LYBX00000000
Providencia rustigianii DSM 4541	GG703851
Providencia stuartii 50655837	LNHS00000000
Rahnella aquatilis HX2	CP003403
Rahnella aquatilis OV588	JUHL00000000
Raoultella ornithinolytica 10 5246	JH603146
Raoultella ornithinolytica 18	CP012555
Raoultella terrigena NZ133	MUBF00000000
Rouxiella chamberiensis 130333	JRWU00000000
Salmonella bongori N268 08	CP006608
Salmonella bongori NCTC 12419	FR877557
Salmonella enterica subsp arizonae serovar 62 z36 str RKS2983	CP006693
Salmonella enterica subsp enterica serovar Choleraesuis C500	CP007639
Salmonella enterica subsp enterica serovar Typhimurium SO4698-09	LN999997
Salmonella enterica subsp houtenae 01 0133	JWSP00000000
Salmonella enterica subsp indica serovar 11 b 1 7 BCW 1559	MXOA00000000
Salmonella enterica subsp salamae RKS2993	JXTT00000000
Salmonella enterica subsp VII serovar 1 40 g z51 2439 64	MXLH00000000
Serratia fonticola 5I	MQRH00000000
Serratia grimesii A2	JGVP00000000
Serratia liquefaciens 20 SPLY	JVQG00000000
Serratia marcescens CAV1492	CP011642
Serratia plymuthica 4Rx13	CP006250
Serratia proteamaculans 568	CP000826
Serratia symbiotica SCt VLC	FR904230
Shimwellia blattae DSM 4481 NBRC 105725	CP001560
Siccibacter colletis 1383	JMSQ00000000
Siccibacter turicensis LMG 23730	AWFZ00000000
Tatumella morbirosei LMG 23360	CM003276
Tatumella ptyseos ATCC 33301 2	JMPR00000000
Tatumella saanichensis NML 06 3099	ATMI00000000
Xenorhabdus bovienii CS03	FO818637
Xenorhabdus doucetiae FRM16	FO704550
Xenorhabdus eapokensis DL20	MKGQ00000000
Xenorhabdus hominickii ANU1	CP016176
Xenorhabdus mauleonii DSM 17908	FORG00000000
Xenorhabdus nematophila AN6 1	LN681227
Xenorhabdus poinarii G6	FO704551
Xenorhabdus thuongxuanensis 30TX1	MKGR00000000
Yersinia enterocolitica FORC 002 2	CP009456
Yersinia pestis 2944	CP006792
Yersinia pseudotuberculosis EP2	CP009759
Yokenella regensburgei ATCC 43003	JH417859
Yokenella regensburgei ATCC 49455	JMPS00000000

6.3.2. ETT2 homologue presence and absence

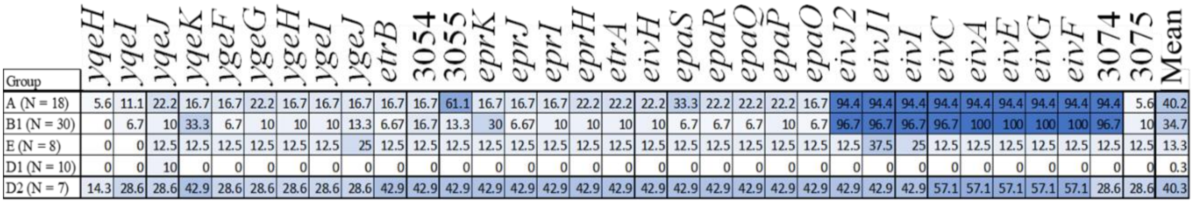
To identify gene homologous to ETT2 genes in the 100 *E. coli* set, the cryptic clade set, and the 200 *Enterobacteriaceae* strain set, BLAST analysis was carried out. To do this a 95% amino acid identity cut-off value was used, as this was determined as the appropriate value to identify orthologues in *E. coli* identified in Chapter 3 and 33 ETT2 locus genes obtained from *E. coli* strain 042 (Genbank accession: FN554766) were used as reference gene sequences (loci *yqeH* – 3075 for ETT2 genes). No length cut-off was used so homologous sequences which were truncated and had partial gene sequence matches could be identified in sequences. Genotypes were then visually inspected using Artemis (Rutherford et al. 2000). 116 ETT2 varied genotypes found in previously unreported strains were determined. 18 A, 30 B1, 8 E, 10 D1, and 7 D2 genotypes from phylogenetic groups A-D2 were described. Among phylogenetic groups A–G, presence of 'complete' *E. coli* 042-like homologues were most prevalent in D1 strains followed by groups E and D2 strains when no multiple deletions had occurred, with a complete deletion of ETT2 genes in groups G and B2 (Table 6.3, Table 6.4, Table 6.5).

Table 6.3. Mean amino acid percentage identity values of complete and partial ETT2 homologues relative to the whole of each *E. coli* strain 042 reference gene for all strains of each of the phylogenetic groups A-G in the 100-strain set.



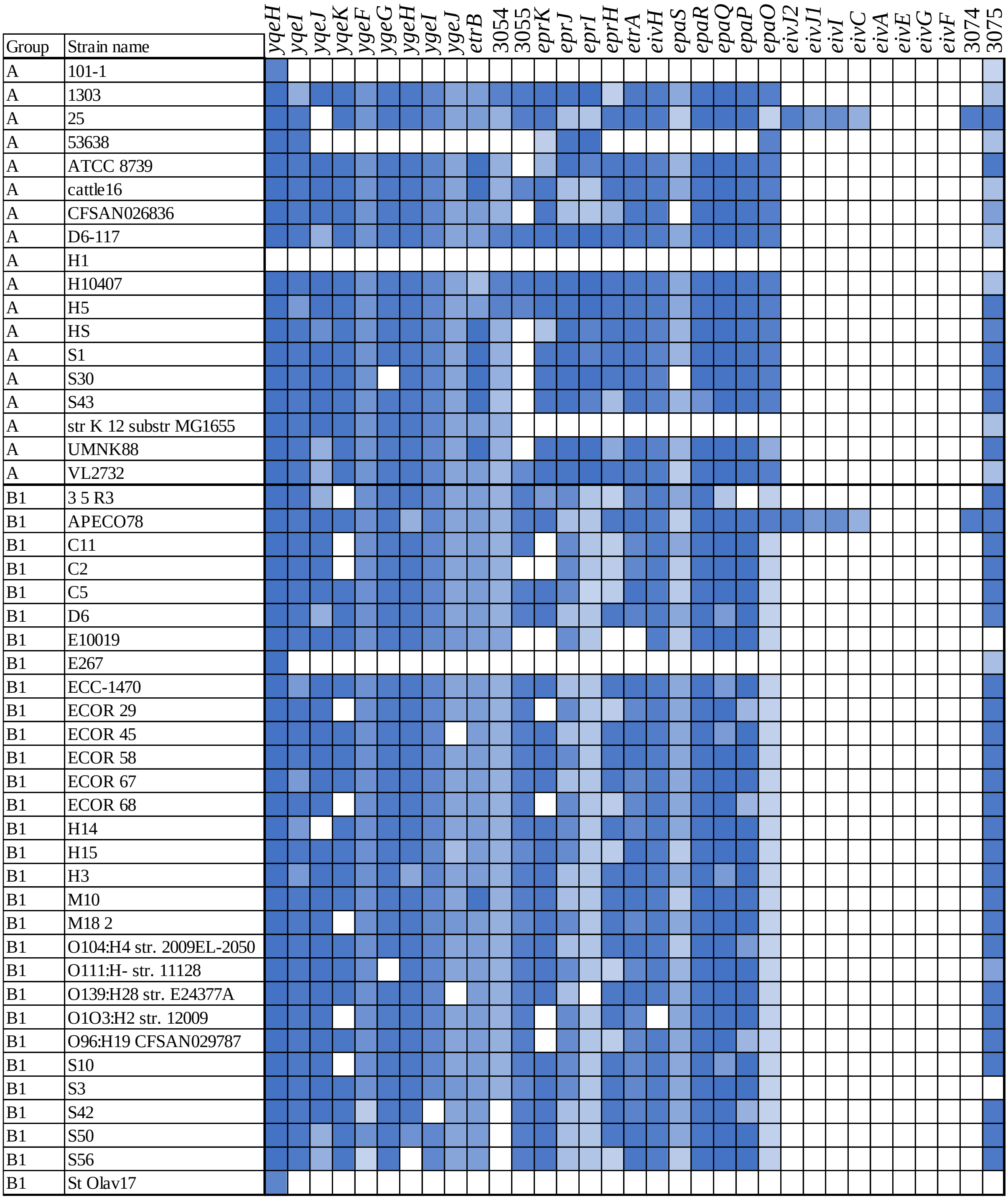
Group	*yqeH*	*yqeI*	*yqeJ*	*yqeK*	*ygeF*	*ygeG*	*ygeH*	*ygeI*	*ygeJ*	*etrB*	3054	3055	*eprK*	*eprJ*	*eprI*	*eprH*	*etrA*	*eivH*	*epaS*	*epaR*	*epaQ*	*epaP*	*epaO*	*eivJ2*	*eivJ1*	*eivI*	*eivC*	*eivA*	*eivE*	*eivG*	*eivF*	3074	3075	Mean
A (N = 18)	99.9	91.6	87.9	96.8	76.1	94	95.1	85.2	64.2	81.7	64.3	89.9	86.1	87.3	85.1	82.8	95.6	91	54.3	96.9	99.8	97	85	92.2	73.4	80.3	57.4					93.2	71.8	84.7
B1 (N = 30)	99.4	92.3	91.6	96.8	75.6	94	91.9	84.5	64.6	70.4	56.6	90.6	94.7	66.6	41.4	73.4	90.5	92.7	54.8	97.3	92.4	91.2	35.6	92.2	73.4	80.3	57.2					93.2	92.1	80.3
E (N = 8)	99.9	88.1	91.2	98.1	90.8	93.4	86.6	85.3	73.2	99.3	84.5	94.5	97	96.9	100	91.4	98.1	67.5	97.2	92.8	100	98.5	93.1	95.3	67.6	98.1	96.4	96.3	98.8	93.7	99.9	96.6	90.9	92.4
D1 (N = 10)	100	99.5	94.8	99.1	99.4	99.5	95.7	98.1	97.6	100	98.7	98.2	96	99	100	99.4	99.9	98.2	93.4	99.5	100	99.3	99.2	98.3	82.2	99.7	99.2	99.7	99.3	99.8	100	98.5	94.8	98.1
D2 (N = 7)	99.7	99.1	97.3	98	85.5	99	98.2	98.3	96	99.6	97.7	96.8	98.6	99.1	100	100	100	98.2	98.2	99.3	100	99.3	96.1	100	82.9	98.8	99.1	99.8	99.3	99.3	99.9	95.9	95.7	97.7

Table 6.4. Percentage of strains from each of the A-G phylogenetic groups from the 100-strain set with inferred deleted ETT2 homologues within their genomes, for each ETT2 gene.



Group	*yqeH*	*yqeI*	*yqeJ*	*yqeK*	*ygeF*	*ygeG*	*ygeH*	*ygeI*	*ygeJ*	*etrB*	3054	3055	*eprK*	*eprJ*	*eprI*	*eprH*	*etrA*	*eivH*	*epaS*	*epaR*	*epaQ*	*epaP*	*epaO*	*eivJ2*	*eivJ1*	*eivI*	*eivC*	*eivA*	*eivE*	*eivG*	*eivF*	3074	3075	Mean
A (N = 18)	5.6	11.1	22.2	16.7	16.7	22.2	16.7	16.7	16.7	16.7	16.7	61.1	16.7	16.7	16.7	22.2	22.2	22.2	33.3	22.2	22.2	22.2	16.7	94.4	94.4	94.4	94.4	94.4	94.4	94.4	94.4	94.4	5.6	40.2
B1 (N = 30)	0	6.7	10	33.3	6.7	10	10	10	13.3	6.67	16.7	13.3	30	6.67	10	10	10	10	6.7	6.7	6.7	10	6.7	96.7	96.7	96.7	96.7	100	100	100	100	96.7	10	34.7
E (N = 8)	0	0	12.5	12.5	12.5	12.5	12.5	12.5	25	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	37.5	25	12.5	12.5	12.5	12.5	12.5	12.5	12.5	13.3
D1 (N = 10)	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3
D2 (N = 7)	14.3	28.6	28.6	42.9	28.6	28.6	28.6	28.6	28.6	42.9	42.9	42.9	42.9	42.9	42.9	42.9	42.9	42.9	42.9	42.9	42.9	42.9	42.9	42.9	42.9	42.9	57.1	57.1	57.1	57.1	57.1	28.6	28.6	40.3

Table 6.5. *E. coli* strain 042 ETT2 PI presence indicated by percentage amino acid identity of 33 complete and partial ETT2 gene sequences within the genome set of 100 phylogenetically diverse *E. coli* from groups A-G and Cryptic clade strains.*



Group	Strain name	*yqeH*	*yqeI*	*yqeJ*	*yqeK*	*ygeF*	*ygeG*	*ygeH*	*ygeI*	*ygeJ*	*etrB*	3054	3055	*eprK*	*eprJ*	*eprI*	*eprH*	*etrA*	*eivH*	*epaS*	*epaR*	*epaQ*	*epaP*	*epaO*	*eivJ2*	*eivJ1*	*eivI*	*eivC*	*eivA*	*eivE*	*eivG*	*eivF*	3074	3075
A	101-1	■	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■
A	1303	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
A	25	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	■	■
A	53638	■	■	□	□	□	□	□	□	□	□	□	□	■	■	■	□	□	□	□	□	□	□	■	□	□	□	□	□	□	□	□	□	■
A	ATCC 8739	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
A	cattle16	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
A	CFSAN026836	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	□	■	■	■	■	□	□	□	□	□	□	□	□	□	■
A	D6-117	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
A	H1	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
A	H10407	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
A	H5	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
A	HS	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
A	S1	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
A	S30	■	■	■	■	■	□	■	■	■	■	■	□	■	■	■	■	■	■	□	■	■	■	■	□	□	□	□	□	□	□	□	□	■
A	S43	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
A	str K 12 substr MG1655	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■
A	UMNK88	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
A	VL2732	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	3 5 R3	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	□	□	□	□	□	□	□	□	□	■
B1	APECO78	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	■	■
B1	C11	■	■	■	□	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	C2	■	■	■	□	■	■	■	■	■	■	■	□	□	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	C5	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	D6	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	E10019	■	■	■	■	■	■	■	■	■	■	■	□	□	■	■	□	□	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	□
B1	E267	■	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■
B1	ECC-1470	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	ECOR 29	■	■	■	□	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	ECOR 45	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	ECOR 58	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	ECOR 67	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	ECOR 68	■	■	■	□	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	H14	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	H15	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	H3	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	M10	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	M18 2	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	O104:H4 str. 2009EL-2050	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	O111:H- str. 11128	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	O139:H28 str. E24377A	■	■	■	■	■	■	■	■	□	■	■	■	■	■	□	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	O1O3:H2 str. 12009	■	■	■	□	■	■	■	■	■	■	■	■	□	■	■	■	■	□	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	O96:H19 CFSAN029787	■	■	■	□	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	S10	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	S3	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	□
B1	S42	■	■	■	■	■	■	■	□	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	S50	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	S56	■	■	■	■	■	■	□	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	■
B1	St Olav17	■	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□

* ETT2 gene names and strain groups for each genome are labelled. Darkness of blue (scale: ≤ 30% identity (white) to 100% identity (dark blue)) indicates increasing percentage identity of each ETT2 reference gene to a homologous gene present in each genome. Only genes with > 30% identity and length relative to ETT2 reference genes are shown.

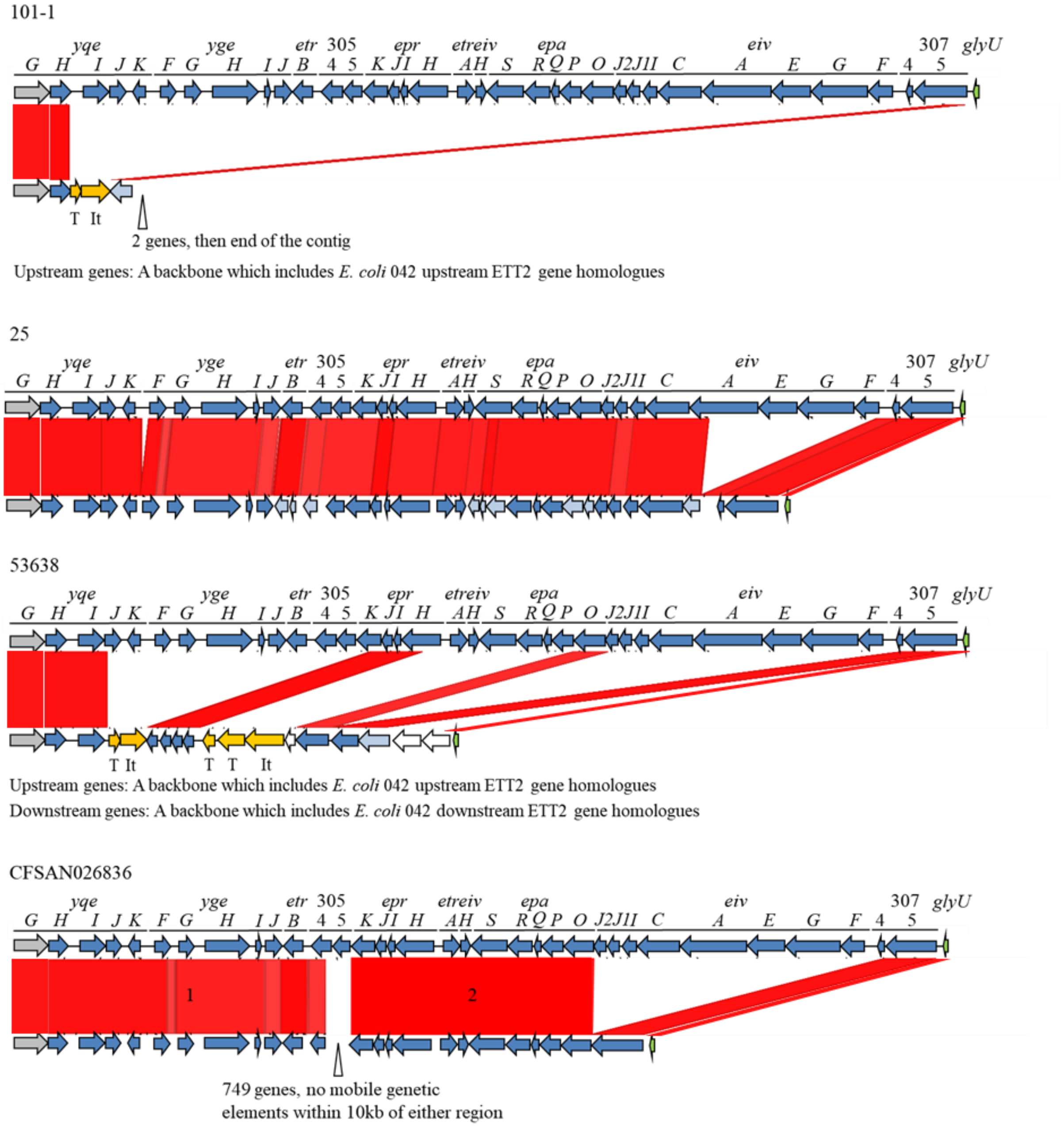
Table 6.5 continued.

Group	Strain name	*yqeH*	*yqeI*	*yqeJ*	*yqeK*	*ygeF*	*ygeG*	*ygeH*	*ygeI*	*ygeJ*	*etrB*	3054	3055	*eprK*	*eprJ*	*eprI*	*eprH*	*etrA*	*eivH*	*epaS*	*epaR*	*epaQ*	*epaP*	*epaO*	*eivJ2*	*eivJ1*	*eivI*	*eivC*	*eivA*	*eivE*	*eivG*	*eivF*	3074	3075
E	400654	■	■	■	■	■	■	░	■	□	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■
E	AF85	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■
E	B185	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■
E	C161 11	■	░	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	░	□	□	□	□	▓	▓	■	■	■	■	■
E	D6-113	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■
E	O157:H16 str. Santai	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□
E	O157:H7 str. Sakai	■	■	■	■	■	■	■	■	▓	■	░	■	■	■	■	░	■	□	■	░	■	■	■	■	□	■	■	■	■	■	■	■	▓
E	O169:H41 str. F9792	■	■	░	■	■	■	■	■	▓	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■
D1	042	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
D1	B354	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■
D1	C1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■
D1	C4	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■
D1	EC2	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■
D1	ECOR 48	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■
D1	TA255	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	▓
D1	TA280	■	■	░	■	■	■	░	■	■	■	■	■	░	■	■	■	■	■	░	■	■	■	■	■	▓	■	■	■	■	■	■	■	■
D1	UMN026	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■
D1	upec 213	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■
D2	24 1 R1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■
D2	BIDMC 19C	■	■	▓	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■
D2	HVH 87 4	■	■	■	■	░	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	□	□	□	□	□	■	■
D2	IAI39	■	■	■	□	▓	■	■	■	■	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■
D2	SMS35	■	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
D2	Swine 65	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
D2	UCI 57	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	■	■	■

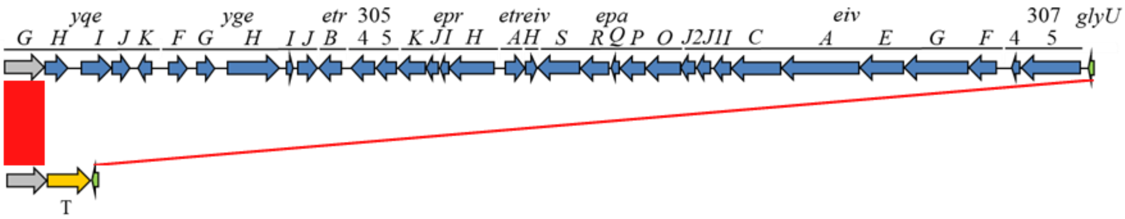
Premature stop codons (truncations) were determined for homologues based on the presence of a reduced identity value for them relative to the full-length *E. coli* strain 042 reference ETT2 gene sequences in BLAST, compared to that of full-length homologue sequences, followed by manual inspection sequences. Premature stop codons were determined to be most prevalent in A and B1 strains of A-G groups relative to other groups, indicated by their lowest mean percentage homologue identity values relative to full length *E. coli* strain 042 reference gene sequences (Table 6.3). Genes *ygeF*, *ygeJ*, 3054, *epaS*, *eivJ1*, *eivC*, and 3075 were the most frequently truncated genes in group A strains, and *ygeF*, *ygeJ*, *etrB*, 3054, *eprJ*, *eprI*, *eprH*, *epaS*, *epaI*, *eivC*, and 3075 the most frequently truncated in group B1 (evidenced by ≤ 80% identity (Table 6.3, Table 6.5)). The genes *ygeJ* and *epaS* were truncated in all group A strains, *ygeJ*, *etrB*, 3054, *eprI*, and *epaS* are truncated in all B1 strains (Table 6.3, Table 6.5), and *eivJ1* was truncated in all group E strains (Table 6.3, Table 6.5). Groups A and B1 strains also exhibited a full homologue absence of all genes in at least one strain, apart from *yqeH* in group B1 (Table 6.5). An 8.7 kb 9-gene *eivJ2* to 3074 region was also absent in all but one strain in each of groups A and B1 (strains 25 and APECO78 respectively), which lacked the four homologue region *eivA* to *eivF* within the 8.7 kb region (Table 6.5). Other absences of 2 or more genes were present in 6 A and 12 B1 strains, and of 4 or more genes were present in 4 A and 3 B1 strains (Table 6.5). One group A strain exhibited a full ETT2 island deletion (H1), and 1 A and 2 B1 strains exhibited complete absences other than *yqeH* (Table 6.5). Of the group A and B1 deletion genotypes (Figure 6.10), 4 group A and 3 group B1 strains exhibited different genotypes with an insertion sequence annotated as transposon (6 and 1 across A and B1 genotypes respectively) and integrase (3 and 2 across A and B1 genotypes respectively) encoding genes in different locations (Figure 6.10). Group B1 additionally exhibited 2 genotypes with 5 phage-related genes and 2 with 2 insertion sequences within 10 kb of the homologues (Figure 6.10). 1 group D1 genotype exhibited a complete absence of ETT2 homologues (Table 6.5, Figure 6.10). 4 group E and D2 strains each exhibited deletion of 2 or more genes, and 2 E and 4 D2 strains exhibited homologue deletions of 4 or more (Table 6.5, Figure 6.10). Of the group E and D2 genotypes deletions of 4 or more, 1 D2 strain exhibited a full homologue deletion (Table 6.5, Figure 6.10), and 2 and 3 genes encoding transposons were found across E and D2 genotypes respectively with an integrase-encoding gene within 10 kb of ETT2 homologues in 1 group E strain (Figure 6.10). Where homologues were present in E, D1, and D2 strains, they were not truncated to the point of exhibiting < 95% identity in a BLAST against the full length strain 042 reference genes in D2 and D1 strains, and only exhibited < 95% identity in a BLAST against the full length strain 042 reference genes with 3 homologues amongst group E strains (*ygeJ*, *eivH*, and *eivJ1*) (Table 6.3, Table 6.5, Figure 6.10). All strains of phylogenetic groups G and B2 exhibited a full ETT2 homologue deletion.

All homologues of genotypes across groups A-D2 were located within 10 kb of one another apart from in two strains (CFSAN026836 (group A) and E110019 (group B1) (Figure 6.10).

E. coli phylogenetic group A


101-1
yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K J I H A H S R Q P O J2 J1 I C A E G F 4 5
T It
2 genes, then end of the contig
Upstream genes: A backbone which includes E. coli 042 upstream ETT2 gene homologues
25
yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K J I H A H S R Q P O J2 J1 I C A E G F 4 5
53638
yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K J I H A H S R Q P O J2 J1 I C A E G F 4 5
T It T T It
Upstream genes: A backbone which includes E. coli 042 upstream ETT2 gene homologues
Downstream genes: A backbone which includes E. coli 042 downstream ETT2 gene homologues
CFSAN026836
yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K J I H A H S R Q P O J2 J1 I C A E G F 4 5
1
2
749 genes, no mobile genetic elements within 10kb of either region

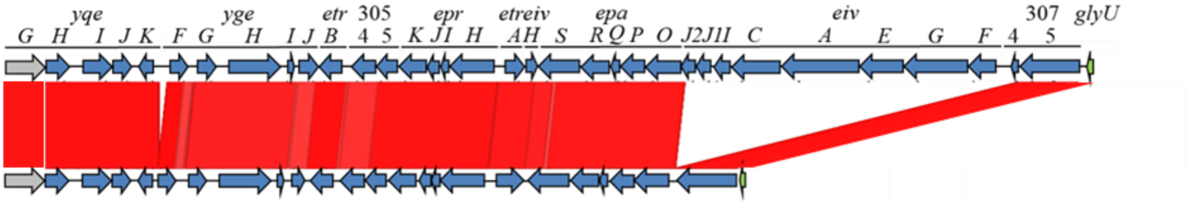
H1


yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K JI H AH S RQ P O J2J1I C A E G F 4 5
T

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: A backbone which includes *E. coli* 042 downstream ETT2 gene homologues

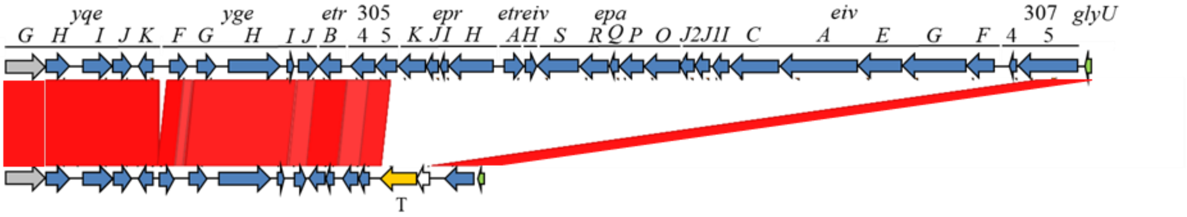
H5


yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K JI H AH S RQ P O J2J1I C A E G F 4 5

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: A backbone which includes *E. coli* 042 downstream ETT2 gene homologues

Str K-12 MG1655

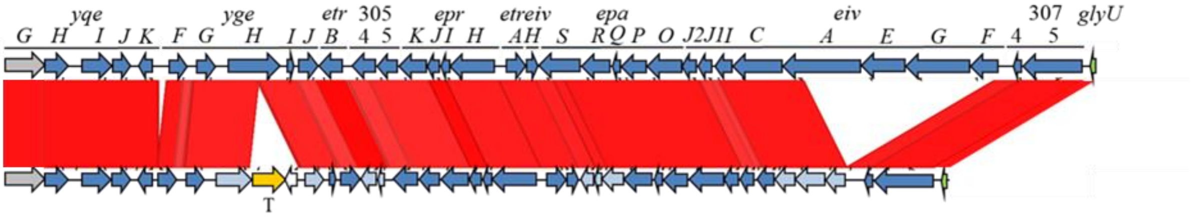

yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K JI H AH S RQ P O J2J1I C A E G F 4 5
T

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: A backbone which includes *E. coli* 042 downstream ETT2 gene homologues

E. coli phylogenetic group B1

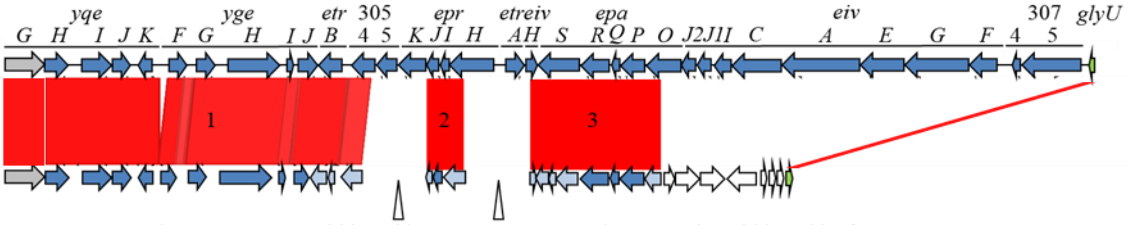
APECO78


yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K JI H AH S RQ P O J2J1I C A E G F 4 5
T

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: A backbone which includes *E. coli* 042 downstream ETT2 gene homologues

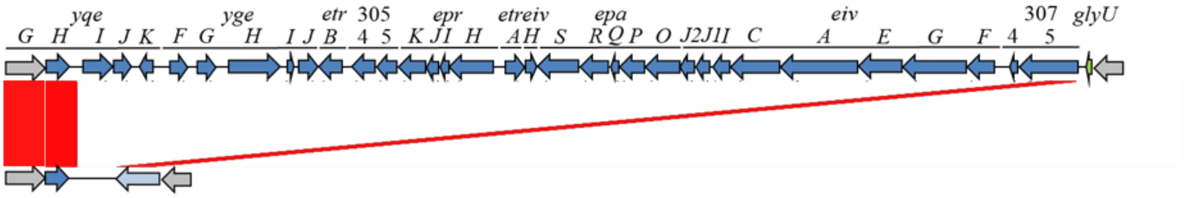
E110019


yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K JI H AH S RQ P O J2J1I C A E G F 4 5
1
2
3

499 genes, 1 integrase present within 10kb of region 1, insertion element and type IV secretion proteins within 10kb of region 2

94 genes, 2 phage proteins within 10kb of region 2, no mobile genetic elements within 10kb of region 3

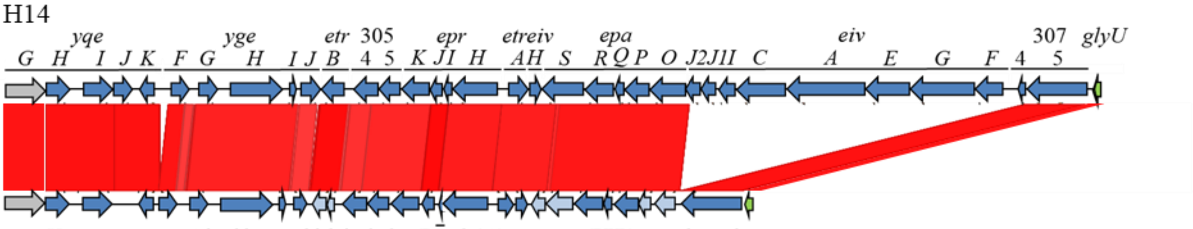
E267


yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K JI H AH S RQP O J2J1I C A E G F 4 5

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: 2 *E. coli* 042 downstream ETT2 gene homologues, then a backbone without *E. coli* 042 downstream ETT2 gene homologues

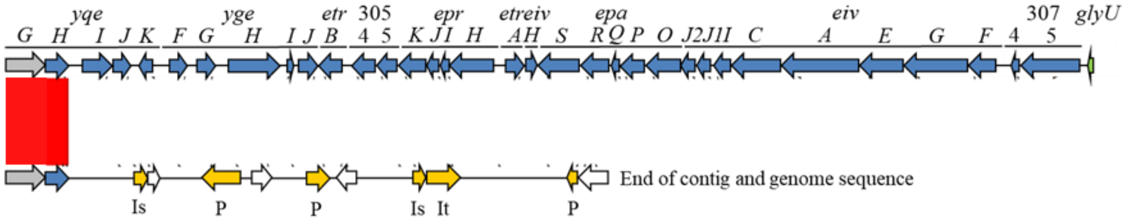
H14


yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K JI H AH S RQP O J2J1I C A E G F 4 5

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: A backbone which includes *E. coli* 042 downstream ETT2 gene homologues

St Olav 17

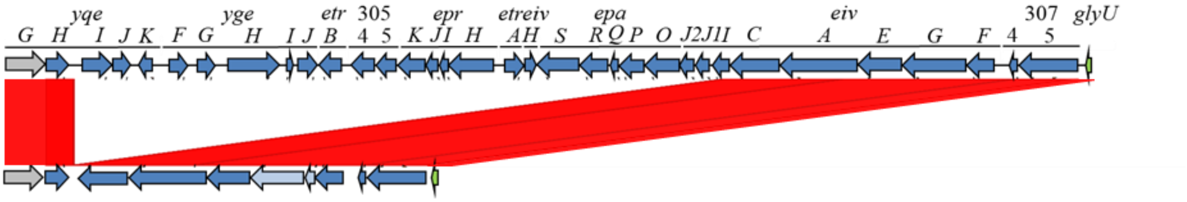

yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K JI H AH S RQP O J2J1I C A E G F 4 5
Is
P
P
Is It
P

End of contig and genome sequence

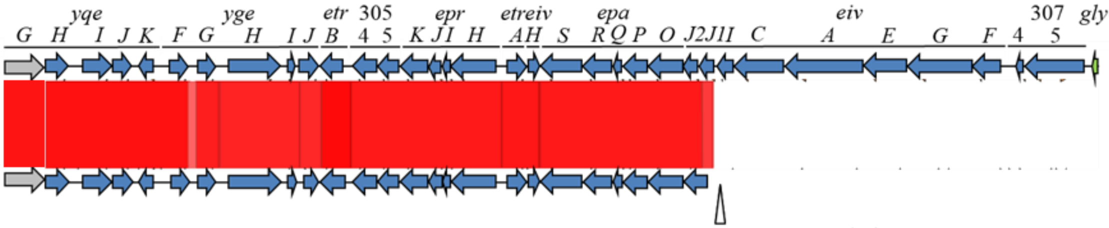
Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

E. coli phylogenetic group E

C161 11


yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K JI H AH S RQP O J2J1I C A E G F 4 5

O157 H7 Santai


yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K JI H AH S RQP O J2J1I C A E G F 4 5

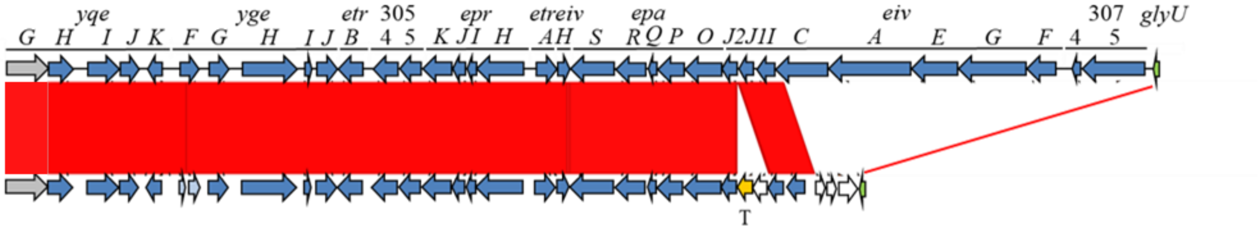
2 transposons and 1 integrase within 10kb of homologues

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: No glyU, a backbone without *E. coli* 042 downstream ETT2 gene homologues

E. coli phylogenetic group D2

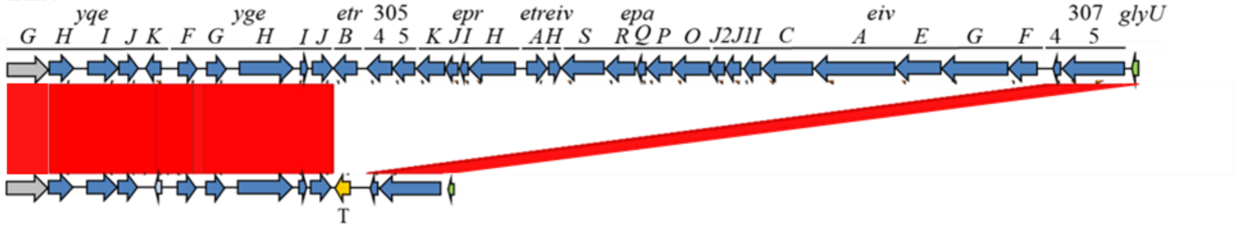
HVH 87 4


yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K J I H A H S R Q P O J2 J1 I C A E G F 4 5
T

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: A backbone of non-*E. coli* 042 downstream ETT2 gene homologues

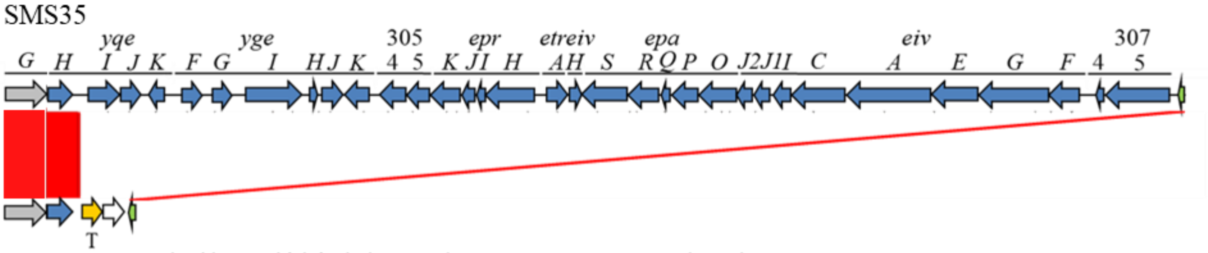
IAI39


yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K J I H A H S R Q P O J2 J1 I C A E G F 4 5
T

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: A backbone which includes *E. coli* 042 downstream ETT2 gene homologues

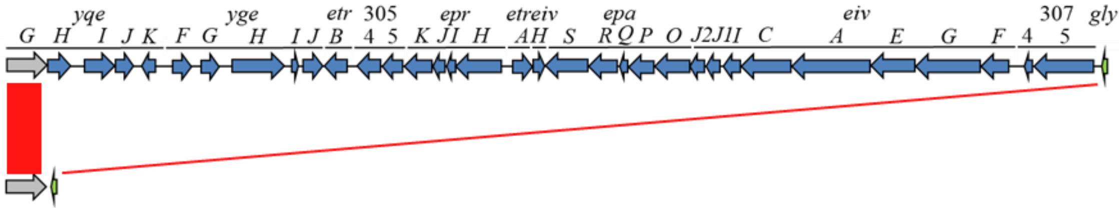
SMS35


yqe yge 305 epr etreiv epa eiv 307
G H I J K F G I H J K 4 5 K J I H A H S R Q P O J2 J1 I C A E G F 4 5
T

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

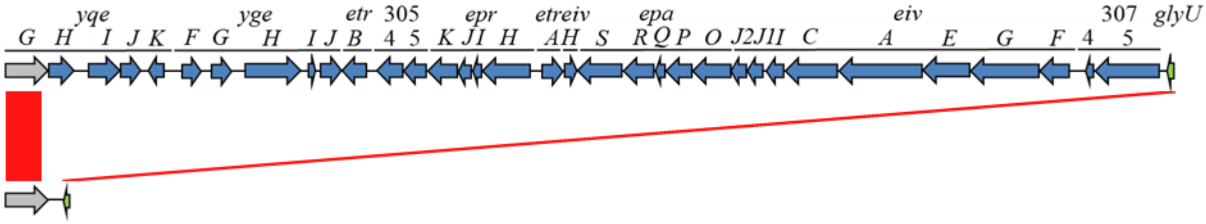
Downstream genes: A backbone which includes *E. coli* 042 downstream ETT2 gene homologues

Swine 65


yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K J I H A H S R Q P O J2 J1 I C A E G F 4 5

E. coli phylogenetic group G

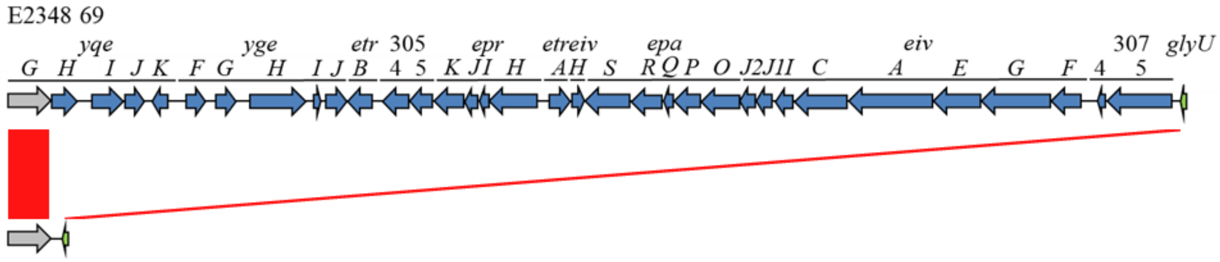
MDR 56


yqe yge etr 305 epr etreiv epa eiv 307 glyU
G H I J K F G H I J B 4 5 K J I H A H S R Q P O J2 J1 I C A E G F 4 5

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: A backbone which includes *E. coli* 042 downstream ETT2 gene homologues

E. coli phylogenetic group B2


E2348 69
yqe
yge
etr
305
epr
etreiv
epa
eiv
307
glyU
G H I J K F G H I J B 4 5 K JI H AH S RQP O J2J1I C A E G F 4 5

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: A backbone which includes *E. coli* 042 downstream ETT2 gene homologues

Figure 6.10 Genotypes of ETT2 homologues as present in *E. coli* from phylogenetic groups A-G. ETT2 *E. coli* 042 reference genes are present and labelled in the top row of each figure, with the genes present in the given strain on the bottom row in the order they are present in the genome. Red bars between the top and bottom rows indicate regions of homology between ETT2 genes of *E. coli* 042 and the given strain identified with BLAST. Genes with identity and length ≥ 30% are coloured dark blue, and homologues disrupted by premature stop codons are coloured light blue. Genes in the strain between homologues with no ETT2 homologue are coloured white, and genes encoding mobile genetic elements including transposons, bacteriophages, or insertion sequences are coloured yellow with a T, P, Is, It, annotated underneath denoting its annotation as a transposon, phage-related gene, insertion sequence, or integrase. An account of whether the 160 genes upstream and downstream of the first and last ETT2 homologue shown in the figure is homologous to the 160 upstream and downstream genes of ETT2 in *E. coli* 042 is described under each figure also.

The presence of four homologues (*yqeH*, *yqeI*, *yqeJ*, and *ygeH*) located at the *yqeG glyU* tRNA locus were found in cryptic clade groups C-III, C-IV, and C-V, and an almost complete *E. coli* strain 042-like ETT2 was found in two group C-I strains (Table 6.6, Figure 6.11). Each of these genotypes were present as separated 4 and 5 block regions of homologues in strains TW10509 and TW15838 respectively, separated by at least 50 kb (Figure 6.11). Each genotype included 5 and 3 mobile genetic elements within 10 kb of blocks respectively (Figure 6.11).

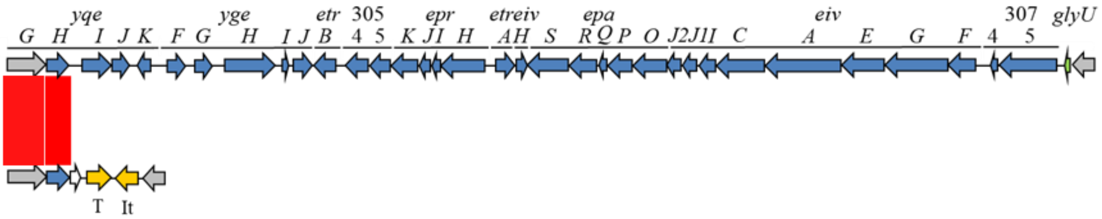
Table 6.6 *E. coli* strain 042 ETT2 PI presence indicated by percentage identity of 33 ETT2 genes within the genome set of 20 strains from cryptic clade phylogenetic groups C-I to C-V.*

Group	Strain name	*yqeH*	*yqeI*	*yqeJ*	*yqeK*	*ygeF*	*ygeG*	*ygeH*	*ygeI*	*ygeJ*	*etrB*	3054	3055	*eprK*	*eprJ*	*eprI*	*eprH*	*etrA*	*eivH*	*epaS*	*epaR*	*epaQ*	*epaP*	*epaO*	*eivJ2*	*eivJ1*	*eivI*	*eivC*	*eivA*	*eivE*	*eivG*	*eivF*	3074	3075
C-I	2 011 08 S1 C1	■																																
C-I	TW10509	■	■	▒	■	■	■	■	■	▓	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	■	■	■	■	▓	■	
C-I	TW15838	■	■	▓	■	■	■	■		▓	■	■	■	■	■	■	■	■	■	■	■	■	■	▓	■	▓	■	■	■	■	■	■	■	■
C-III	KTE31	■	▓	▓				░																										
C-III	TW09231	■	▒	▓																														
C-III	TW09276	■	▓	▓				░																										
C-IV	1 176 05 S3 C2	■	▓	▓				░																										
C-IV	TW11588	■	▓	▓				▒																										
C-IV	TW14182	■	▓																															
C-V	KTE52	■	▒	▓				▒																										
C-V	KTE96	■	▒	▓				▒																										
C-V	TW09308	■	▒	▓				▒																										

* ETT2 gene names and strain groups for each genome are labelled. Darkness of blue indicates increasing percentage identity of each ETT2 reference gene to a homologous gene present in each genome. Only genes with ≥ 30% identity and length relative to ETT2 reference genes are shown.

Cryptic clade group C-I

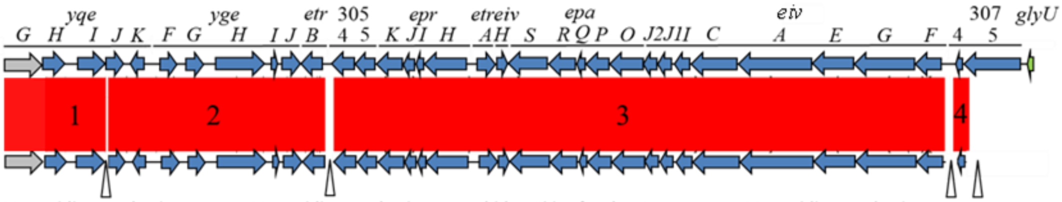
2 011 08 S1 C1


yqe
yge
etr
305
epr
etreiv
epa
eiv
307
glyU
G H I J K F G H I J B 4 5 K J I H A H S R Q P O J2 J1 I C A E G F 4 5
T
It

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: A backbone of non-*E. coli* 042 downstream ETT2 gene homologues

TW10509


yqe
yge
etr
305
epr
etreiv
epa
eiv
307
glyU
G H I J K F G H I J B 4 5 K J I H A H S R Q P O J2 J1 I C A E G F 4 5
1
2
3
4

No mobile genetic elements within 10kb of regions 1 or 2

No mobile genetic elements within 10kb of region 2, 4 bacteriophage-related genes within 10kb of region 3

No mobile genetic elements within 10kb of region 3, 1 gene encoding a transposase within 10kb of region 4

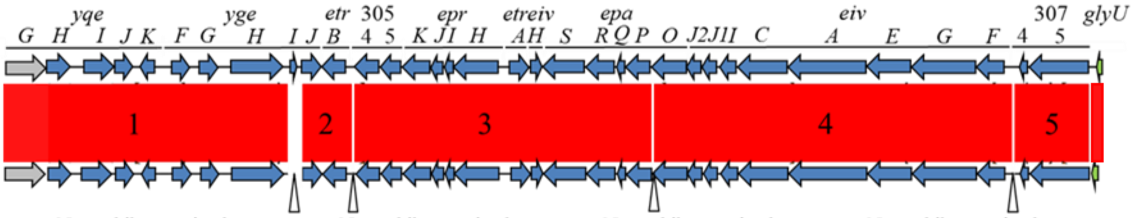
No mobile genetic elements within 10kb of region 4

Order of regions: 1, then 237 genes, then 3, then 669 genes, then 2, then 621 genes, then 4

Upstream genes of region 1: A backbone without *E. coli* 042 upstream ETT2 gene homologues

Downstream genes of region 4: No glyU, a backbone without *E. coli* 042 downstream ETT2 gene homologues

TW15838


yqe
yge
etr
305
epr
etreiv
epa
eiv
307
glyU
G H I J K F G H I J B 4 5 K J I H A H S R Q P O J2 J1 I C A E G F 4 5
1
2
3
4
5

No mobile genetic elements within 10kb of regions 1, 1 gene encoding a transposase within 10kb of region 2

No mobile genetic elements within 10kb of regions 2 or 3

No mobile genetic elements within 10kb of regions 3, 2 genes encoding transposases within 10kb of region 4

No mobile genetic elements within 10kb of regions 4 or 5

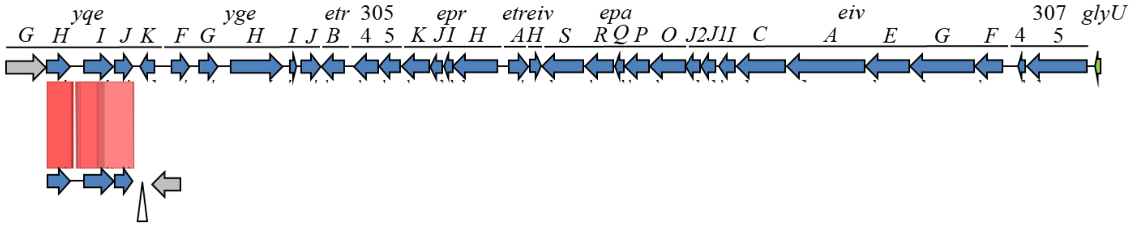
Order of regions: 1, then 1107 genes, then 5, then 722 genes, then 3, then 4, then 187 genes, then 2

Upstream genes of region 1: A backbone without *E. coli* 042 upstream ETT2 gene homologues

Downstream genes of region 5: A backbone without *E. coli* 042 downstream ETT2 gene homologues

Cryptic clade group C-III

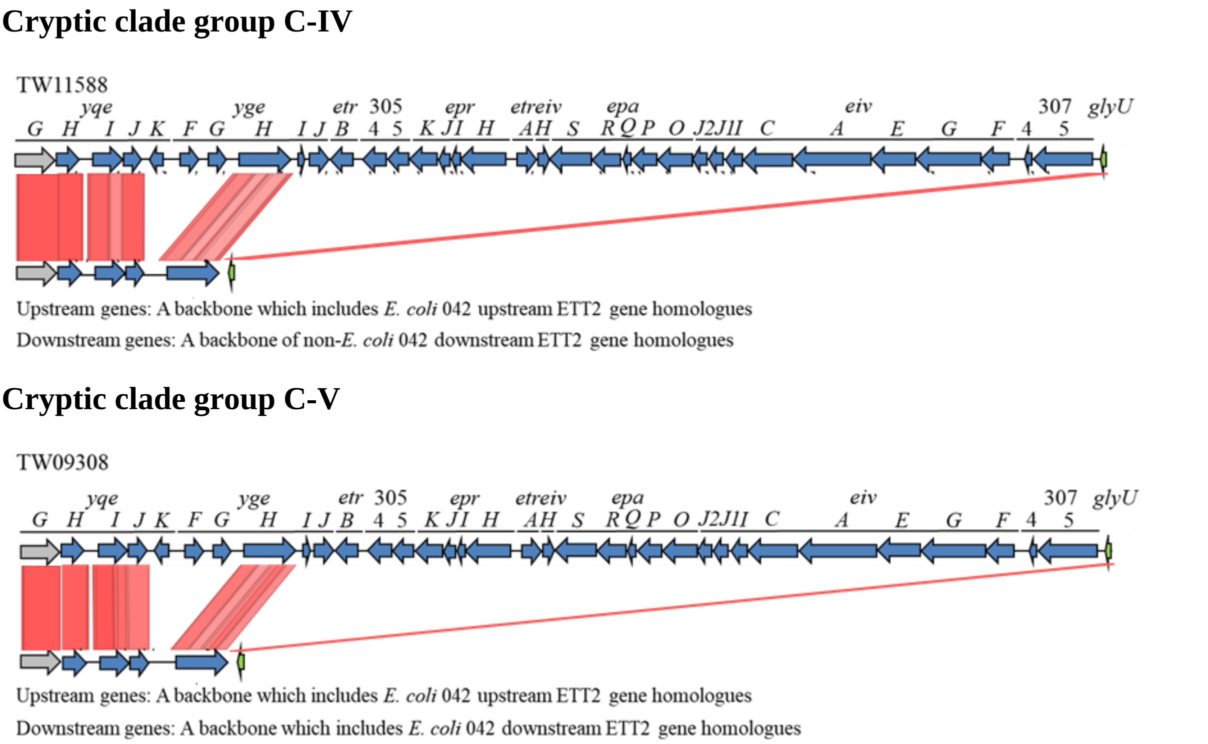
TW09276


yqe
yge
etr
305
epr
etreiv
epa
eiv
307
glyU
G H I J K F G H I J B 4 5 K J I H A H S R Q P O J2 J1 I C A E G F 4 5

9 genes, no mobile genetic elements

Upstream genes: A backbone of non-*E. coli* 042 upstream ETT2 gene homologues

Downstream genes: No glyU, a backbone without *E. coli* 042 downstream ETT2 gene homologues



Cryptic clade group C-IV

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: A backbone of non-*E. coli* 042 downstream ETT2 gene homologues

Cryptic clade group C-V

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: A backbone which includes *E. coli* 042 downstream ETT2 gene homologues

Figure 6.11 Genotypes of ETT2 homologues as present in *E. coli* from cryptic clade phylogenetic groups C-I-C-V. ETT2 *E. coli* 042 reference genes are present and labelled in the top row of each figure, with the genes present in the given strain on the bottom row in the order they are present in the genome. Red bars between the top and bottom rows indicate regions of homology between ETT2 genes of *E. coli* 042 and the given strain identified with BLAST. Genes with identity and length ≥ 30% are coloured dark blue, and homologues disrupted by premature stop codons are coloured light blue. Genes in the strain between homologues with no ETT2 homologue are coloured white, and genes encoding mobile genetic elements including transposons, bacteriophages, or insertion sequences are coloured yellow with a T, P, Is, It, annotated underneath denoting its annotation as a transposon, phage-related gene, insertion sequence, or integrase. An account of whether the 160 genes upstream and downstream of the first and last ETT2 homologue shown in the figure is homologous to the 160 upstream and downstream genes of ETT2 in *E. coli* 042 is described under each figure also.

The presence of *yqeH*, *yqeI*, *yqeJ* homologues were found at the *glyU* tRNA locus within three of 4 *E. fergusonii* genomes (Table 6.7, Figure 6.11), each with genes encoding 2 phage-related proteins, 1 transposase, and integrase within 10 kb of the homologous region (Figure 6.11). A complete *E. coli* strain 042-like ETT2 with absent *ygeF* and *eivJ1* genes were found in 3 of 6 *E. albertii* strains (Table 6.7, Figure 6.11). The same homologues *yqeH*, *yqeI*, *yqeJ* were also found within the genomes of *C. amalonaticus*, *C. farmeri*, *C. rodentium*, and *C. sedlakii* (Table

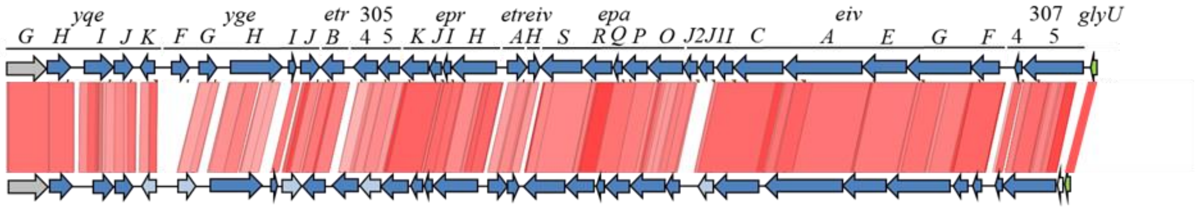
6.7, Figure 6.12), which included tyrosine recombinases within 10 kb of the homologous region (Figure 6.12). *yqeH* was present in *C. braakii*, *C. freundii*, *C. koseri*, *C. pasteuri*, and *C. werkmanii* (Table 6.7, Figure 6.12) with all but the latter 2 being located in close proximity to the *yqeG* homologue and *glyU* tRNA locus (Figure 6.12). Of these genotypes, 3 of 4 contained mobile genetic elements within 10 kb of the homologous regions (Figure 6.12).

Table 6.7 *E. coli* strain 042 ETT2 PI presence indicated by percentage identity of 33 ETT2 genes within the genome set of 10 non-*coli Escherichia* strains, 25 *Citrobacter* species strains. ETT2 gene names and strain groups for each genome are labelled. *

Group	Strain name	*yqeH*	*yqeI*	*yqeJ*	*yqeK*	*ygeF*	*ygeG*	*ygeH*	*ygeI*	*ygeJ*	*etrB*	3054	3055	*eprK*	*eprJ*	*eprI*	*eprH*	*etrA*	*eivH*	*epaS*	*epaR*	*epaQ*	*epaP*	*epaO*	*eivJ2*	*eivJ1*	*eivI*	*eivC*	*eivA*	*eivE*	*eivG*	*eivF*	3074	3075
Escherichia albertii	CB9791	▓	░	▒	░		▒	▒	░	▒	▓	░	▒	▓	▒	▓	▒	▒	▒	▓	▓	▓	▓	░	▒		▒	▓	▓	▓	▓	░	▒	▒
E. albertii	EC06 170	▓	▒	▒	░		▒	▒	░	▒	▓	░	▒	▓	▒	▓	▒	▒	▒	▓	▓	▓	▓	▒	▒		▒	▓	▓	▓	▓	▓	▒	▓
E. albertii	HIPH08472	▓	▒	▒	░		▒	▒	░	▒	▓	░	▒	▓	▒	▓	▒	▒	▒	▓	▓	▓	▓	▒	▒		▒	▓	▓	▓	▓	▓	▒	▓
E. albertii	KF1	▓	▒	▒				▒	░	▒	▓	░	▒	▓	▒	▓	▒	▒	▒	▓	▓	▓	▓	▒	▒		▒	▓	▓	▓	▓	▓	▒	▓
E. albertii	NIAH Bird 5	▓	▒	▒	░		▒	▒	░	▒	▓	░	▒	▓	▒	▓	▒			░	▓	▓	▓	▒	▒		▒	▓	▓	▓	▓	▓	▒	▒
E. albertii	TW07627																																	
E. fergusonii	ATCC 35469	▒	▒	░																														
E. fergusonii	ECD227	▒	▒	░																														
E. fergusonii	FDAARGOS 170																																	
E. fergusonii	GTA EF03	▒	▒	░																														
Citrobacter amalonaticus	FDAARGOS 122	▒	▒	░																														
C. amalonaticus	FDAARGOS 166	▒	▒	░																														
C. amalonaticus	L8A	▒	▒	░																														
C. amalonaticus	Y19	▒	▒	░																														
C. braakii	641 SENT	░																																
C. braakii	GTA CB04	░																																
C. braakii	SCC4	░																																
C. farmeri	GTC 1319	▒	▒	░																														
C. freundii	4 7 47CFAA	░																																
C. freundii	B38	░																																
C. freundii	BD	░																																
C. freundii	CAV1321	░																																
C. freundii	CF04	░																																
C. freundii	P10159	░																																
C. freundii	RU2 BHI16	░																																
C. freundii	UCI 31	░																																
C. koseri	2	▒																																
C. koseri	ATCC BAA 895	▒																																
C. koseri	DNF00568	▒																																
C. pasteurii	CIP 55 13	░																																
C. rodentium	ICC168	▒	▒	░																														
C. sedlakii	NBRC 105722	▒	▒	░																														
C. sp	MGH106																																	
C. werkmanii	NBRC 105721	░																																
C. youngae	ATCC 29220	░																																
Salmonella bongori	N268 08	░	▒	░			░			░				▒		▓				▒	▒	▓	▓	░				▓	▓	░	▒	░		
S. bongori	NCTC 12419	░	▒	░			░			░				▒		▓				▒	▒	▓	▓	░				▓	▓	░	▒	░		
S. enterica subsp arizonae	RKS2983						░			░	░			▒		▓				▒	▒	▓	▓	░				▓	▓	░	▒	░		
S. enterica subsp enterica	C500	░	▒	▒			░			░	░			▒		▓				▒	▒	▓	▓	░				▓	▓	░	▒	░		
S. enterica subsp enterica	SO4698-09	░	▒	▒			░			░	░			▒		▓				▒	▒	▓	▓	░				▓	▓	░	▒	░		
S. enterica subsp houtenae	01 0133						░			░	░			▒		▓				▒	▒	▓	▓	░				▓	▓	░	▒	░		
S. enterica subsp indica serovar	11 b 1 7 BCW 1559	░	▒	▒			░			░	░			▒		▓				▒	▒	▓	▓	░				▓	▓	░	▒	░		
S. enterica subsp salamae	RKS2993	░	▒	▒			░			░	░			▒		▓				▒	▒	▓	▓	░				▓	▓	░	▒	░		
S. enterica subsp VII	1 40 g z51 2439 64						░			░	░			▒		▓				▒	▒	▓	▓	░				▓	▓	░	▒	░		

* Darkness of blue indicates increasing percentage identity of each ETT2 reference gene to a gene present in each genome, and for comparison, SPI-1 and SPI-3 homologues in *Salmonella* are coloured in green. Only genes with ≥ 30% identity in protein BLAST analysis relative to full length ETT2 reference genes are shown. No strains other strains from the 200-strain *Enterobacteriaceae* set contained ETT2 gene homologues.

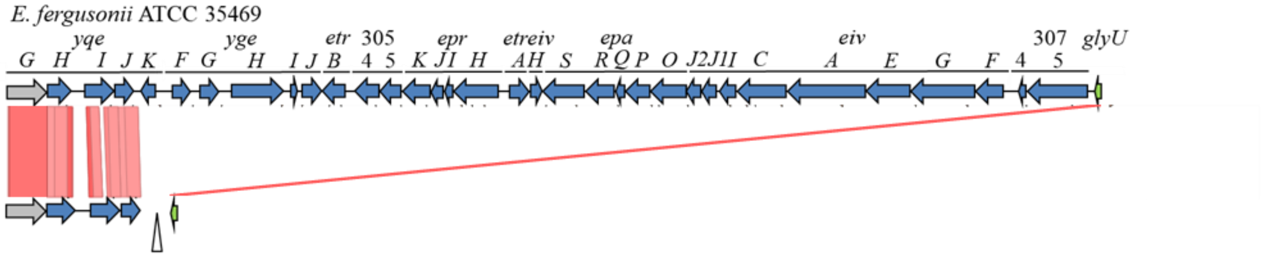
E. albertii CB791


yqe
yge
etr 305
epr
etreiv
epa
eiv
307 glyU
G H I J K F G H I J B 4 5 K JI H AH S RQ P O J2J1I C A E G F 4 5

Upstream genes: yqeG, then a backbone without *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: glyU, then a backbone which includes *E. coli* 042 downstream ETT2 gene homologues

E. fergusonii ATCC 35469

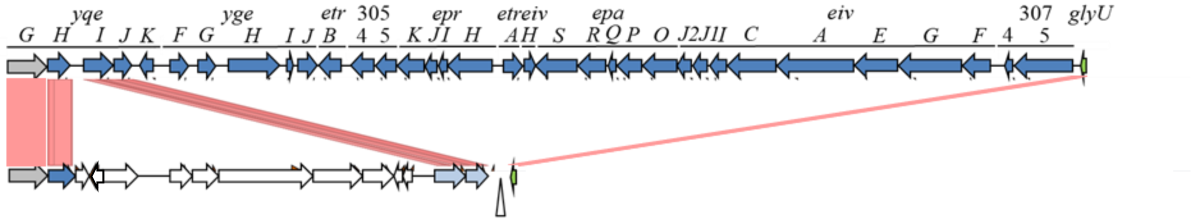

yqe
yge
etr 305
epr
etreiv
epa
eiv
307 glyU
G H I J K F G H I J B 4 5 K JI H AH S RQ P O J2J1I C A E G F 4 5

22 genes, 4 genes encoding mobile genetic elements (2 x P, T, It) within 10kb of homologues

Upstream genes: yqeG, then a backbone without *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: glyU, then a backbone which includes *E. coli* 042 downstream ETT2 gene homologues

C. amalonaticus FDAARGOS 122

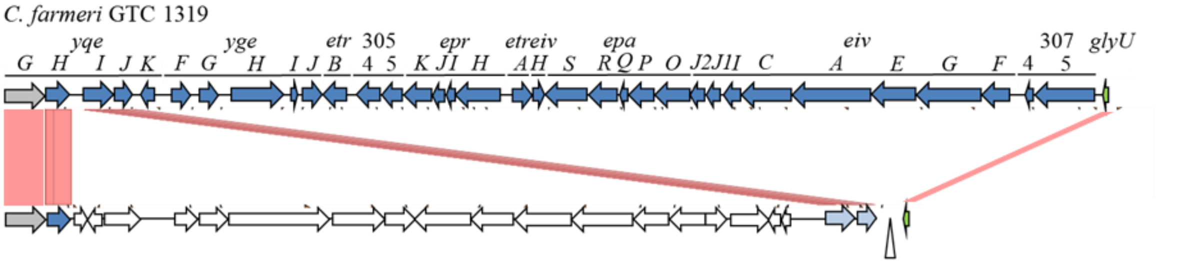

yqe
yge
etr 305
epr
etreiv
epa
eiv
307 glyU
G H I J K F G H I J B 4 5 K JI H AH S RQ P O J2J1I C A E G F 4 5

62 genes, 2 genes encoding tyrosine recombinases within 10kb of homologues

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: A backbone including *E. coli* 042 downstream ETT2 gene homologues

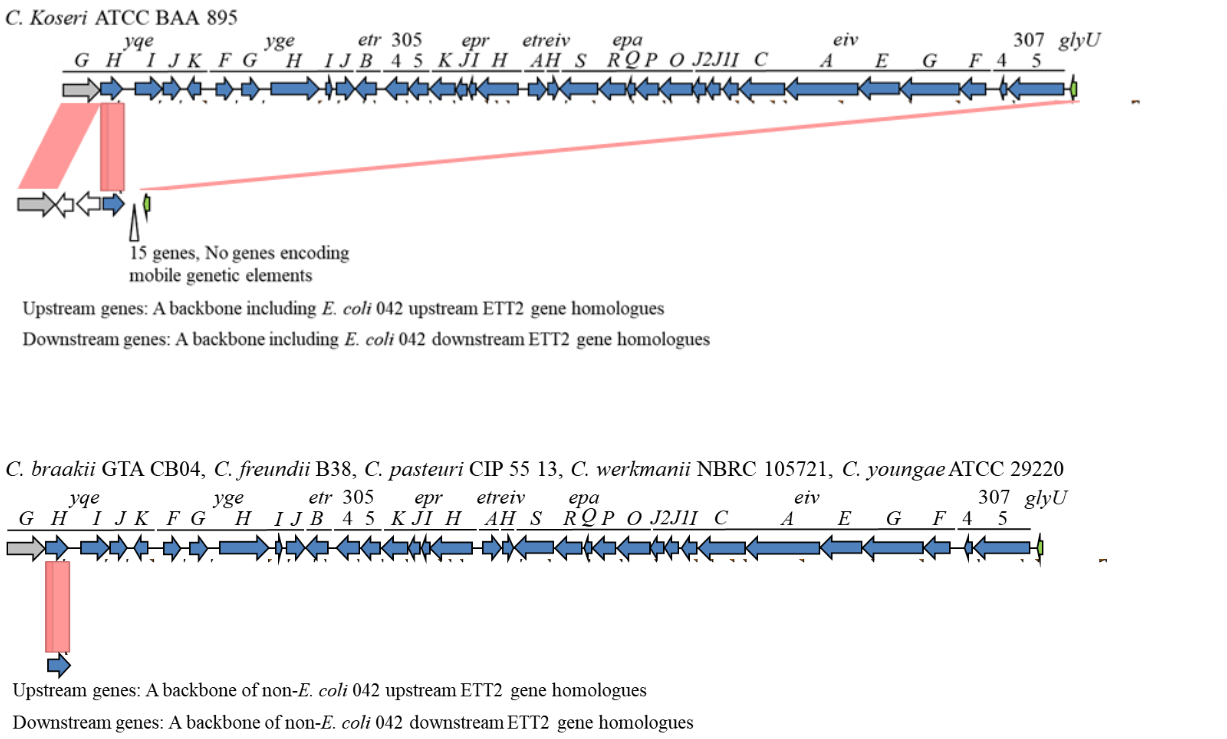
C. farmeri GTC 1319


yqe
yge
etr 305
epr
etreiv
epa
eiv
307 glyU
G H I J K F G H I J B 4 5 K JI H AH S RQ P O J2J1I C A E G F 4 5

57 genes, 2 genes encoding tyrosine recombinases within 10kb of homologues

Upstream genes: A backbone which includes *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: Abackbone including *E. coli* 042 downstream ETT2 gene homologues



C. Koseri ATCC BAA 895

yqe
yge
etr 305
epr
etreiv
epa
eiv
307 glyU
G H I J K F G H I J B 4 5 K JI H AH S R Q P O J2J1I C A E G F 4 5
15 genes, No genes encoding mobile genetic elements

Upstream genes: A backbone including *E. coli* 042 upstream ETT2 gene homologues

Downstream genes: A backbone including *E. coli* 042 downstream ETT2 gene homologues

C. braakii GTA CB04, *C. freundii* B38, *C. pasteuri* CIP 55 13, *C. werkmanii* NBRC 105721, *C. youngae* ATCC 29220

yqe
yge
etr 305
epr
etreiv
epa
eiv
307 glyU
G H I J K F G H I J B 4 5 K JI H AH S R Q P O J2J1I C A E G F 4 5

Upstream genes: A backbone of non-*E. coli* 042 upstream ETT2 gene homologues

Downstream genes: A backbone of non-*E. coli* 042 downstream ETT2 gene homologues

Figure 6.12. Genotypes of ETT2 homologues as present in *E. coli* strains from cryptic clade phylogenetic group C-I to C-V, *Citrobacter*, and non-*coli Escherichia*. ETT2 *E. coli* 042 reference genes are present and labelled in the top row of each figure, with the genes present in the given strain on the bottom row in the order they are present in the genome. Red bars between the top and bottom rows indicate regions of homology between ETT2 genes of *E. coli* strain 042 and the given strain identified with BLAST. Genes with identity and length ≥ 30% are coloured dark blue. Separate regions of adjacent genes are numbered and their order and number of genes between each region in the strain's genome is described under each figure. An account of whether the 160 genes upstream and downstream of the first and last ETT2 homologue shown in the figure is homologous to the 160 upstream and downstream genes of ETT2 in *E. coli* 042 is described under each figure also.

6.3.3. Phylogenetic trees created using ETT2 gene sequence

To determine the relatedness of ETT2 genes, phylogenies of ETT2 cluster genes which were shared across strains, homologous ETT2 gene sequences, were created. This was done by aligning homologous ETT2 gene sequences using Muscle before constructing phylogenies with RAxML. To determine the phylogenetic position of each group in each a manual approach was taken by colouring strains of each group a different colour for identification of the group, before noting the internal node denoting the clade in which ≥ 50% of strains of a given phylogenetic group were present. This node was noted as the topological position of the phylogenetic group in the phylogeny. For each phylogenetic group a strain which represented the central clustered topological position relative to all strains of the group under this node was chosen and ETE Toolkit (Huerta-Cepas et al. 2016) was used to prune each group's representative from the phylogeny. A new phylogeny was then inferred using just the group reference strain sequences. Phylogenetic trees of genes *yqeH* and *yqeI* created using an amino acid alignment (Figure 6.13 and Figure 6.14 respectively) showed *Citrobacter*, Cryptic clade, and non-*coli Escherichia* to be clustered as an outgroup to *E. coli* with the exception of C-I strains. C-I strains were shown to cluster with D1 and D2 strains in the *yqeH* to 3055, yqeH to *epaP*, and the *yqeH* to 3075 phylogenies, and in the core ETT2 sequence phylogeny from strains of groups E, D1, D2, C-I, and *E. albertii* strains (15,436 bp) (Figure 6.15).

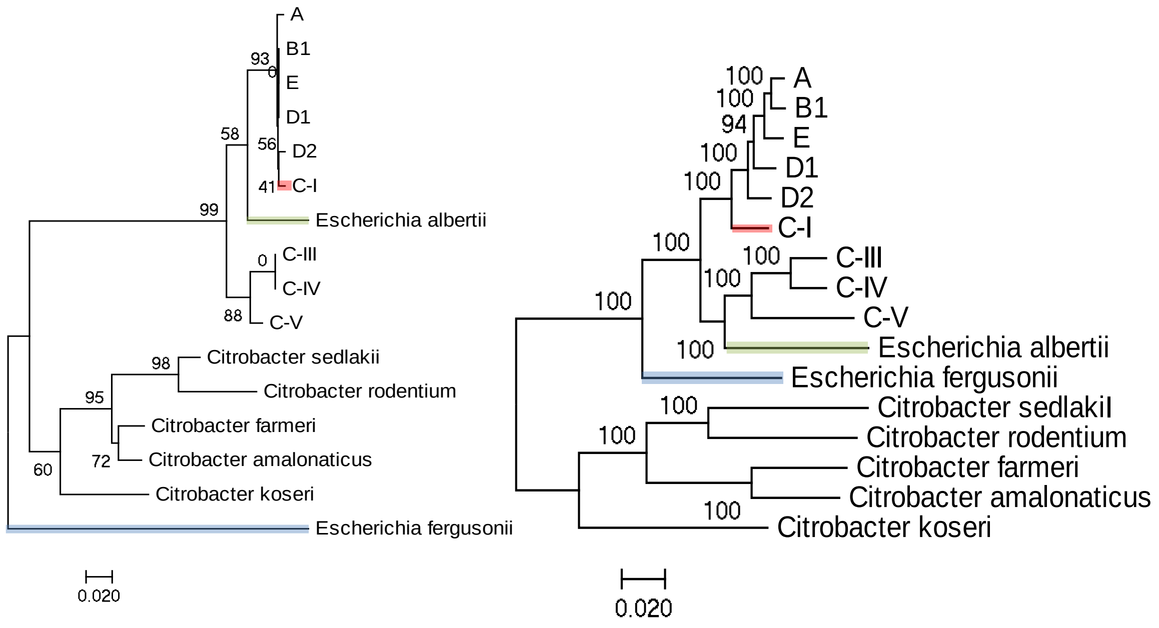


Figure 6.13. Left: A midpoint rooted Maximum likelihood RAxML phylogeny constructed using a core gene alignment of 210 amino acids identified as homologues of ETT2 gene *yqeH* in *E. coli* strain 042, which are present in 86 non-*coli Escherichia* and *Citrobacter* strains, and *E. coli* strains from groups A, B1, E, D1, D2, C-I, C-III, C-IV, and C-V. Each phylogenetic *E. coli* group or species is represented by a strain which represents phylogenetic position for > 50% of strains of that group with homologous amino acids to the mentioned ETT2 genes. Right: a midpoint rooted RAxML core genome phylogeny constructed using a 2.2 Mb alignment. Topological differences between the two phylogenies are highlighted with a red, green, or blue to indicate the different branching topology for groups C-I, *E. albertii,* and *E. fergusonii* respectively. Percentage bootstrap support values are shown on internal branches. The scale bar indicates the number of substitutions per site represented by the branch length shown.

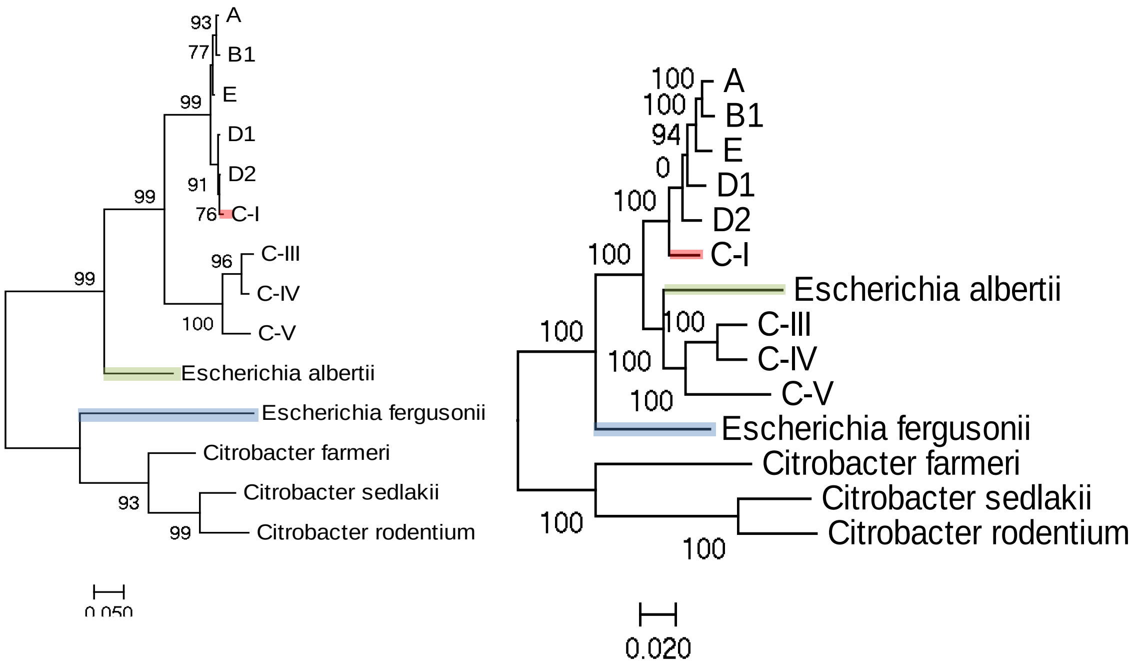


Figure 6.14. Left: A midpoint rooted Maximum likelihood RAxML phylogeny constructed using a core gene alignment of 497 amino acids identified as homologues of ETT2 genes *yqeH* and *yqeI* in *E. coli* strain 042, which are present in 80 non-*coli Escherichia* and *Citrobacter* strains, and *E. coli* strains from groups A, B1, E, D1, D2, C-I, C-III, C-IV, and C-V. Each phylogenetic *E. coli* group or species is represented by a strain which represents phylogenetic position for >50% of strains of that group with homologous amino acids to the mentioned ETT2 genes. Right: A midpoint rooted RAxML core genome phylogeny constructed using a 2.2 Mb alignment. Topological differences between the two phylogenies are highlighted with a red, green, or blue to indicate the different branching topology for groups C-I, *E. albertii*, and *E. fergusonii* respectively. Percentage bootstrap support values are shown on internal branches. The scale bar indicates the number of substitutions per site represented by the branch length shown.

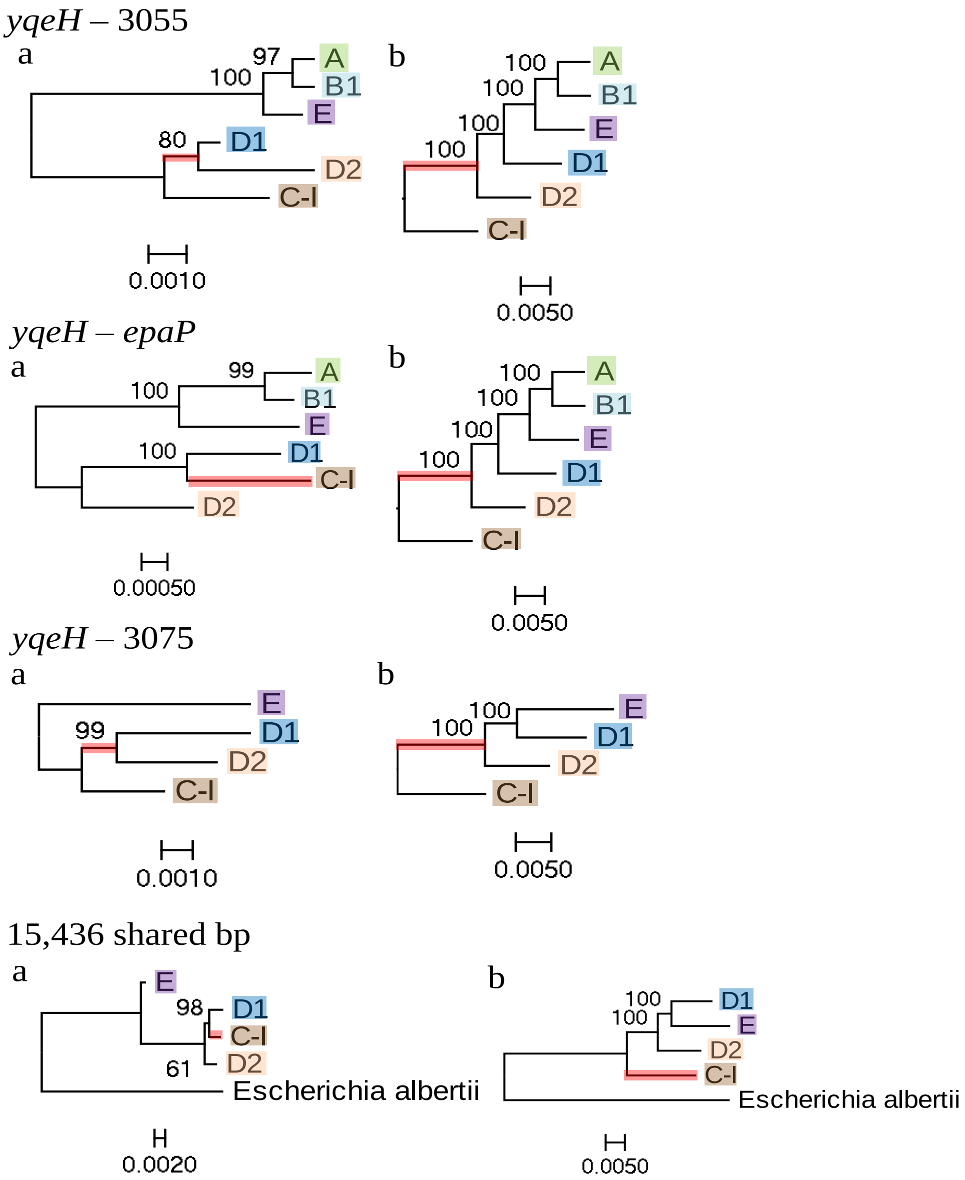
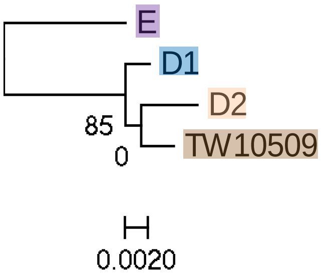


Figure 6.15. Four midpoint rooted phylogenies constructed using shared sequence across respective phylogenetic groups (a) and the core genome phylogeny of the same strains for topological comparison (b). The phylogenies are constructed using a core gene alignment of 1: 7,310 bp from 12 adjacent ETT2 gene homologues of *yqeH* through until 3055 in 66 *E. coli* strains from groups A, B1, E, D1, D2, and C-I, 2: 13,402 bp from 22 adjacent homologues of *yqeH* to *epaP* from 62 *E. coli* strains from groups A, B1, E, D1, D2, and C-I, 3: 24, 285 bp from 33 adjacent homologues of 22 *E. coli* strains from groups E, D1, D2, and C-I, and 4: 15,436 bp from ETT2 homologue sequence shared by strains of groups E, D1, D2, C-I, and *E. albertii*. Each phylogenetic *E. coli* group or species is represented by a strain which represents the phylogenetic position for > 50% strains of that group with homologous nucleotides to the mentioned ETT2 genes. The core genome phylogeny shown (in b) shares a red highlighted branch which indicates a topological difference for the group attached to the branch between the phylogenies. Percentage bootstrap support values are shown on internal branches. The scale bar indicates the number of substitutions per site represented by the branch length shown.

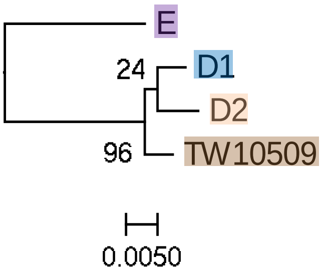
Phylogenetic trees of DNA sequence from each block of ETT2 homologues in the C-I strain genomes TW10509 and TW5838 revealed blocks are clustered closest to phylogenetic groups as follows for TW10509: 1: D2, 2: outgroup to D1 and D2 groups, 3: D2, 4: D1, and for TW15838: 1: outgroup to D1 and D2 groups, 2: D1, 3: D2, 4: D2, 5: E (Figure 6.16). To infer evidence of recombination of ETT2 sequence between strains from *E. coli* phylogenetic groups D1, E, C-I, and *E. albertii*, pan genome analysis was first carried out using Roary before core gene quartet recombination analysis was carried out using the resulting core gene alignment as input. The analysis between D1, E, C-I and *E. albertii* strains revealed that both TW10509 and TW15838 C-I strains clustered in a non-clonal manner, closest to D1 strains relative to group E and *E. albertii* strains in 200 kb (7.4%) of core genome sequence, and closest to E strains relative to the D1 and *E. albertii* in 10 kb (0.37%) of core genome sequence (Figure 6.17).

TW10509

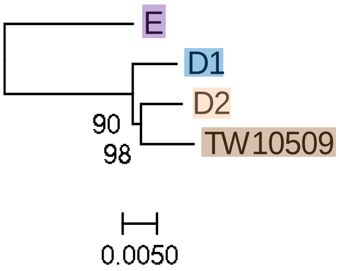
1



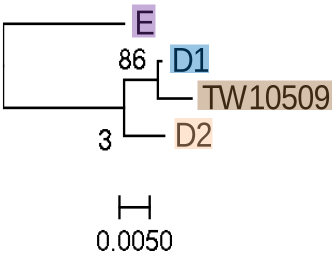
2



3

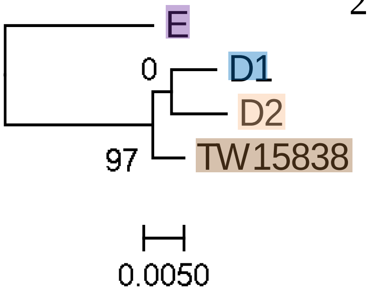


4

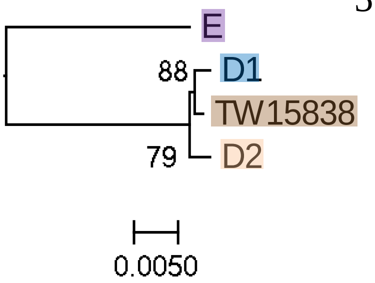


TW15838

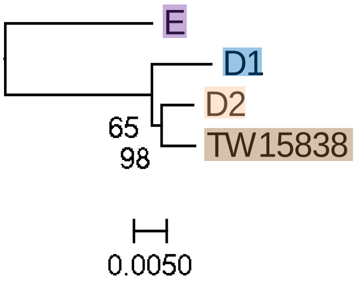
1



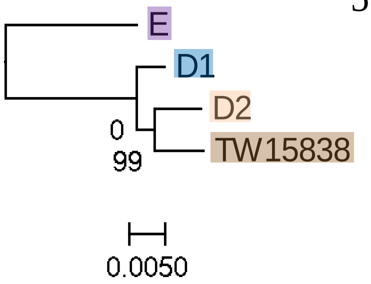
2



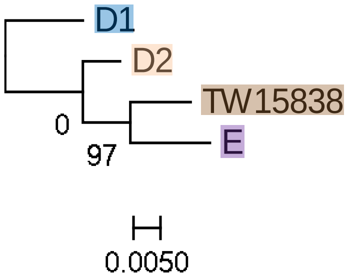
3



4



5



Core genome

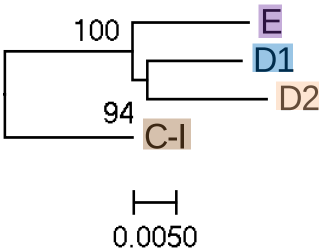


Figure 6.16. Maximum likelihood RAxML phylogenies constructed using the DNA sequence of each block region of adjacent ETT2 homologous genes which are present in strains TW10509 and TW15838 from group C-I, with the same DNA sequence in each case present in *E. coli* strains in groups E, D1, and D2. For TW10509, region 1 includes *yqeH* to *yqeI*, region 2 is *yqeJ* to *ygeK*, region 3 is 3054 to *eivF*, and region 4 is gene 3074. For TW15838, region 1 includes *yqeH* to *ygeH*, region 2 is *ygeI* to *ygeJ*, region 3 is *ygeK* to *epaQ*, region 4 is *epaP* to *eivF*, and region 5 is 3074 to 3075. Groups are coloured consistently between trees and the core genome phylogeny for groups E, D1, D2, and C-I is shown at the bottom so topologies can be compared to it. Percentage bootstrap support values are shown on internal branches. The scale bar indicates the number of substitutions per site represented by the branch length shown.

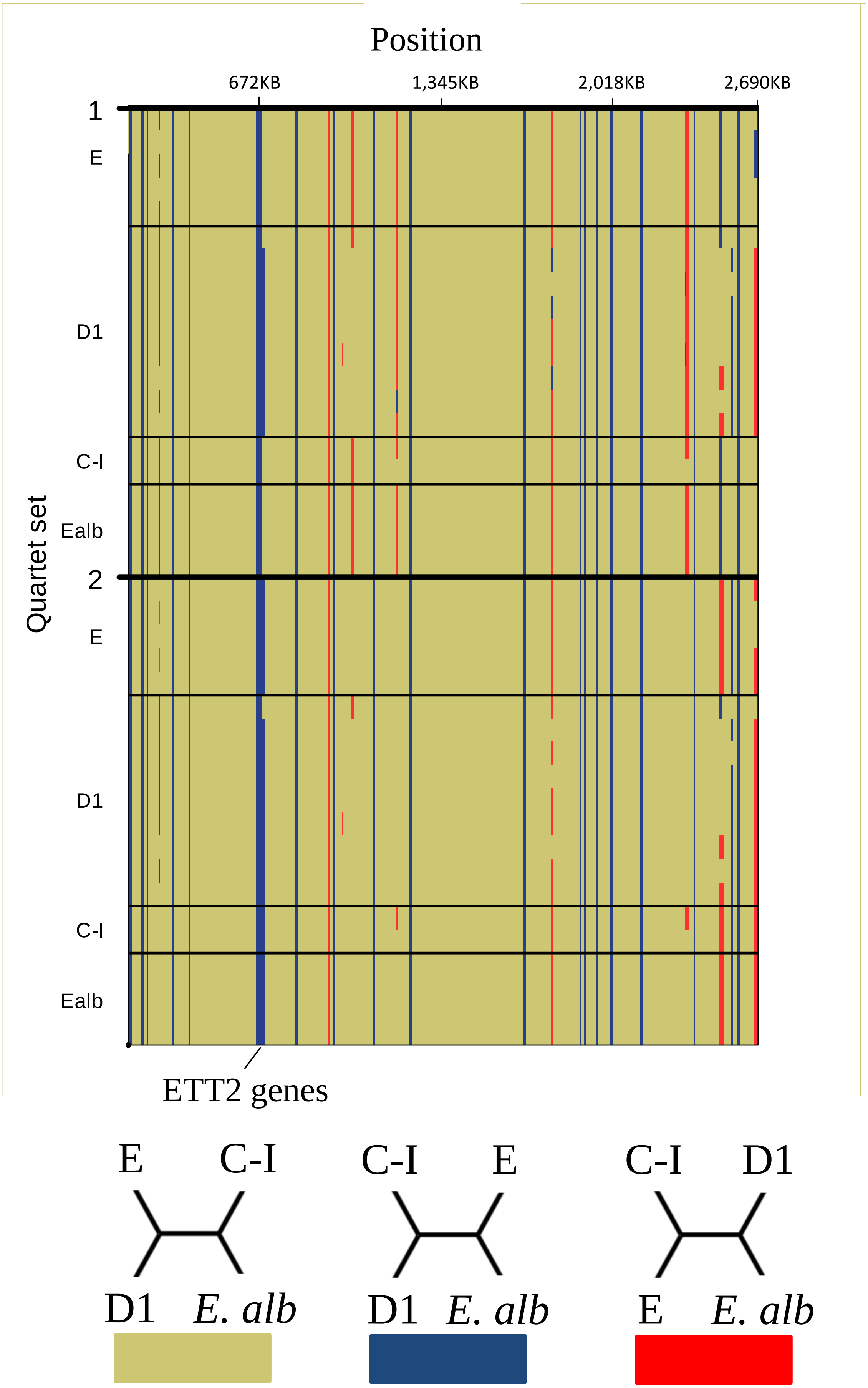

Position
672KB
1,345KB
2,018KB
2,690KB
Quartet set
1
E
D1
C-I
Ealb
2
E
D1
C-I
Ealb
ETT2 genes
E
C-I
D1
E. alb
C-I
E
D1
E. alb
C-I
D1
E
E. alb

Figure 6.17. Sliding window 10 kb quartet recombination plot of potential recombination events between *E. coli* phylogenetic groups C-I (2), E (5), and D1 (9), and *E. albertii* (4) strains with a complete ETT2 island of homologues present. a. The x axis is the length of the core genome (2,690,705 bp) ordered as in strain TW10509, and the y axis indicates rows of quartet unrooted phylogenies constructed from the DNA sequence of each sliding window, each with different combinations of strains, one from each group C-I, E, D1, and *E. albertii* in each quartet. Non-overlapping windows move across the core gene in 10 kb intervals and the sequence from each sliding window was used to construct a quartet. The relationship of the four groups in the quartet were then coloured in the figure in accordance with their relationships shown in b. The lines dividing the plot mark when the group labelled on the left of the plot is changing in each quartet and the other three strains in the quartet are always the same. The thick line labelled ‘1’ and ‘2’ indicate when the reference strains, the strains which do not change in each quartet when they are the non-changing group for a section, change and a new reference set of four strains is used. The plot shows 200 kb (7.4 %) of C-I core genome clusters closer to D1 than other groups, and 10 kb (0.37%) clusters with group E closest. ETT2 gene sequence highlighted in the core genome is highlighted.

6.3.4. *Eip* locus homologue presence

To identify homologous genes to *eip* locus genes present in the reference genome *E. coli* 042 (locus *eicA* – *air*) the same approach that was used to identify ETT2 gene homologues was used. *Eip* cluster homologues were found to be absent from strains of groups A, B1, G, B2, C-II, C-IV, CV, and all strains in the 200 *Enterobacteriaceae* set other than *E. albertii,* where *eicA* and *eilA* homologues were present (Table 6.8). All *E. coli* strain 042 homologues were present in two E, two D1, and four D2 strains, and all genes other than *air* were present in three E, five D1, seven D2, and one C-I group strains (Table 6.8, Table 6.9, Table 6.10). Full locus absence was seen in two group E strains, and one C-I and *E. albertii* strain (Table 6.8, Table 6.10). Within groups E, D1, and D2, genes *eicA*, *eipD*, and *eilA* were present in all but two strains with no truncations evident by <80% identity (Table 6.8, Table 6.9, Table 6.10), while *air* was truncated in the majority of E, D1, and D2 strains, with a complete deletion in 50% of D1 and 75% of E strains and full absence in C-I and *E. albertii* strains (Table 6.8, Table 6.9, Table 6.10). All *eip* locus gene homologues were found to be located within 10 kb of one another in each given genome.

Table 6.8. *E. coli* strain 042 *eip* full length gene cluster presence indicated by percentage identity of 6 complete and partial *eip* genes within the genomes of 8, 10, and 7 *E. coli* strains from phylogenetic groups E, D1, and D2 respectively, 3 *E. coli* cryptic clade group C-I strains, and 6 *E. albertii* strains. *

Group	Strain name	*eicA*	*eipB*	*eipC*	*eipD*	*eilA*	*air*
E	400654	█	█	█	█	█	░
E	AF85	█	█	█	█	█	
E	B185	█	▓		█	█	
E	C161 11						
E	D6-113	█	▓		█	█	
E	H16 Santai	█	█	█	█	█	▒
E	O157:H7 str. Sakai						
E	O169:H41 str. F9792	█	▓		█	█	
D1	042	█	█	█	█	█	█
D1	B354	█	█	▓	█	█	
D1	C1	█	▓		█	█	▒
D1	C4	█	▓		█	█	▒
D1	EC2	█	▓		█	█	▒
D1	ECOR 48	█	▓		█	█	
D1	TA255	█	█	▓	█	█	
D1	TA280	█	░	▓	█	█	
D1	UMN026	█	█	█	█	█	▒
D1	upec 213	█			█	█	
D2	24 1 R1	█	█	▓	█	█	▒
D2	BIDMC 19C	█	█	▓	█	█	▒
D2	HVH 87 4	█	█	▓	█	░	
D2	IAI39	█	█	▓	█	█	░
D2	SMS35	█	█	▓	█	█	░
D2	swine65	█	░	▓	█	█	
D2	UCI 57	█	█	▓	█	█	
C-I	2 011 08 S1 C1						
C-I	TW10509	█	▓		█	█	
C-I	TW15838	█	█	▓	█	█	
Escherichia albertii	CB9791	▒				▒	
E. albertii	EC06 170	▒				▒	
E. albertii	HIPH08472	▒				▒	
E. albertii	KF1	▒				▒	
E. albertii	NIAH Bird 5	▒				▒	
E. albertii	TW07627						

* *Eip* gene names and strain groups for each genome are labelled. Darkness of blue (scale: < 30% identity (white) to 100% identity (dark blue)) indicates increasing percentage identity of each *eip* reference gene to a gene present in each genome. Only genes with ≥ 30% identity in protein BLAST analysis relative to full length *eip* reference genes are shown. No other strains from the *E. coli* phylogenetic groups A-G, cryptic clade groups C-I - C-V, or the 200-strain *Enterobacteriaceae* set contained *eip* gene homologues.

Table 6.9. Mean percentage identity values of *eip* homologues relative to *E. coli* strain 042 reference genes for all strains of each of the phylogenetic groups A-G in the 100-strain set.

Group	*eicA*	*eipB*	*eipC*	*eipD*	*eilA*	*air*
E (N = 8)	99	77	60	83	85	54
D1 (N = 10)	99	84	69	98	100	72
D2 (N = 7)	99	90	84	97	93	53

Table 6.10. Percentage of strains from each of the A-G phylogenetic groups from the 100-strain set with inferred deleted *eip* homologues within their genomes, for each *eip* gene.

Group	*eicA*	*eipB*	*eipC*	*eipD*	*eilA*	*air*
E (N = 8)	25	25	25	25	25	75
D1 (N = 10)	0	0	0	0	0	50
D2 (N = 7)	0	0	0	0	0	0

6.3.5. Phylogenetic trees created using *eip* locus gene sequence

To identify create phylogenies using *eip* locus gene sequence for the purpose of inferring evolutionary history, the same approach that was used to for creating phylogenies of ETT2 gene homologue sequences was used. Phylogenies constructed using *eip* locus DNA sequence showed that group C-I strains clustered closest to group D2 strains in all cases other than for the gene *eilA* (Figure 6.17). *E. albertii* gene sequences were shown to cluster as an outgroup in the *eicA* and *eilA* phylogenies (Figure 6.18).

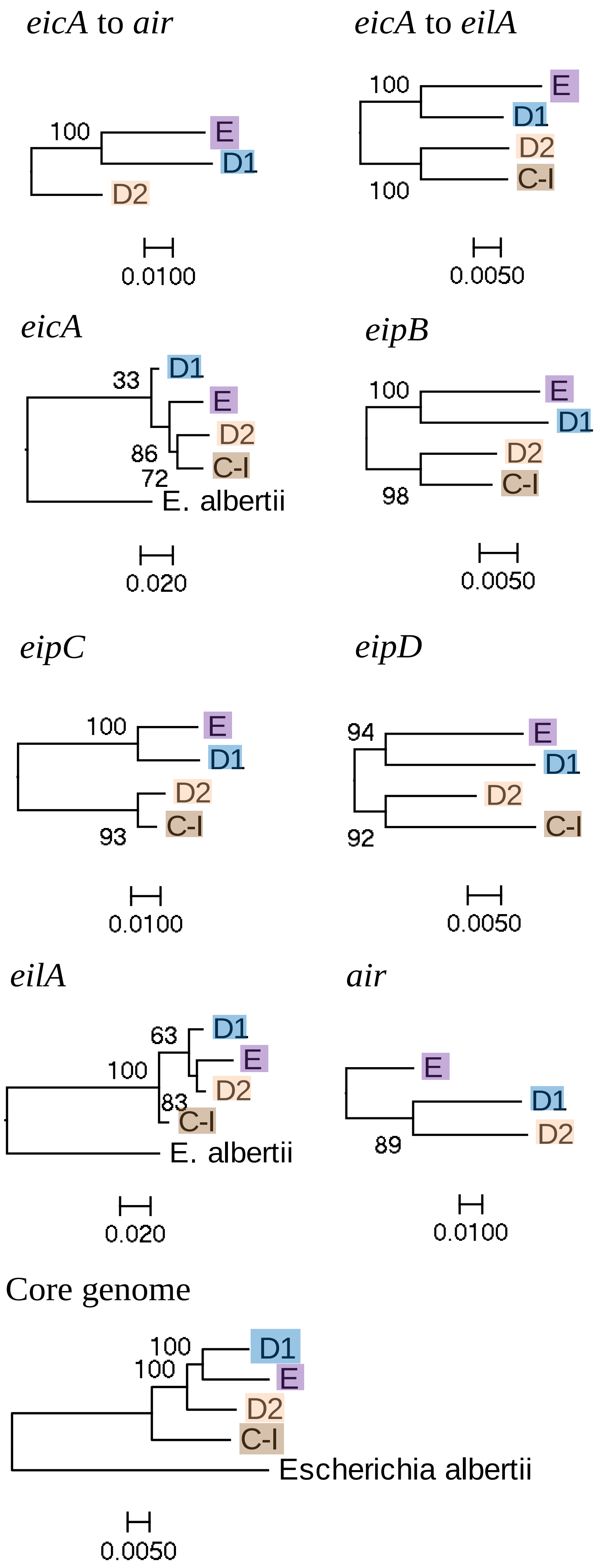

eicA to *air*
100
E
D1
D2
0.0100
eicA to *eilA*
100
E
D1
D2
C-I
100
0.0050
eicA
33
D1
E
D2
86
C-I
72
E. albertii
0.020
eipB
100
E
D1
D2
C-I
98
0.0050
eipC
100
E
D1
D2
C-I
93
0.0100
eipD
94
E
D1
D2
C-I
92
0.0050
eilA
63
D1
100
E
D2
83
C-I
E. albertii
0.020
air
E
D1
D2
89
0.0100
Core genome
100
D1
100
E
D2
C-I
Escherichia albertii
0.0050

Figure 6.18. Midpoint rooted maximum likelihood RAxML phylogenetic trees constructed using DNA sequence obtained from *E. coli* strain 042 *eip* locus gene homologues shared by strains of E, D1, D2, C-I, and *E. albertii* strains. A strain from each group which represents the phylogenetic position of > 50% of group members with homologous nucleotides to the given *Eip* locus gene(s), is labelled with its group label. Phylogenies are labelled as follows: *EicA to air* genes (constructed using 17,901 bp from homologues of *eicA, eipB, eipC, eipD, eilA*, and *air* present in 11 strains), *eicA* to *eilA* (constructed using 6,480 bp from homologues of *eicA, eipB, eipC, eipD,* and *eilA* present in 18 strains). Phylogenies labelled *eicA, eipB, eipC, eipD, eilA,* and *air* are constructed using 498 bp from 26 strains, 1,782 bp from 18 strains, 1,152 bp from 18 strains, 1,350 bp from 18 strains, 1,698 bp from 23 strains, and 11,421 bp from 13 strains respectively from the gene of their name. Groups are coloured identically between phylogenies for ease of comparison and the core genome phylogeny of groups D1, E, D2, C-I, and *E. albertii* is provided at the bottom for topological comparison to each phylogeny. The core genome phylogeny is shown (right) for comparison. Percentage bootstrap support values are shown on internal branches. The scale bar indicates the number of substitutions per site represented by the branch length shown.

6.4. Discussion

The purpose of the work in this chapter was to provide an up-to-date account of ETT2 and *eip* cluster revealing the extent of its complex evolution and genotypic diversity, confirm the results of previous research, and provide possible phenotypic insights for the clusters, using the set of 120 phylogenetically diverse *E. coli* strains described in Chapter 3 and 200 other species representatives of the *Enterobacteriaceae* family. This was effectively addressed through the testing of the hypothesis and 3 objectives.

The first objective was addressed effectively by characterising and recording the number of unique ETT2 cluster genotypes using computational genome sequence comparison and analysis of differences using in-house programs. The 116 ETT2 varied genotypes found in previously unreported strains illustrate that ETT2 is more diverse between and within groups than reported previously. The previous studies characterising ETT2 genotypes described in Table 6.1 have not had the availability of genome sequence data to create samples of sufficiently large phylogenetic diversity to reveal inter and intragroup ETT2 genotypic differences at high resolution. The more representative level of sampling used in this study has enabled identification of 18 A, 30 B1, 8 E, 10 D1, and 7 D2 genotypes from phylogenetic groups A-D2. This is an increase from 3 A, D1, D2, 2 B1, and 1 E genotype described by Ren et al. (2004) (Figure 6.4), and the 4 identified in groups A and B1, and 5 identified in groups B2 and D1 by Wang et al (2016). It has also aided previously unreported ETT2 homologues to be found in the genus *Citrobacter*, *E. fergusonii*, and in *E. coli* cryptic clade strains which are likely to be remnants of previously existing complete clusters.

The second objective was effectively addressed through reviewing the presence of the ETT2 cluster in *non-E. coli* bacterial species. Methods of phylogenetic construction were used with manual inspection of genotypes to infer the evolutionary point of origin of both the ETT2 and *eip* cluster. The presence of ETT2 homologues were found to be located at the *yqeG glyU* tRNA locus in 6 *Citrobacter* species and the phylogenies created using ETT2 sequence indicates that ETT2 was acquired prior to the divergence of *Citrobacter* and *Escherichia* genera. This is earlier than the previous inference of acquisition in the most recent common ancestor of *E. coli* and *E. albertii* (Ooka et al. 2015, Ren et al. 2004). Presence of two tyrosine recombinases within 10 kb of ETT2 sequence of four genotypes of *Citrobacter* indicate that is possible that they were involved in the deletion (Figure 6.11).

The phylogenies of ETT2 gene sequence exhibit have a largely consistent topological relationships of phylogenetic groups relative to one another with those in the core genome phylogeny in Figure 6.14, Figure 6.15, and Figure 6.16 for *E. coli* from groups A-G groups, cryptic clades, non-*coli Escherichia*, and *Citrobacter* groups. The exception is ETT2 sequence from group cryptic clade group C-I stains, which clustered closest with D1, D2, and E group strains in these phylogenies, indicating that no ETT2 sequence recombination occurred between non-C-I ancestors of each group. However, in the *yqeH* and *yqeI* homologue phylogeny (Figure 6.15), *E. fergusonii* clusters closer to *Citrobacter* strains which is a topology that differs to that in the core genome phylogeny where it clusters closest to *E. coli* and *E.*

albertii (Figure 6.15). If the genes were clonally inherited though by these groups and did not undergo ancestral recombination, it is likely that alignment of DNA sequence used in the phylogeny (497 amino acids) exhibited an insufficient number of amino acid polymorphisms to adequately distinguish genes from each of the groups in phylogenetic analysis. If this was the case, the insufficient number of polymorphisms would have been the cause for the ETT2 sequences from the groups clustering together in the phylogeny and therefore having a topology which conflicted to the topology for the same groups in the core genome phylogeny. However, this cannot be confirmed unless further DNA sequencing reveals other *E. fergusonii* strains to contain more ETT2 homologues.

After acquisition in the ancestor of *Citrobacter* and *Escherichia* strains, it can be proposed that ETT2 was inherited as a complete 33-gene *E. coli* 042-like island by *Citrobacter* strains, followed by deletions of all genes but *yqeH*, *yqeI*, and *yqeJ* either through the action of tyrosine recombinases in *C. amalonaticus*, *C. farmeri*, *C. rodentium*, and *C. sedlakii* and all but *yqeH* in others through an unknown mechanism. It can also be inferred that in the *Citrobacter* sister group, the genus *Escherichia*, inherited a complete ETT2 island by its ancestor. *E. fergusonii*, the outgroup to *E. coli* and *E. albertii*, then inherited the complete island after which a deletion of all genes but *yqeH*, *yqeI*, and *yqeJ* mediated most likely through the activity of bacteriophages, an integrase sequence, and a transposon in this specific region. Also, it can be inferred that *E. albertii* clonally inherited a complete island after which deletion of *ygeF* and *eivJ1* coding sequences occurred most likely as a result of frameshifts in sequence after indel mutations occurred at each gene’s locus. The complete ETT2 island was then inherited clonally by the ancestor of C-V, C-IV, and C-III Cryptic clades which underwent two deletion events from *yqeK* to *ygeB*, and *ygeI* to 30755, leaving four genes *yqeH*, *yqeI*, *yqeJ*, and *ygeH* when the 3 groups diverged, after which *ygeH* was deleted in C-III group strains all through an unknown mechanism. The results indicate the ETT2 cluster underwent homologous recombination in C-I strains. The phylogenies of ETT2 sequence from C-I strains (Figure 6.12) indicates that in strain 2 011 08 S1 C1 all genes but a *yqeH* homologue was lost through the action of a transposon and an integrase but not in the other C-I strains. At least 3 different recombination acquisitions of sequence likely occurred in the ancestor of C-I strain TW10509 and at least 4 in the ancestor of C-I strain TW15838 likely through the action of transposon, integrase, and bacteriophages, genes encoding each found in each genotype (Figure 6.10). This is also inferred as each block region of ETT2 homologues clusters either with D1, D2, as an out group to D1 and D2, or with group E strains (Figure 6.12). This is inconsistent with these groups' relationships in the core genome phylogeny (also Figure 6.12), which most likely reflects clonal relationships where group C-I clusters as an outgroup to all A-G phylogenetic groups. Regions 1 and 3 from TW10509 and 3 and 4 from TW15838 both cluster with group D2 strains and cover the same 21 homologues ranging 3054 to *eivF* so likely originate from the same recombination event in the ancestor of both strain TW10509 and TW15838 (Figure 6.12). Similarly region 2 of TW10509 and region 1 of TW15838 both cluster as outgroups to D1 and D2 strains (Figure 6.12) and include the five homologues ranging *yqeJ* to *ygeH* (Figure 6.11) so may have been acquired in a single event in the ancestor also. The quartet analysis of recombination between E, D1, C-I, and *E. albertii* strains indicates that recombination between C-I and D1 and D2 strains makes up 7.4 % of the core genome including ETT2 genes, and between C-I and E strains makes up 0.34% of the genome (Figure 6.13), so horizontally transferred DNA between the ancestor of C-I, D1, and D2 like that in ETT2 has had a small but significant impact in structuring C-I group core genomes.

The results also indicate that upon clonal inheritance of the ETT2 by the ancestor of *E. coli*, the island most likely underwent a full deletion in a single event in the ancestor of G and B2 group strains as both groups have an almost identical ETT2 deletion genotype. Absence in group B2 has previously been reported (Ren et al. 2004). It was likely inherited by the ancestors of D1 and D2 strains as a complete island, but underwent homologue truncations of *ygeJ* in the ancestor of group E, A, and B1 strains, *epaS* in the ancestor of group A strains, and 3054, *eprI*, *epaS*, *epaO* in the ancestor of group B1 (Table 6.5). A previously reported 8.7 kb deletion is likely to have occurred in the ancestor of groups A and B1 (Ren et al. 2004), although it may have occurred in two stages, where the four homologues *eivA* to *eivF* were deleted first in the ancestor to both groups, and then a further deletion of the five homologues ranging *eivJ2* to 3074 in the ancestor of all strains other than 25 (group A) and APECO78 (group B1), which possess the same deletion genotype, although a quartet analysis would need to be carried out to remove recombinant sequence and create a non-recombination phylogeny to determine if these 2 strains are an outgroup to groups A and B1. Overall, these reported characteristic genotypic differences between phylogenetic groups do indicate that ETT2 cluster genotypes could theoretically be used as a phylogenetic marker in some cases for groups A, B1, E, D1, and D2, with absence of ETT2 as a marker for groups G and B2. This supports the chapter hypothesis and the result reported by Ren et al. (2004).

After ancestral lineage divergence within groups A and B1, further premature stop codons and deletions occurred can be inferred to have occurred in homologues including 3054, 3055, *ygeF*, *eivJ1*, eivC, and 3075 in group A, and *eprH*, *epaI*, *eivC*, and 3075 in group B1 most frequently in individual strain lineages (Table 6.5). Partial island deletions of at least four genes occurred in 4 A, 3 B1, 2 E, and 4 D2 strains, including a complete deletion in 1 A and 1 D2 strain also occurred in individual strain lineages. Deletions were likely mediated by 5 transposon and 3 integrases in group A, a transposon, 2 integrases, 3 insertion elements, and 5 bacteriophages in group B1, and 2 transposons, and 1 integrase in group E strain O157 H7 Santai (Figure 6.10). Also, recombination in strains CFSAN026836 (group A) and E110019 (group B1) likely occurred after strain divergence as each's ETT2 genotype is comprised of 2 and 3 block regions respectively and genes encoding 1 integrase and insertion element, and 2 phages are present within 10 kb of homologues in the latter strain (Figure 6.10). Phylogenetic analysis of the regions revealed each originate within the same group of each respective strain (not shown). Other than these two occurrences, all ETT2 homologues in *E. coli* from A-G groups retained a single structure as a single genomic island block, evidenced by the observation that no homologue was further than 10 kb from any other in each genotype.

The third objective was effectively addressed through manually reviewing ETT2 and *eip* cluster genotypes to identify genes without an inferred loss-of-function mutation and collecting likely functional annotations for proteins encoded by such genes to infer phenotypes for the genotypes. It was also effectively addressed by manually determining if genotypes with likely phenotypes were associated with the phylogenetic group of the strain with the given genotype.

The presence of a truncated *ygeJ* and *epaS* homologue in both groups A and B1 may be evidence indicating that either premature stop codons in these genes, or the four-homologue deletion in the common ancestor of the groups A and B1, strain 25, and strain APECO78 deemed the T3SSs of these groups to be non-functional. If so, the loss of the homologues did not reduce the survival of the ancestor, which meant after the homologue losses occurred, they were passed on to the descendant groups A and B1 when both appeared. Similarly, the truncation of *ygeJ* in group E strains may be the reason why a higher proportion of premature stop codons and deletions have accumulated in other genes in group E strains compared to D1 strains, which mostly contained intact ETT2 homologue islands (Table 6.5). The single premature stop codon may have deemed the ancestor's T3SS non-functional and enabled the accumulation of multiple other mutations in T3SS-related genes in the island as an intact T3SS was not crucial for survival. In groups D2 and D1, no homologue truncations or deletions are present in all strains so cannot be inferred to have occurred in the most recent common ancestor of strains for each group (Table 6.5). The inheritance of an intact ETT2 island by all strains of the group would have occurred and the ETT2 island may have worked as a functional unit encoding a T3SS that was crucial for improved fitness, survival, and proliferation during adaptation when the group was diverging. Any homologue losses would have been associated with high fitness costs, meaning such genotypes are not inherited and a complete ETT2 genotype was maintained.

Studies disrupting the open reading frame of ETT2 genes have shown that an *E. coli* 042-like intact ETT2 in a genome indicate that the island is functional (Zhang et al. 2004, Wang et al. 2017, Luzander et al. 2016, Ideses et al. 2015, Sheikh et al. 2016). However, with the assumption that all ETT2 genes must be intact to encode a functional T3SS, an ETT2 genotype with truncations and deletions may retain some function despite the island as a whole not encoding a functional T3SS. Such incomplete but functioning genotypes may exist in phylogenetic groups A, B1, E, D1, and D2, Cryptic clade, non-coli *Escherichia*, and *Citrobacter* genomes as the patterns of remaining genes indicate that ETT2 may represent an island which coordinates non-ETT2 gene regulation. ETT2 regulator genes *etrA*, *etrB*, and *eivF* have previously been found to have important roles in virulence (Zhang et al. 2004, Wang et al. 2017, Luzander et al. 2016), and the ETT2 regulator genes *yqeH*, *yqeI*, *ygeH*, *etrA*, *etrB*, and *eivF* are among the least deleted, least truncated, and most conserved ETT2 genes among phylogenetic groups A-E. Within groups A, B1, E, D1, and D2, *yqeH* is almost identical to that in *E. coli* strain 042 when present, and all other regulators exhibit little to no truncation and at least 86.6% average identity to reference genes (Table 6.9). The exception is the truncation of *etrB* in some strains (81.7 % and 70.4% mean identity) in groups A and B1 respectively, and complete absence of *eivF* (Table 6.5). Consistent with this, *yqeH, yqeI,* or just *yqeH,* have been retained after the deletion of most or all other ETT2 genes in ten strains from groups A–G, and C-III, C-IV, C-V, non-*coli Escherichia,* and *Citrobacter* strains. This may indicate that the two regulators *yqeH*, and *yqeJ* have a functional fitness value as a regulatory island after the loss of a functional ETT2 T3SS.

Non-regulator encoding genes which exhibit fewer truncations and deletions than other genes include (by group) A: *ygeG* (chaperone), *eivH*, and *epaRQPO* (putative surface presentation of antigen proteins), and B1: also *ygeG*, *eivH*, and *epaR* (Table 6.5). This may indicate that additionally to roles in regulation, genotypes in groups A and B1 may have roles in chaperoning non-ETT2 related proteins and the presentation of antigens on the cell surface. Also, the non-regulatory genes *eprHIJK* with a previously found role in host serum survival (Ideses et al. 2015) are on average not truncated among strains evidenced by >82% identity relative to reference genes in groups A, E, D1, and D2 (Table 7.6), indicating that these genes may be important for intracellular survival among strains in these groups.

The results indicate that the *eip* locus island was acquired prior to the divergence of *E. coli* and *E. albertii* groups as homologues are absent in *Citrobacter* strains. It can be proposed that the *eip* locus underwent deletions of *eipB*, *eipC*, *eipD*, and *air* in the ancestor of *E. albertii* and a complete deletion in strain TW07627. It underwent a complete deletion in the ancestor of C-III, C-IV, and C-V clade strains, and in the ancestor of C-I strains underwent a deletion of *air*, and an acquisition via homologous recombination from D2 strains in C-I group strains, replacing the 4 homologues *eicA* to *eipD* and leaving *eilA*, which shows C-I strains clustering as an outgroup to E, D1, and D2. C-I group strain 2 011 08 S1 C1 also underwent a complete deletion. The ancestor of phylogenetic groups A-G can be inferred to have clonally inherited a complete *eip* locus but it underwent a full deletion independently in the ancestor of groups A and B1, and G and B2. The lineages leading to strains C161 11 and O157:H7 str. Sakai (group E) underwent full deletions of the *eip* locus, *eipC* was independently deleted in groups E and D1 at least 3 times, and *air* was independently deleted at least 4 times in groups E, D1, D2, and C-I, indicating that these genes are unlikely to be important for survival in these groups.

The *eip* locus may have been independently lost in strains of *E. coli* which lack the ETT2 island due to a functional dependence on ETT2. Sheikh et al. (2016) found that expression of the *eip* locus gene *eilA* increased expression of ETT2 genes *eivF* and *eivA*, showing that the two genomic islands interact. If so, a deletion of ETT2 genes could result in *eip* locus becoming obsolete and increase the likelihood they accumulate truncations through frameshift mutations through lack of use. In groups D1 and D2, *eicA*, *eipD*, and *eilA* are 'complete' homologues with no truncation in all strains, and *eipB* is also present in all but one strain, which may indicate that the homologues have remained intact as a result of an intact ETT2 island being present in groups D1 and D2 (Table 6.5). It could be that the two genomic islands have a continued interaction which promotes survival and proliferation in many of the strains within groups D1 and D2. As mentioned previously, group E may have a potentially non-functional T3SS in all strains due to a putative *ygeJ* truncation in the group ancestor, potentially meaning maintaining the *eip* locus can no longer be a functional genomic island in group E strains. Other than the *air* homologue, group E exhibits the largest strain percentage of truncation events in *eipB, eipC, eipD,* and *eilA* homologues (Table 6.10), and the lowest mean identity values for present genes compared to groups D1 and D2 (Table 6.9). The group also has the highest rate of *air* gene deletion compared to strains in groups D1 and D2 and contains strains with two full deletions (Table 6.8), which suggests that these genes may be not functional in group E. The *eicA* homologue is highly conserved (99% identity, Table 6.9) with no manually observed premature stop codon however, indicating it may be functional despite a mostly degraded *eip* locus in group E (Table 6.8). Furthermore, the absence of the *eip* locus in groups A and B1, and G and B2 could be due to the partial or complete absence of ETT2 in the ancestor of those lineages.

In summary, the hypothesis, overall aim, and objectives were addressed which resulted in the creation of an up-to-date account of ETT2 and *eip* cluster genetic diversity and evolutionary history. The account was also then compared to that generated in reports published in previous research. Overall, the ETT2 island appears to have undergone widespread deletion events mediated largely by mobile genetic elements and undergone multiple recombination events in the C-I strain group. The island has several group-specific characteristics including a complete absence in G and B2 strains, an 8.7 kb deletion in A and B1 strains, truncations of *ygeJ* in group E, A, and B1 strains, *epaS* in group A strains, and 3054, *eprI*, *epaS*, *epaO* in group B1 strains. These deletions and truncations thus potentially have some utility in being used as markers for phylogenetically assigning groups to *E. coli* strains from groups A-G. This supports the chapter hypothesis and the result reported by Ren et al. (2004), that ETT2 cluster genotypes can be used as a phylogenetic marker in some cases. ETT2 is mostly likely not a functioning T3SS in groups A, B1, and E but quite possibly functions as a regulatory island that coordinates gene expression in other areas of the genome. Furthermore, it likely has this role in the groups C-III, C-IV, C-V, non-coli *Escherichia*, and *Citrobacter* where intact regulator homologues exist despite degradation of the island as a whole. Results indicate the point of acquisition of ETT2 to be in the ancestor of *Citrobacter* and *Escherichia* strains, a point earlier than previously reported, but there is no evidence that the *eip* locus originated at this point as the earliest point of occurrence is in *E. albertii* strains. There is some evidence that the *eip*-locus may have a functional dependence on the existence of a complete 042-like ETT2 island, as its presence is positively associated with more complete ETT2 genotypes. The results of this study will likely enhance understanding of the origins, evolution, diversity, and functionality of ETT2 and the *eip* locus which can be considered as genomic clusters with a probable importance to survival and proliferation in many commensal and pathogenic *E. coli*.

Chapter 7: Discussion

7.1. Summary

The findings presented in Chapters 3, 4, 5, and 6 include previously unreported insights into *E. coli* genetic diversity, evolutionary history, pathogenicity, and phylogenetic group assignment. These findings relate to strains representing the *E. coli* species as a whole (Chapter 3, Chapter 4, and Chapter 6), and to *E. coli* strains which have been previously linked to virulence (Chapter 5 and Chapter 6). In Chapter 3 an analysis of the *E. coli* pan genome, the core genome phylogeny, and the clonal frame phylogeny was carried out using a set of *E. coli* strains which exhibited a level of phylogenetic and phenotypic diversity that reflects the species. The analysis provided new estimates for the size of the core genome and revealed the proportion and numbers of accessory genes shared between groups. It also revealed the core genome topology to be as in Figure 3.5 and the clonal frame phylogeny to be as in Figure 3.8. The work in Chapter 3 provided an up-to-date estimate for the number of genes which have undergone recent and ancestral recombination between the phylogenetic group lineages also. This revealed that recent inter-group recombination between phylogenetic groups A-G has occurred in a greater number of genes than inferred ancestral inter-group recombination events. The types of genes which were inferred to exhibit the least ancestral recombination in their history were also reported, together with those which were inferred to have undergone ancestral recombination between phylogenetically distant groups. Furthermore, the analyses carried out supported the existence of a new *E. coli* phylogenetic group which clusters most closely to group B2 and was tentatively named ‘G’.

Chapter 4 was an analysis to provide evidence to support the use of a proposed cgMLST as a reliable schema for assigning clonal and phylogenetic groups to group A-G and cryptic clade *E. coli* strains respectively, and as an alternative to using a core gene phylogeny, a cgMLST, or a 7-15 locus MLST or multiplex schema. Through analysis of novel MLST schemas it was determined that 100% correct group assignment was only reliably achievable if 800 genes at least were utilised for the MLST. As using this number of genes was deduced to take a prohibitively long time, which compared to the 256 of the proposed cgMLST, it was proposed that the cgMLST be developed for use as an *in-silico* cgMLST either to function as a standalone program or one which forms a component of a larger bioinformatic sequence typing program like BIGSdb (Jolley et al. 2010). Furthermore a 7-gene MLST schema and a 10-locus multiplex schema were also developed which were inferred at their current stage of development to provide 100% correct group assignment (except for the inability of the multiplex to differentiate groups A and B1). Like the 256-gene proposed cgMLST, these were also suggested to be used as *in-silico* tools.

Chapter 5 was an investigation into the genetic basis for UPEC urinary tract infection virulence phenotypes, specifically those which mediate mild to significant decreases in ureter contractility in 20 UPEC strains. Genes were determined which were significantly associated with different phenotypes through grouping strains by their exhibited phenotypes at 5 and 9 hours and comparing the genomic contents of strains across the groups. It was concluded that multiple phenotypes were the result of the expression of a range of genes shared across different subsets of strains which have been acquired through HT in many cases. Out of the genes identified, of note was a 9-gene island, termed ‘1A1’, which was potentially responsible for the phenotype in 8 strains, possibly through the use of the haemolysin operon, a *kdpD* operon regulator present in 8 strains which allows the uptake of potassium and may directly inhibit ureter contractility, and an oxidoreductase enzyme which may have been involved in reducing the impact of oxidative stress defences initiated by host cells. Also of note was the identification of the two previously reported haemolysin operons *hlyI* and *hlyII* (Velasco et al. 2018) as part of an 8 gene cluster in strains 536 and J96, which exhibited the earliest decreases in ureter contractility. The haemolysin operons were inferred to play a significant role in both strains exhibiting an early ureter contractility inhibition phenotype.

Chapter 6 was a re-examination of the evolutionary history and genotypic diversity of the ETT2 and *eip* genetic clusters across *E. coli* and other groups using up-to-date sets of genomes. These sets consisted of strains which represented the phylogenetic diversity present across *E. coli* and other closely related *Citrobacter* and *Escherichia* species. 116 varied ETT2 genotypes were identified and it was determined that the ETT2 cluster has undergone widespread deletion and gene truncation events mediated mostly by mobile genetic elements. Multiple recombination events of the cluster between C-I strains and groups E, D1, and D2 seem to have occurred, and genotypic characteristics specific to phylogenetic groups were found to exist. These included truncations of the genes *ygeJ* in groups A, B1, and E strains, *epaS* in group A strains, and ECs3054, *eprI*, *epaS*, *epaO* in group B1 strains. Such truncations could potentially be used as a marker for phylogenetic group assignment. It was inferred based on the arrangement of intact genes across phylogenetic groups that the ETT2 cluster most likely does not encode a functioning T3SS in groups A, B1, E but is potentially functional in group D1. ETT2 possibly has an alternative function as a regulatory island which coordinates gene expression in other areas of the genome in strains with incomplete ETT2 genotypes. The fragmented presence in groups C-III, C-IV, C-V, non-*coli Escherichia*, and *Citrobacter* indicate the point of acquisition of ETT2 to be in an ancestor of *Citrobacter* and *Escherichia*, earlier than previously reported. However, there was no evidence that the associated *eip* locus originated at this point, as it was only found in *E. coli* and *E. albertii* strains. Based on *eip*-locus genotypes the *eip*-locus was speculated to have a functional dependence on an 042-like complete ETT2 cluster due to more complete *eip* genotypes being found when a complete ETT2 cluster was present.

7.2. The implications and contribution of this work

The findings presented in Chapter 3 contribute to understanding of the genetic diversity and evolutionary history. The strain set of 100 *E. coli* exhibited a level of phylogenetic diversity which has not previously been reported and which could potentially be used as a reference set for studies of *E. coli* evolution and pathogenicity. The genetic diversity of the strains also meant the results most likely provide a level of detail not previously investigated, regarding *E. coli* phylogenetic group unique and shared core and accessory gene contents, ancestral and recent recombination, and the clonal frame phylogeny. Also, the finding of a previously unreported phylogenetic group (tentatively named 'G') in the clonal frame phylogeny means the inclusion of the group could be considered in future genetic studies of *E. coli*.

In Chapter 4 a novel 256 gene cgMLST, 7 gene MLST, and 10 loci multiplex schema and set were determined which could be used to correctly assign clonal groups to A-G group strains and phylogenetic groups to cryptic clade strains in future studies. As tools they have potential to facilitate accurate clonal and phylogenetic group assignment to groups A-G and cryptic clade groups respectively, within a reduced computational time period if used as *in-silico* tools with hundred or thousands number of genome sequences.

The genes identified in Chapter 5 which are significantly associated with ureter contractility decrease phenotypes could be further investigated to corroborate the inferred links to pathogenesis in this study. If gene knockout experiments are carried out and one of the designated phenotypes were found to be less virulent in the knockout strain through being associated with less ureter contractility in the experiment compared to a wild type strain, that could be the stimulus to initiate investigation to corroborate the existence of a gene to phenotype link. Further to this, it can be speculated that such further investigation could potentially result in the development of a therapeutic drug to target the strains with the genes through interacting with the protein expressed by the gene.

The findings in Chapter 6 about the ETT2 and *eip* locus contribute to understanding of the genotypic variation and evolutionary history of the two clusters, specifically the point at which their acquisitions occurred. The findings provide an insight into the ways in which type 3 secretion system (T3SS) gene clusters can undergo significant changes in terms of deletions and truncation in different lineages to potentially take on new roles which are not necessarily secretion related. This is as previously unreported genotypes were determined across *E. coli*, *Citrobacter*, and non-*coli Escherichia* lineages, some of which have many gene truncations and deletions and some of which have genes with a previously reported link to virulence which were inferred to potentially express virulence-associated proteins as the genes are intact and truncation-free.

7.3. Recommendations and future directions

The next step to continue the work carried out in Chapter 3 would be to investigate the rates of ancestral and recent recombination of genes by functional category. It would be interesting to investigate whether genes associated with virulence undergo more regular recombination compared to genes not directly involved such as metabolism-related genes. This could be carried out through programmatically comparing the frequency of topological differences of gene phylogeny inconsistencies to the core genome phylogeny. It would be a way of measuring the frequency of genes of a certain clonal or phylogenetic group exhibiting inferred recombination. In this future study, this could be done with genes associated with virulence compared to those not associated with virulence. The frequency of inferred recombination could also be compared between genes associated with fitness in different environments such as those associated with the different ExPEC pathovars or those associated with survival in different non-animal environments like on different soils, waters, or plants.

The determination of the new group labelled as ‘G’ also provides potential for future research. Strains of the group originate from varied commensal and pathogenic pathovars so it appears in the first instance that they are not a group which is specifically adapted for exploitation of a specific environmental or host niche. The behaviours and metabolic differences of strains of this group could therefore be characterised to determine the specific features which differentiate them from other *E. coli* when interacting with different substrates and in varied environments.

When sampling genomes to include in the set of 100 *E. coli* strains representing groups A-G in Chapter 3 it was clear that there is an over representation of publicly available whole genome sequences genomes from infection-related sources, and of strains from phylogenetic groups B2, A, and B1 compared to other groups. Increasing the representation from phylogenetic groups E, D1, B2, D2, G, and also cryptic clade groups could be recommended to obtain a more representative picture of *E. coli* genetic diversity. It can be proposed that this could be achieved with increased whole genome sequencing of strains from a range of soil, water, mineral substrate, plant, and wild animal sources in forests, wetlands, grasslands, and desert environments in tropical, sub-tropical, and temperate geographical locations around the world. *E. coli* have previously been sampled representatively for studies in areas such as these but whole genome sequencing was not carried out meaning phylogenetically unique and potentially highly interesting genetic insights in these genomes are not available for study. Examples include *E. coli* sampled at freshwater beaches (Walk et al. 2007), in cultivated soils (Hartmann et al. 2012), and from Australian domestic and wild animals including kangaroo, possum, waterfowl, emu, and deer (Ahmed et al. 2015). The reduced costs of whole genome sequencing

(Quainoo et al. 2018) mean that obtaining genome sequences from such isolates is now feasible.

In terms of the Chapter 4, it can be recommended that the developed MLST schemas and multiplex be developed into research tools. Future work would focus both developing these into stand-alone or components of existing functional programmatic tools and also carrying out tests to determine their accuracy in assigning clonal groups A-G, and cryptic clade phylogenetic groups to hundreds or thousands of *E. coli* genome sequences.

The potential for future work following the findings of Chapters 5 and 6 are similar. For ETT2 and the *eip*-locus, cluster phenotype experiments could be carried out where complete genes are knocked-out to cause a loss of function mutation while the strain is exposed to a range of variable environments and *in vitro* and *in vivo* animal models. These studies could provide more evidence to link a given highlighted gene and a certain pathogenic or ecological phenotype. For the ureter phenotype data, similar work could involve application of pathogenic and mutant strains to a wider range of *in vitro* and *in vivo* animal models. Such studies could provide more evidence to link each of the highlighted genes to specific urinary tract infection phenotypes and potentially pave the way for the development of treatments which target the products of these genes.

8. References

AHMED, W., GYAWALI, P. & TOZE, S. 2015. Quantitative PCR measurements of *Escherichia coli* including shiga toxin-producing *E. coli* (STEC) in animal feces and environmental waters. *Environmental science & technology,* 49**,** 3084-3090.

AHRENFELDT, J., SKAARUP, C., HASMAN, H., PEDERSEN, A. G., AARESTRUP, F. M. & LUND, O. 2017. Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods. *BMC genomics,* 18**,** 19.

AIZAWA, S.-I. 2001. Bacterial flagella and type III secretion systems. *FEMS Microbiology Letters,* 202**,** 157-164.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *Journal of molecular biology,* 215**,** 403-410.

ANDERSON, G. J. & VULPE, C. D. 2009. Mammalian iron transport. *Cellular and Molecular Life Sciences,* 66**,** 3241.

ARIFUZZAMAN, M., MAEDA, M., ITOH, A., NISHIKATA, K., TAKITA, C., SAITO, R., ARA, T., NAKAHIGASHI, K., HUANG, H.-C. & HIRAI, A. 2006. Large-scale identification of protein–protein interaction of *Escherichia coli* K-12. *Genome research,* 16**,** 686-691.

ARITA, M. 2003. In silico atomic tracing by substrate-product relationships in *Escherichia coli* intermediary metabolism. *Genome Research,* 13**,** 2455-2466.

BARRICK, J. E., YU, D. S., YOON, S. H., JEONG, H., OH, T. K., SCHNEIDER, D., LENSKI, R. E. & KIM, J. F. 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature,* 461**,** 1243.

BENS, M., VIMONT, S., BEN MKADDEM, S., CHASSIN, C., GOUJON, J. M., BALLOY, V., CHIGNARD, M., WERTS, C. & VANDEWALLE, A. 2014. Flagellin/TLRS 5 signalling activates renal collecting duct cells and facilitates invasion and cellular translocation of uropathogenic *Escherichia coli*. *Cellular microbiology,* 16**,** 1503-1517.

BENSON, D. A., CAVANAUGH, M., CLARK, K., KARSCH-MIZRACHI, I., OSTELL, J., PRUITT, K. D. & SAYERS, E. W. 2017. GenBank. *Nucleic acids research*.

BENSON, D. A., CLARK, K., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J. & SAYERS, E. W. 2015. GenBank. *Nucleic acids research,* 43**,** D30.

BERGTHORSSON, U. & OCHMAN, H. 1995. Heterogeneity of genome sizes among natural isolates of *Escherichia coli*. *Journal of bacteriology,* 177**,** 5784-5789.

BERGTHORSSON, U. & OCHMAN, H. 1998. Distribution of chromosome length variation in natural isolates of *Escherichia coli*. *Molecular Biology and Evolution,* 15**,** 6-16.

BIEN, J., SOKOLOVA, O. & BOZKO, P. 2012. Role of uropathogenic *Escherichia coli* virulence factors in development of urinary tract infection and kidney damage. *International journal of nephrology,* 2012.

BINGEN, E., BONACORSI, S., BRAHIMI, N., DENAMUR, E. & ELION, J. 1997. Virulence patterns of *Escherichia coli* K1 strains associated with neonatal meningitis. *Journal of clinical microbiology,* 35**,** 2981-2982.

BISERCIĆ, M., FEUTRIER, J. Y. & REEVES, P. R. 1991. Nucleotide sequences of the *gnd* genes from nine natural isolates of *Escherichia coli*: evidence of intragenic recombination as a contributing factor in the evolution of the polymorphic *gnd* locus. *Journal of bacteriology,* 173**,** 3894-3900.

BLATTNER, F. R., PLUNKETT, G., BLOCH, C. A., PERNA, N. T., BURLAND, V., RILEY, M., COLLADO-VIDES, J., GLASNER, J. D., RODE, C. K. & MAYHEW, G. F. 1997. The complete genome sequence of *Escherichia coli* K-12. *science,* 277**,** 1453-1462.

BLOCKER, A., KOMORIYA, K. & AIZAWA, S.-I. 2003. Type III secretion systems and bacterial flagella: insights into their function from structural similarities. *Proceedings of the National Academy of Sciences,* 100**,** 3027-3030.

BLYTON, M. D., BANKS, S. C., PEAKALL, R. & GORDON, D. M. 2013. High temporal variability in commensal *Escherichia coli* strain communities of a herbivorous marsupial. *Environmental microbiology,* 15**,** 2162-2172.

BOHLIN, J., BRYNILDSRUD, O. B., SEKSE, C. & SNIPEN, L. 2014. An evolutionary analysis of genome expansion and pathogenicity in *Escherichia coli*. *BMC genomics,* 15**,** 882.

BRUMBAUGH, A. R., SMITH, S. N., SUBASHCHANDRABOSE, S., HIMPSL, S. D., HAZEN, T. H., RASKO, D. A. & MOBLEY, H. L. 2015. Blocking yersiniabactin import attenuates extraintestinal pathogenic *Escherichia coli* in cystitis and pyelonephritis and represents a novel target to prevent urinary tract infection. *Infection and immunity***,** IAI. 02904-14.

BRYNILDSRUD, O., BOHLIN, J., SCHEFFER, L. & ELDHOLM, V. 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome biology,* 17**,** 238.

BUCHRIESER, C., BROSCH, R., BACH, S., GUIYOULE, A. & CARNIEL, E. 1998. The high-pathogenicity island of *Yersinia pseudotuberculosis* can be inserted into any of the three chromosomal *asn* tRNA genes. *Molecular microbiology,* 30**,** 965-978.

BURDET, C., CLERMONT, O., BONACORSI, S., LAOUÉNAN, C., BINGEN, E., AUJARD, Y., MENTRÉ, F., LEFORT, A., DENAMUR, E. & GROUP, C. 2014. *Escherichia coli* bacteremia in children: age and portal of entry are the main predictors of severity. *The Pediatric infectious disease journal,* 33**,** 872-879.

BURRUS, V., PAVLOVIC, G., DECARIS, B. & GUÉDON, G. 2002. Conjugative transposons: the tip of the iceberg. *Molecular microbiology,* 46**,** 601-610.

CARTER, M. Q. & PHAM, A. 2018. Complete Genome Sequence of a Natural *Escherichia coli* O145: H11 Isolate That Belongs to Phylogroup A. *Genome announcements,* 6**,** e00349-18.

CARVER, T. J., RUTHERFORD, K. M., BERRIMAN, M., RAJANDREAM, M.-A., BARRELL, B. G. & PARKHILL, J. 2005. ACT: the Artemis comparison tool. *Bioinformatics,* 21**,** 3422-3423.

CHAIN, P., GRAFHAM, D., FULTON, R., FITZGERALD, M., HOSTETLER, J., MUZNY, D., ALI, J., BIRREN, B., BRUCE, D. & BUHAY, C. 2009. Genome project standards in a new era of sequencing. *Science,* 326**,** 236-237.

CHATZOU, M., FLODEN, E. W., DI TOMMASO, P., GASCUEL, O. & NOTREDAME, C. 2018. Generalized bootstrap supports for phylogenetic analyses of protein sequences incorporating alignment uncertainty. *Systematic Biology***,** syx096.

CHAUDHURI, R. R. & HENDERSON, I. R. 2012. The evolution of the *Escherichia coli* phylogeny. *Infection, Genetics and Evolution,* 12**,** 214-226.

CHAUDHURI, R. R., SEBAIHIA, M., HOBMAN, J. L., WEBBER, M. A., LEYTON, D. L., GOLDBERG, M. D., CUNNINGHAM, A. F., SCOTT-TUCKER, A., FERGUSON, P. R. & THOMAS, C. M. 2010. Complete genome sequence and comparative metabolic profiling of the prototypical enteroaggregative *Escherichia coli* strain 042. *PloS one,* 5**,** e8801.

CHENG, D., ZHU, S., SU, Z., ZUO, W. & LU, H. 2012. Prevalence and isoforms of the pathogenicity island ETT2 among *Escherichia coli* isolates from colibacillosis in pigs and mastitis in cows. *Current microbiology,* 64**,** 43-49.

CHO, S., HIOTT, L. M., BARRETT, J. B., MCMILLAN, E. A., HOUSE, S. L., HUMAYOUN, S. B., ADAMS, E. S., JACKSON, C. R. & FRYE, J. G. 2018. Prevalence and characterization of *Escherichia coli* isolated from the Upper Oconee Watershed in Northeast Georgia. *PloS one,* 13**,** e0197005.

CLERMONT, O., BONACORSI, S. & BINGEN, E. 2000. Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Applied and environmental microbiology,* 66**,** 4555-4558.

CLERMONT, O., CHRISTENSON, J. K., DENAMUR, E. & GORDON, D. M. 2013. The C lermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environmental microbiology reports,* 5**,** 58-65.

CLERMONT, O., GORDON, D. & DENAMUR, E. 2015. Guide to the various phylogenetic classification schemes for *Escherichia coli* and the correspondence among schemes. *Microbiology,* 161**,** 980-988.

COCK, P. J., ANTAO, T., CHANG, J. T., CHAPMAN, B. A., COX, C. J., DALKE, A., FRIEDBERG, I., HAMELRYCK, T., KAUFF, F. & WILCZYNSKI, B. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics,* 25**,** 1422-1423.

COPELAND, P. R. 2003. Regulation of gene expression by stop codon recoding: selenocysteine. *Gene*, *312*, 17-25.

COREN, J. S., PIERCE, J. C., & STERNBERG, N. (1995). Headful packaging revisited: the packaging of more than one DNA molecule into a bacteriophage P1 head. *Journal of molecular biology*, 249(1), 176-184.

CORNELIS, G. R. & VAN GIJSEGEM, F. 2000. Assembly and function of type III secretory systems. *Annual Reviews in Microbiology,* 54**,** 735-774.

CORNELLS, G. R. 2000. Type III secretion: a bacterial device for close combat with cells of their eukaryotic host. *Philosophical Transactions of the Royal Society of London B: Biological Sciences,* 355**,** 681-693.

COVE-SMITH, A. & ALMOND, M. 2007. Management of urinary tract infections in the elderly. *Trends in Urology, Gynaecology & Sexual Health,* 12**,** 31-34.

COWLEY, L. A., DALLMAN, T. J., FITZGERALD, S., IRVINE, N., ROONEY, P. J., MCATEER, S. P., DAY, M., PERRY, N. T., BONO, J. L. & JENKINS, C. 2016. Short-term evolution of Shiga toxin-producing *Escherichia coli* O157: H7 between two food-borne outbreaks. *Microbial genomics,* 2.

CRICK, F. H. 1968. The origin of the genetic code *Journal of molecular biology*, *38*(3), 367-379.

CROSA, J. H. 1989. Genetics and molecular biology of siderophore-mediated iron transport in bacteria. *Microbiological reviews,* 53**,** 517-530.

CROSSMAN, L. C., CHAUDHURI, R. R., BEATSON, S. A., WELLS, T. J., DESVAUX, M., CUNNINGHAM, A. F., PETTY, N. K., MAHON, V., BRINKLEY, C. & HOBMAN, J. L. 2010. A commensal gone bad: complete genome sequence of the prototypical enterotoxigenic *Escherichia coli* strain H10407. *Journal of bacteriology,* 192**,** 5822-5831.

CROXEN, M. A., LAW, R. J., SCHOLZ, R., KEENEY, K. M., WLODARSKA, M. & FINLAY, B. B. 2013. Recent advances in understanding enteric pathogenic *Escherichia coli*. *Clinical microbiology reviews,* 26**,** 822-880.

DA SILVA, G. J. & MENDONÇA, N. 2012. Association between antimicrobial resistance and virulence in *Escherichia coli. Virulence,* 3**,** 18-28.

DANESE, P. N., PRATT, L. A. & KOLTER, R. 2001. Biofilm formation as a developmental process. *Methods in enzymology.* Elsevier.

DAVIS, J. M., CARVALHO, H. M., RASMUSSEN, S. B. & O'BRIEN, A. D. 2006. Cytotoxic necrotizing factor type 1 delivered by outer membrane vesicles of uropathogenic *Escherichia coli* attenuates polymorphonuclear leukocyte antimicrobial activity and chemotaxis. *Infection and immunity,* 74**,** 4401-4408.

DAVIS, J. M., RASMUSSEN, S. B., & O'BRIEN, A. D. 2005. Cytotoxic necrotizing factor type 1 production by uropathogenic *Escherichia coli* modulates polymorphonuclear leukocyte function. *Infection and immunity,* 73, 5301-5310.

DE BEEN, M., PINHOLT, M., TOP, J., BLETZ, S., MELLMANN, A., VAN SCHAIK, W., BROUWER, E., ROGERS, M., KRAAT, Y. & BONTEN, M. 2015. A core genome MLST scheme for high-resolution typing of *Enterococcus faecium*. *Journal of clinical microbiology***,** JCM. 01946-15.

DENG, W., LI, Y., VALLANCE, B. A. & FINLAY, B. B. 2001. Locus of enterocyte effacement from *Citrobacter rodentium*: sequence analysis and evidence for horizontal transfer among attaching and effacing pathogens. *Infection and immunity,* 69**,** 6323-6335.

DHAKAL, B., KULESUS, R. & MULVEY, M. 2008. Mechanisms and consequences of bladder cell invasion by uropathogenic *Escherichia coli*. *European journal of clinical investigation,* 38**,** 2-11.

DHAKAL, B. K. & MULVEY, M. A. 2012. The UPEC pore-forming toxin α-hemolysin triggers proteolysis of host proteins to disrupt cell adhesion, inflammatory, and survival pathways. *Cell host & microbe,* 11**,** 58-69.

DIDELOT, X. & FALUSH, D. 2006. Inference of bacterial microevolution using multilocus sequence data. *Genetics*.

DIDELOT, X. & MAIDEN, M. C. 2010. Impact of recombination on bacterial evolution. *Trends in microbiology,* 18**,** 315-322.

DIDELOT, X., MÉRIC, G., FALUSH, D. & DARLING, A. E. 2012. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC genomics,* 13**,** 256.

DIJKSHOORN, L., URSING, B. & URSING, J. 2000. Strain, clone and species: comments on three basic concepts of bacteriology. *Journal of medical microbiology,* 49**,** 397-401.

DOBRINDT, U. 2005. (Patho-) genomics of *Escherichia coli*. *International Journal of Medical Microbiology,* 295**,** 357-371.

DOBRINDT, U., HOCHHUT, B., HENTSCHEL, U. & HACKER, J. 2004. Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology,* 2**,** 414.

DONNENBERG, M. S. 2002. *Escherichia coli*: virulence mechanisms of a versatile pathogen. Amsterdam; Boston: Academic press.

DOZOIS, C. M., DAIGLE, F. & CURTISS, R. 2003. Identification of pathogen-specific and conserved genes expressed in vivo by an avian pathogenic *Escherichia coli* strain. *Proceedings of the National Academy of Sciences,* 100**,** 247-252.

DUCHÊNE, S., GEOGHEGAN, J. L., HOLMES, E. C. & HO, S. Y. 2016. Estimating evolutionary rates using time-structured data: a general comparison of phylogenetic methods. *Bioinformatics,* 32**,** 3375-3379.

DUNNE, K. A., CHAUDHURI, R. R., ROSSITER, A. E., BERIOTTO, I., BROWNING, D. F., SQUIRE, D., CUNNINGHAM, A. F., COLE, J. A., LOMAN, N. & HENDERSON, I. R. 2017. Sequencing a piece of history: complete genome sequence of the original *Escherichia coli* strain. *Microbial genomics,* 3.

DZIVA, F., HAUSER, H., CONNOR, T. R., VAN DIEMEN, P. M., PRESCOTT, G., LANGRIDGE, G. C., ECKERT, S., CHAUDHURI, R. R., EWERS, C. & MELLATA, M. 2013. Sequencing and functional annotation of avian pathogenic *Escherichia coli* serogroup O78 strains reveal the evolution of *E. coli* lineages pathogenic for poultry via distinct mechanisms. *Infection and immunity,* 81**,** 838-849.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research,* 32**,** 1792-1797.

ELLIOTT, S. J., WAINWRIGHT, L. A., MCDANIEL, T. K., JARVIS, K. G., DENG, Y., LAI, L. C., MCNAMARA, B. P., DONNENBERG, M. S. & KAPER, J. B. 1998. The complete sequence of the locus of enterocyte effacement (LEE) from enteropathogenic *Escherichia coli* E2348/69. *Molecular microbiology,* 28**,** 1-4.

ENRIGHT, M. C., & SPRATT, B. G. 1999. Multilocus sequence typing. *Trends in microbiology,* 7(12), 482-487.

ENRIGHT, A. J., VAN DONGEN, S. & OUZOUNIS, C. A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research,* 30**,** 1575-1584.

ENRIGHT, M. C. & SPRATT, B. G. 1999. Multilocus sequence typing. *Trends in microbiology,* 7**,** 482-487.

EPSTEIN, W. 2016. The KdpD sensor kinase of *Escherichia coli* responds to several distinct signals to turn on expression of the Kdp transport system. *Journal of bacteriology,* 198**,** 212-220.

ESCOBAR-PÁRAMO, P., GRENET, K., LE MENAC'H, A., RODE, L., SALGADO, E., AMORIN, C., GOURIOU, S., PICARD, B., RAHIMY, M. C. & ANDREMONT, A. 2004. Large-scale population structure of human commensal *Escherichia coli* isolates. *Applied and environmental microbiology,* 70**,** 5698-5700.

FEIST, A. M., HENRY, C. S., REED, J. L., KRUMMENACKER, M., JOYCE, A. R., KARP, P. D., BROADBELT, L. J., HATZIMANIKATIS, V. & PALSSON, B. Ø. 2007. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular systems biology,* 3**,** 121.

FERDOUS M, FRIEDRICH AW, GRUNDMANN H, DE BOER RF, CROUGHS PD, ISLAM MA, KLUYTMANS-VAN DEN BERGH MF, KOOISTRA-SMID AM, & ROSSEN JW. 2016. Molecular characterization and phylogeny of Shiga toxin–producing Escherichia coli isolates obtained from two Dutch regions using whole genome sequencing. *Clinical Microbiology and Infection,* 1, 22.

FLEISCHMANN, R. D., ADAMS, M. D., WHITE, O., CLAYTON, R. A., KIRKNESS, E. F., KERLAVAGE, A. R., BULT, C. J., TOMB, J.-F., DOUGHERTY, B. A. & MERRICK, J. M. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science,* 269**,** 496-512.

FLOYD, R. V., UPTON, M., HULTGREN, S. J., WRAY, S., BURDYGA, T. V. & WINSTANLEY, C. 2012. *Escherichia coli–*mediated impairment of ureteric contractility is uropathogenic *E. coli* specific. *The Journal of infectious diseases,* 206**,** 1589-1596.

FLOYD, R. V., WINSTANLEY, C., BAKRAN, A., WRAY, S. & BURDYGA, T. V. 2010. Modulation of ureteric Ca signaling and contractility in humans and rats by uropathogenic *E. coli. American Journal of Physiology-Renal Physiology,* 298**,** F900-F908.

FORDE, B. M., ZAKOUR, N. L. B., STANTON-COOK, M., PHAN, M.-D., TOTSIKA, M., PETERS, K. M., CHAN, K. G., SCHEMBRI, M. A., UPTON, M. & BEATSON, S. A. 2014. The complete genome sequence of *Escherichia coli* EC958: a high quality reference sequence for the globally disseminated multidrug resistant *E. coli* O25b: H4-ST131 clone. *PLoS One,* 9**,** e104400.

FOXMAN, B. 2002. Epidemiology of urinary tract infections: incidence, morbidity, and economic costs. *The American journal of medicine,* 113**,** 5-13.

FOXMAN, B. 2010. The epidemiology of urinary tract infection. *Nature Reviews Urology,* 7**,** 653.

FU, L., NIU, B., ZHU, Z., WU, S. & LI, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics,* 28**,** 3150-3152.

GALTIER, N., GOUY, M. & GAUTIER, C. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Bioinformatics,* 12**,** 543-548.

GARCÍA-FERNÁNDEZ, I., MIRALLES-CUEVAS, S., OLLER, I., MALATO, S., FERNÁNDEZ-IBÁÑEZ, P. & POLO-LÓPEZ, M. I. 2018. Inactivation of *E. coli* and *E. faecalis* by solar photo-Fenton with EDDS complex at neutral pH in municipal wastewater effluents. *Journal of hazardous materials*.

GENEVAUX, P., BAUDA, P., DUBOW, M. S. & OUDEGA, B. 1999. Identification of Tn 10 insertions in the *dsbA* gene affecting *Escherichia coli* biofilm formation. *FEMS microbiology letters,* 173**,** 403-409.

GERLACH, R. G. & HENSEL, M. 2007. Protein secretion systems and adhesins: the molecular armory of Gram-negative pathogens. *International Journal of Medical Microbiology,* 297**,** 401-415.

GIRÓN, J. A., TORRES, A. G., FREER, E. & KAPER, J. B. 2002. The flagella of enteropathogenic *Escherichia coli* mediate adherence to epithelial cells. *Molecular microbiology,* 44**,** 361-379.

GOMES, C. M., GIUFFRE, A., FORTE, E., VICENTE, J. B., SARAIVA, L. M., BRUNORI, M. & TEIXEIRA, M. 2002. A novel type of nitric oxide reductase: *Escherichia coli* flavorubredoxin. *Journal of Biological Chemistry,* 277, 25273-25276.

GOMEZ-CRUZ, J., NAIR, S., MANJARREZ-HERNANDEZ, A., GAVILANES-PARRA, S., ASCANIO, G. & ESCOBEDO, C. 2018. Cost-effective flow-through nanohole array-based biosensing platform for the label-free detection of uropathogenic *E. coli* in real time. *Biosensors and Bioelectronics,* 106**,** 105-110.

GONZÁLEZ-GONZÁLEZ, A., SÁNCHEZ-REYES, L. L., SAPIEN, G. D., EGUIARTE, L. E. & SOUZA, V. 2013. Hierarchical clustering of genetic diversity associated to different levels of mutation and recombination in *Escherichia coli*: a study based on Mexican isolates. *Infection, Genetics and Evolution,* 13**,** 187-197.

GOTO, D. K. & YAN, T. 2011. Genotypic diversity of *Escherichia coli* in the water and soil of tropical watersheds in Hawaii. *Applied and environmental microbiology***,** AEM. 02140-10.

GRANA, L., DONNELLAN, W. L. & SWENSON, O. 1968. Effects of gram-negative bacteria on ureteral structure and function. *The Journal of urology,* 99**,** 539-550.

GUY, L., JERNBERG, C., NORLING, J. A., IVARSSON, S., HEDENSTRÖM, I., MELEFORS, Ö., LILJEDAHL, U., ENGSTRAND, L. & ANDERSSON, S. G. 2013. Adaptive mutations and replacements of virulence traits in *the Escherichia coli* O104: H4 outbreak population. *PLoS One,* 8**,** e63027.

GUYER, D. M., HENDERSON, I. R., NATARO, J. P. & MOBLEY, H. L. 2000. Identification of sat, an autotransporter toxin produced by uropathogenic *Escherichia coli*. *Molecular microbiology,* 38**,** 53-66.

GUYER, D. M., RADULOVIC, S., JONES, F.-E. & MOBLEY, H. L. 2002. Sat, the secreted autotransporter toxin of uropathogenic *Escherichia coli*, is a vacuolating cytotoxin for bladder and kidney epithelial cells. *Infection and immunity,* 70**,** 4539-4546.

HACKER, J., BLUM-OEHLER, G., MÜHLDORFER, I. & TSCHÄPE, H. 1997. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Molecular microbiology,* 23**,** 1089-1097.

HACKER, J. & KAPER, J. B. 1999. The concept of pathogenicity islands. *Pathogenicity islands and other mobile virulence elements.* American Society of Microbiology.

HAGEN, J. B. 2000. The origins of bioinformatics. *Nature Reviews Genetics,* 1**,** 231. HAIKO, J. & WESTERLUND-WIKSTRÖM, B. 2013. The role of the bacterial flagellum in adhesion and virulence. *Biology,* 2**,** 1242-1267.

HANAHAN, D. 1983. Studies on transformation of *Escherichia coli* with plasmids. *Journal of molecular biology,* 166**,** 557-580.

HARTLEIB, S., PRAGER, R., HEDENSTROM, I., LOFDAHL, S. & TSCHAPE, H. 2003. Prevalence of the new, SPI1-like, pathogenicity island ETT2 among *Escherichia coli. International journal of medical microbiology,* 292**,** 487.

HARTMANN, A., AMOUREUX, L., LOCATELLI, A., DEPRET, G., JOLIVET, C., GUENEAU, E. & NEUWIRTH, C. 2012. Occurrence of CTX-M producing *Escherichia coli* in soils, cattle, and farm environment in France (Burgundy region). *Frontiers in microbiology,* 3**,** 83.

HASEGAWA, M., KISHINO, H. & SAITOU, N. 1991. On the maximum likelihood method in molecular phylogenetics. *Journal of molecular evolution,* 32**,** 443-445.

HAYASHI, T., MAKINO, K., OHNISHI, M., KUROKAWA, K., ISHII, K., YOKOYAMA, K., HAN, C.-G., OHTSUBO, E., NAKAYAMA, K. & MURATA, T. 2001. Complete genome sequence of enterohemorrhagic *Eschelichia coli* O157: H7 and genomic comparison with a laboratory strain K-12. *DNA research,* 8**,** 11-22.

HAYASHI, T., MAKINO, K., OHNISHI, M., KUROKAWA, K., ISHII, K., YOKOYAMA, K., HAN, C.-G., OHTSUBO, E., NAKAYAMA, K. & MURATA, T. 2001c. Complete genome sequence of enterohemorrhagic *Eschelichia coli* O157: H7 and genomic comparison with a laboratory strain K-12. *DNA research,* 8**,** 11-22.

HERZER, P. J., INOUYE, S., INOUYE, M. & WHITTAM, T. S. 1990. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. *Journal of bacteriology,* 172**,** 6175-6181.

HO, S. N., HUNT, H. D., HORTON, R. M., PULLEN, J. K., & PEASE, L. R. 1989. Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene*, 77(1), 51-59.

HOLZINGER, A., DEHMER, M. & JURISICA, I. 2014. Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC bioinformatics,* 15**,** I1.

HOOTON, T. M. & STAMM, W. E. 1997. Diagnosis and treatment of uncomplicated urinary tract infection. *Infectious Disease Clinics,* 11**,** 551-581.

HOSOKI, R., MATSUKI, N. & KIMURA, H. 1997. The possible role of hydrogen sulfide as an endogenous smooth muscle relaxant in synergy with nitric oxide. *Biochemical and biophysical research communications,* 237**,** 527-531.

HOWARD, C. & GLYNN, A. 1971. The virulence for mice of strains of *Escherichia coli* related to the effects of K antigens on their resistance to phagocytosis and killing by complement. *Immunology,* 20**,** 767.

HUECK, C. J. 1998. Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiology and molecular biology reviews,* 62**,** 379-433.

HUERTA-CEPAS, J., SERRA, F. & BORK, P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution,* 33**,** 1635-1638.

HUJA, S., OREN, Y., TROST, E., BRZUSZKIEWICZ, E., BIRAN, D., BLOM, J., GOESMANN, A., GOTTSCHALK, G., HACKER, J. & RON, E. Z. 2015. Genomic avenue to avian colisepticemia. *MBio,* 6**,** e01681-14.

HYATT, D., CHEN, G.-L., LOCASCIO, P. F., LAND, M. L., LARIMER, F. W. & HAUSER, L. J. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics,* 11**,** 119.

IDESES, D., GOPHNA, U., PAITAN, Y., CHAUDHURI, R. R., PALLEN, M. J. & RON, E. Z. 2005. A degenerate type III secretion system from septicemic *Escherichia coli* contributes to pathogenesis. *Journal of bacteriology,* 187**,** 8164-8171.

IGUCHI, A., THOMSON, N. R., OGURA, Y., SAUNDERS, D., OOKA, T., HENDERSON, I. R., HARRIS, D., ASADULGHANI, M., KUROKAWA, K. & DEAN, P. 2009. Complete genome sequence and comparative genome analysis of enteropathogenic *Escherichia coli* O127: H6 strain E2348/69. *Journal of Bacteriology,* 191**,** 347-354.

JAIN, M., OLSEN, H. E., PATEN, B., & AKESON, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome biology*, 17(1), 239.

JANECKO, N., HALOVA, D., JAMBOROVA, I., PAPOUSEK, I., MASARIKOVA, M., DOLEJSKA, M. & LITERAK, I. 2018a. Occurrence of plasmid-mediated quinolone resistance genes in *Escherichia coli* and *Klebsiella spp*. recovered from *Corvus brachyrhynchos* and *Corvus corax* roosting in Canada. *Letters in applied microbiology*.

JARVIS, K. G., GIRON, J. A., JERSE, A. E., MCDANIEL, T. K., DONNENBERG, M. S. & KAPER, J. B. 1995. Enteropathogenic *Escherichia coli* contains a putative type III secretion system necessary for the export of proteins involved in attaching and effacing lesion formation. *Proceedings of the National Academy of Sciences,* 92**,** 7996-8000.

JAY, M. T., COOLEY, M., CARYCHAO, D., WISCOMB, G. W., SWEITZER, R. A., CRAWFORD-MIKSZA, L., FARRAR, J. A., LAU, D. K., O'CONNELL, J. & MILLINGTON, A. 2007. *Escherichia coli* O157: H7 in feral swine near spinach fields and cattle, central California coast. *Emerging infectious diseases,* 13**,** 1908.

JERSE, A. E., YU, J., TALL, B. D. & KAPER, J. B. 1990. A genetic locus of enteropathogenic *Escherichia coli* necessary for the production of attaching and effacing lesions on tissue culture cells. *Proceedings of the National Academy of Sciences,* 87**,** 7839-7843.

JESSE, H. E., ROBERTS, I. S. & CAVET, J. S. 2014. Metal Ion Homeostasis in Listeria monocytogenes and Importance in Host–Pathogen Interactions. *Advances in microbial physiology.* Elsevier.

JIANG, M., WAN, Q., LIU, R., LIANG, L., CHEN, X., WU, M., ZHANG, H., CHEN, K., MA, J. & WEI, P. 2014. Succinic acid production from corn stalk hydrolysate in an *E. coli* mutant generated by atmospheric and room-temperature plasmas and metabolic evolution strategies. *Journal of industrial microbiology & biotechnology,* 41**,** 115-123.

JORDAN, I. K., ROGOZIN, I. B., WOLF, Y. I. & KOONIN, E. V. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome research,* 12**,** 962-968.

JORES, J., RUMER, L. & WIELER, L. H. 2004. Impact of the locus of enterocyte effacement pathogenicity island on the evolution of pathogenic *Escherichia coli*. *International journal of medical microbiology,* 294**,** 103-113.

JU, J., RUAN, C., FULLER, C. W., GLAZER, A. N., & MATHIES, R. A. 1995. Fluorescence energy transfer dye-labeled primers for DNA sequencing and analysis. Proceedings of the National Academy of Sciences, 92, 4347-4351.

KAAS, R. S., FRIIS, C., USSERY, D. W. & AARESTRUP, F. M. 2012. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC genomics,* 13**,** 577.

KALLONEN, T., BRODRICK, H. J., HARRIS, S. R., CORANDER, J., BROWN, N. M., MARTIN, V., PEACOCK, S. J. & PARKHILL, J. 2017. Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome research*.

KANEHISA, M. & GOTO, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research,* 28**,** 27-30.

KAPER, J. B., NATARO, J. P. & MOBLEY, H. L. 2004a. Pathogenic *Escherichia coli*. *Nature reviews microbiology,* 2**,** 123.

KATOULI, M. 2010. Population structure of gut *Escherichia coli* and its role in development of extra-intestinal infections. *Iranian journal of microbiology,* 2**,** 59.

KENNY, B. 2001. Mechanism of action of EPEC type III effector molecules. *International Journal of Medical Microbiology,* 291**,** 469-477.

KEYSER, P., ELOFSSON, M., ROSELL, S. & WOLF-WATZ, H. 2008. Virulence blockers as alternatives to antibiotics: type III secretion inhibitors against Gram-negative bacteria. *Journal of internal medicine,* 264**,** 17-29.

KLEMM, P. & SCHEMBRI, M. A. 2000a. Bacterial adhesins: function and structure. *International Journal of Medical Microbiology,* 290**,** 27-35.

KOLBE, D. L. & EDDY, S. R. 2011a. Fast filtering for RNA homology search. *Bioinformatics,* 27**,** 3102-3109.

KONDRASHOV, F. A., ROGOZIN, I. B., WOLF, Y. I. & KOONIN, E. V. 2002. Selection in the evolution of gene duplications. *Genome biology,* 3**,** research0008. 1.

KOONIN, E. V., & NOVOZHILOV, A. S. 2009. Origin and evolution of the genetic code: the universal enigma. *IUBMB life, 61*(2), 99-111

KOTLOFF, K. L., NATARO, J. P., BLACKWELDER, W. C., NASRIN, D., FARAG, T. H., PANCHALINGAM, S., WU, Y., SOW, S. O., SUR, D. & BREIMAN, R. F. 2013. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *The Lancet,* 382**,** 209-222.

KUMAR, S., NEI, M., DUDLEY, J. & TAMURA, K. 2008. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in bioinformatics,* 9**,** 299-306.

KUSECEK, B., WLOCH, H., MERCER, A., VAISÄNEN, V., PLUSCHKE, G., KORHONEN, T. & ACHTMAN, M. 1984. Lipopolysaccharide, capsule, and fimbriae as virulence factors among O1, O7, O16, O18, or O75 and K1, K5, or K100 *Escherichia coli*. *Infection and immunity,* 43**,** 368-379.

KÖHLER, C.-D. & DOBRINDT, U. 2011. What defines extraintestinal pathogenic *Escherichia coli*? *International Journal of Medical Microbiology,* 301**,** 642-647.

LAGESEN, K., HALLIN, P., RØDLAND, E. A., STÆRFELDT, H.-H., ROGNES, T. & USSERY, D. W. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research,* 35**,** 3100-3108.

LAING, C., BUCHANAN, C., TABOADA, E. N., ZHANG, Y., KROPINSKI, A., VILLEGAS, A., THOMAS, J. E. & GANNON, V. P. 2010. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC bioinformatics,* 11**,** 461.

LANATA, C. F., FISHER-WALKER, C. L., OLASCOAGA, A. C., TORRES, C. X., ARYEE, M. J. & BLACK, R. E. 2013a. Global causes of diarrheal disease mortality in children< 5 years of age: a systematic review. *PloS one*, 8**,** e72788.

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M. & FITZHUGH, W. 2001. Initial sequencing and analysis of the human genome. *Nature*, 409**,** 860-921.

LANE, M. & MOBLEY, H. 2007. Role of P-fimbrial-mediated adherence in pyelonephritis and persistence of uropathogenic *Escherichia coli* (UPEC) in the mammalian kidney. *Kidney international,* 72**,** 19-25.

LANGRIDGE, G. C., PHAN, M.-D., TURNER, D. J., PERKINS, T. T., PARTS, L., HAASE, J., CHARLES, I., MASKELL, D. J., PETERS, S. E. & DOUGAN, G. 2009. Simultaneous assay of every *Salmonella Typhi* gene using one million transposon mutants. *Genome research*.

LASLETT, D. & CANBACK, B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic acids research,* 32**,** 11-16.

LEE, H., DOAK, T. G., POPODI, E., FOSTER, P. L. & TANG, H. 2016. Insertion sequence-caused large-scale rearrangements in the genome of *Escherichia coli*. *Nucleic acids research,* 44**,** 7109-7119.

LEIMBACH, A., HACKER, J. & DOBRINDT, U. 2013. *E. coli* as an all-rounder: the thin line between commensalism and pathogenicity. *Between pathogenicity and commensalism.* Springer.

LEWIS, J. P. 2010. Metal uptake in host–pathogen interactions: role of iron in *Porphyromonas gingivalis* interactions with host organisms. *Periodontology 2000,* 52**,** 94-116.

LI, W. 1997. *Molecular evolution,* Sinauer associates incorporated.

LITWIN, C. M. & CALDERWOOD, S. 1993. Role of iron in regulation of virulence genes. *Clinical microbiology reviews,* 6**,** 137-149.

LIU, Y., YU, F., WU, W., XIE, Y., WANG, X., ZHANG, X., CHEN, X. & ZONG, Z. 2015. OXA-181-producing *Escherichia coli* in China: the first report and characterization using whole genome sequencing. *Antimicrobial agents and chemotherapy***,** AAC. 00442-15.

LIÉVIN-LE MOAL, V., COMENGE, Y., RUBY, V., AMSELLEM, R., NICOLAS, V. & SERVIN, A. L. 2011. Secreted autotransporter toxin (Sat) triggers autophagy in epithelial cells that relies on cell detachment. *Cellular microbiology,* 13**,** 992-1013.

LOMAN, N. J., CONSTANTINIDOU, C., CHAN, J. Z., HALACHEV, M., SERGEANT, M., PENN, C. W., ROBINSON, E. R. & PALLEN, M. J. 2012. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology,* 10**,** 599.

LOVERING, A. L., DE CASTRO, L. H., LIM, D. & STRYNADKA, N. C. 2007. Structural insight into the transglycosylation step of bacterial cell-wall biosynthesis. *Science,* 315**,** 1402-1405.

LUKJANCENKO, O., WASSENAAR, T. M. & USSERY, D. W. 2010. Comparison of 61 sequenced *Escherichia coli* genomes. *Microbial ecology,* 60**,** 708-720.

LUO, C., WALK, S. T., GORDON, D. M., FELDGARDEN, M., TIEDJE, J. M. & KONSTANTINIDIS, K. T. 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences***,** 201015622.

LUPOLOVA, N., DALLMAN, T. J., MATTHEWS, L., BONO, J. L. & GALLY, D. L. 2016. Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates. *Proceedings of the National Academy of Sciences,* 113**,** 11312-11317.

LUZADER, D. H., WILLSEY, G. G., WARGO, M. J. & KENDALL, M. M. 2016. The ETT2-encoded regulator EtrB modulates enterohemorrhagic *Escherichia coli* virulence gene expression. *Infection and immunity***,** IAI. 00407-16.

MAHILLON, J. & CHANDLER, M. 1998. Insertion sequences. *Microbiology and molecular biology reviews,* 62**,** 725-774.

MAHILLON, J., LÉONARD, C. & CHANDLER, M. 1999. IS elements as constituents of bacterial genomes. *Research in microbiology,* 150**,** 675-687.

MAKINO, S.-I., TOBE, T., ASAKURA, H., WATARAI, M., IKEDA, T., TAKESHI, K. & SASAKAWA, C. 2003. Distribution of the secondary type III secretion system locus found in enterohemorrhagic *Escherichia coli* O157: H7 isolates among Shiga toxin-producing *E. coli* strains. *Journal of clinical microbiology,* 41**,** 2341-2347.

MANGIAMELE, P., NICHOLSON, B., WANNEMUEHLER, Y., SEEMANN, T., LOGUE, C. M., LI, G., TIVENDALE, K. A. & NOLAN, L. K. 2013. Complete genome sequence of the avian pathogenic *Escherichia coli* strain APEC O78. *Genome announcements,* 1**,** e00026-13.

MARGULIES, M., EGHOLM, M., ALTMAN, W. E., ATTIYA, S., BADER, J. S., BEMBEN, L. A., BERKA, J., BRAVERMAN, M. S., CHEN, Y.-J. & CHEN, Z. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437**,** 376.

MARRS, C. F., ZHANG, L. & FOXMAN, B. 2005. *Escherichia coli* mediated urinary tract infections: are there distinct uropathogenic *E. coli* (UPEC) pathotypes? *FEMS microbiology letters,* 252**,** 183-190.

MARTTINEN, P., HANAGE, W. P., CROUCHER, N. J., CONNOR, T. R., HARRIS, S. R., BENTLEY, S. D. & CORANDER, J. 2011. Detection of recombination events in bacterial genomes from large population samples. *Nucleic acids research,* 40**,** e6-e6.

MASLOW, J. N., MULLIGAN, M. E., & ARBEIT, R. D. 1993. Molecular epidemiology: application of contemporary techniques to the typing of microorganisms. *Clinical Infectious Diseases*, 153-162.

MASSON, P., HEREMANS, J. & SCHONNE, E. 1969. Lactoferrin, an iron-binbing protein Ni neutrophilic leukocytes. *Journal of Experimental Medicine,* 130**,** 643-658.

MASSOT, M., DAUBIÉ, A.-S., CLERMONT, O., JAURÉGUY, F., COUFFIGNAL, C., DAHBI, G., MORA, A., BLANCO, J., BRANGER, C. & MENTRÉ, F. 2016. Phylogenetic, virulence and antibiotic resistance characteristics of commensal strain populations of *Escherichia coli* from community subjects in the Paris area in 2010 and evolution over 30 years. *Microbiology,* 162**,** 642-650.

MASTRANGELO, D. & ISELIN, C. 2007. Urothelium dependent inhibition of rat ureter contractile activity. *The Journal of urology,* 178**,** 702-709.

MATAMOUROS, S., HAYDEN, H. S., HAGER, K. R., BRITTNACHER, M. J., LACHANCE, K., WEISS, E. J., POPE, C. E., IMHAUS, A.-F., MCNALLY, C. P. & BORENSTEIN, E. 2018. Adaptation of commensal proliferating *Escherichia coli* to the intestinal tract of young children with cystic fibrosis. *Proceedings of the National Academy of Sciences***,** 201714373.

MCDANIEL, T. K., JARVIS, K. G., DONNENBERG, M. S. & KAPER, J. B. 1995. A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. *Proceedings of the National Academy of Sciences,* 92**,** 1664-1668.

MCDANIEL, T. K. & KAPER, J. B. 1997. A cloned pathogenicity island from enteropathogenic *Escherichia coli* confers the attaching and effacing phenotype on *E. coli* K-12. *Molecular microbiology,* 23**,** 399-407.

MCNALLY, A., CHENG, L., HARRIS, S. R. & CORANDER, J. 2013. The evolutionary path to extraintestinal pathogenic, drug-resistant *Escherichia coli* is marked by drastic reduction in detectable recombination within the core genome. *Genome biology and evolution,* 5**,** 699-710.

MCNAMARA, B. P. & DONNENBERG, M. S. 1998. A novel proline-rich protein, EspF, is secreted from enteropathogenic *Escherichia coli* via the type III export pathway. *FEMS microbiology letters,* 166**,** 71-78.

MEDINI, D., DONATI, C., TETTELIN, H., MASIGNANI, V. & RAPPUOLI, R. 2005. The microbial pan-genome. *Current opinion in genetics & development,* 15**,** 589-594.

MELEKOS, M. D. & NABER, K. G. 2000. Complicated urinary tract infections. *International journal of antimicrobial agents,* 15**,** 247-256.

MIKHEYEV, A. S. & TIN, M. M. 2014. A first look at the Oxford Nanopore MinION sequencer. *Molecular ecology resources,* 14**,** 1097-1102.

MILKMAN, R. 1973. Electrophoretic variation in *Escherichia coli* from natural sources. *Science,* 182**,** 1024-1026.

MILLS, M., MEYSICK, K. C. & O'BRIEN, A. D. 2000. Cytotoxic necrotizing factor type 1 of uropathogenic *Escherichia coli* kills cultured human uroepithelial 5637 cells by an apoptotic mechanism. *Infection and immunity,* 68**,** 5869-5880.

MONK, J. & BOSI, E. 2018. Integration of comparative genomics with genome-scale metabolic modeling to investigate strain-specific phenotypical differences. *Metabolic Network Reconstruction and Modeling.* Springer.

MULVEY, M. A., SCHILLING, J. D., MARTINEZ, J. J. & HULTGREN, S. J. 2000. Bad bugs and beleaguered bladders: interplay between uropathogenic *Escherichia coli* and innate host defenses. *Proceedings of the National Academy of Sciences,* 97**,** 8829-8835.

MURPHY, K. C. 1998. Use of bacteriophage λ recombination functions to promote gene replacement in *Escherichia coli*. *Journal of bacteriology,* 180**,** 2063-2071.

MÜLLER, S., FELDMAN, M. F. & CORNELIS, G. R. 2001. The Type III secretion system of Gram-negative bacteria: a potential therapeutic target? *Expert opinion on therapeutic targets,* 5**,** 327-339.

MÜLLER, S., FELDMAN, M. F. & CORNELIS, G. R. 2001c. The Type III secretion system of Gram-negative bacteria: a potential therapeutic target? *Expert opinion on therapeutic targets,* 5**,** 327-339.

NAGY, G., ALTENHOEFER, A., KNAPP, O., MAIER, E., DOBRINDT, U., BLUM-OEHLER, G., BENZ, R., EMŐDY, L. & HACKER, J. 2006. Both α-haemolysin determinants contribute to full virulence of uropathogenic *Escherichia coli* strain 536. *Microbes and infection,* 8.

NASH, J. H., VILLEGAS, A., KROPINSKI, A. M., AGUILAR-VALENZUELA, R., KONCZY, P., MASCARENHAS, M., ZIEBELL, K., TORRES, A. G., KARMALI, M. A. & COOMBES, B. K. 2010. Genome sequence of adherent-invasive *Escherichia coli* and comparative genomic analysis with other *E. coli* pathotypes. *BMC genomics,* 11**,** 667.

NATARO, J. P. & KAPER, J. B. 1998. Diarrheagenic *Escherichia coli*. *Clinical microbiology reviews,* 11**,** 142-201.

NEHRA, M., SHARMA, R. K. & CHOUDHARY, M. 2017. An Overview on Molecular Basis of Genetic Recombination. *Int. J. Curr. Microbiol. App. Sci,* 6**,** 1154-1167.

NGELEKA, M., KWAGA, J., WHITE, D. G., WHITTAM, T. S., RIDDELL, C., GOODHOPE, R., POTTER, A. A. & ALLAN, B. 1996. *Escherichia coli* cellulitis in broiler chickens: clonal relationships among strains and analysis of virulence-associated factors of isolates from diseased birds. *Infection and immunity,* 64**,** 3118-3126.

NIELSEN, K., VAN ELSAS, J. & SMALLA, K. 2001. Dynamics, horizontal transfer and selection of novel DNA in bacterial populations in the phytosphere of transgenic plants. *Annals of Microbiology,* 51**,** 79-94.

NIPIČ, D., PODLESEK, Z., BUDIČ, M., ČRNIGOJ, M. & ŽGUR-BERTOK, D. 2013. *Escherichia coli* uropathogenic-specific protein, Usp, is a bacteriocin-like genotoxin. *The Journal of infectious diseases,* 208**,** 1545-1552.

OUTI N, JANI HALKILAHTI J, WIKLUND G, OKEKE U, PAULIN L, AUVINEN P, HAUKKA K, & SIITONEN A. 2015. Comparative genomics and characterization of hybrid Shigatoxigenic and enterotoxigenic Escherichia coli (STEC/ETEC) strains. *PLoS One* 10, 8

OCHMAN, H., LAWRENCE, J. G. & GROISMAN, E. A. 2000. Lateral gene transfer and the nature of bacterial innovation. *nature,* 405**,** 299.

OCHMAN, H. & SELANDER, R. K. 1984. Standard reference strains of *Escherichia coli* from natural populations. *Journal of bacteriology,* 157**,** 690-693.

OOKA, T., OGURA, Y., KATSURA, K., SETO, K., KOBAYASHI, H., KAWANO, K., TOKUOKA, E., FURUKAWA, M., HARADA, S. & YOSHINO, S. 2015. Defining the genome features of *Escherichia albertii*, an emerging enteropathogen closely related to *Escherichia coli*. *Genome biology and evolution,* 7**,** 3170-3179.

ORIOL, R., MOLLICONE, R., CAILLEAU, A., BALANZINO, L. & BRETON, C. 1999. Divergent evolution of fucosyltransferase genes from vertebrates, invertebrates, and bacteria. *Glycobiology,* 9**,** 323-334.

ORTH, J. D., CONRAD, T. M., NA, J., LERMAN, J. A., NAM, H., FEIST, A. M. & PALSSON, B. Ø. 2011. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism—2011. *Molecular systems biology,* 7**,** 535.

PACHEPSKY, Y. & SHELTON, D. 2011. *Escherichia coli* and fecal coliforms in freshwater and estuarine sediments. *Critical reviews in environmental science and technology,* 41**,** 1067-1110.

PAGE, A. J., CUMMINS, C. A., HUNT, M., WONG, V. K., REUTER, S., HOLDEN, M. T., FOOKES, M., FALUSH, D., KEANE, J. A. & PARKHILL, J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics,* 31**,** 3691-3693.

PALLEN, M. J., BAILEY, C. M. & BEATSON, S. A. 2006. Evolutionary links between FliH/YscL-like proteins from bacterial type III secretion systems and second-stalk components of the FoF1 and vacuolar ATPases. *Protein science,* 15**,** 935-941.

PALLEN, M. J., BAILEY, C. M. & BEATSON, S. A. 2006c. Evolutionary links between FliH/YscL-like proteins from bacterial type III secretion systems and second-stalk components of the FoF1 and vacuolar ATPases. *Protein science,* 15**,** 935-941.

PALLEN, M. J., CHAUDHURI, R. R. & HENDERSON, I. R. 2003. Genomic analysis of secretion systems. *Current opinion in microbiology,* 6**,** 519-527.

PALMER, L. D. & SKAAR, E. P. 2016. Transition metals and virulence in bacteria. *Annual review of genetics,* 50**,** 67-91.

PARKHILL, J., WREN, B., THOMSON, N., TITBALL, R., HOLDEN, M., PRENTICE, M., SEBAIHIA, M., JAMES, K., CHURCHER, C. & MUNGALL, K. 2001. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature,* 413**,** 523.

PERNA, N. T., PLUNKETT III, G., BURLAND, V., MAU, B., GLASNER, J. D., ROSE, D. J., MAYHEW, G. F., EVANS, P. S., GREGOR, J. & KIRKPATRICK, H. A. 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157: H7. *Nature,* 409**,** 529.

PETERSEN, T. N., BRUNAK, S., VON HEIJNE, G. & NIELSEN, H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature methods,* 8**,** 785.

PHAN, M.-D., PETERS, K. M., SARKAR, S., LUKOWSKI, S. W., ALLSOPP, L. P., MORIEL, D. G., ACHARD, M. E., TOTSIKA, M., MARSHALL, V. M. & UPTON, M. 2013. The serum resistome of a globally disseminated multidrug resistant uropathogenic *Escherichia coli* clone. *PLoS genetics,* 9**,** e1003834.

PICARD, B., GARCIA, J. S., GOURIOU, S., DURIEZ, P., BRAHIMI, N., BINGEN, E., ELION, J. & DENAMUR, E. 1999. The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. *Infection and immunity,* 67**,** 546-553.

POLJAKOVIC, M. & PERSSON, K. 2003. Urinary tract infection in iNOS-deficient mice with focus on bacterial sensitivity to nitric oxide. *American Journal of Physiology-Renal Physiology,* 284**,** F22-F31.

PORCHERON, G., GARÉNAUX, A., PROULX, J., SABRI, M. & DOZOIS, C. M. 2013. Iron, copper, zinc, and manganese transport and regulation in pathogenic *Enterobacteria*: correlations between strains, site of infection and the relative importance of the different metal transport systems for virulence. *Frontiers in cellular and infection microbiology,* 3**,** 90.

PRAGER, R., BAUERFEIND, R., TIETZE, E., BEHREND, J., FRUTH, A. & TSCHÄPE, H. 2004. Prevalence and deletion types of the pathogenicity island ETT2 among *Escherichia coli* strains from oedema disease and colibacillosis in pigs. *Veterinary microbiology,* 99**,** 287-294.

PRATT, L. A. & KOLTER, R. 1998. Genetic analysis of *Escherichia coli* biofilm formation: roles of flagella, motility, chemotaxis and type I pili. *Molecular microbiology,* 30**,** 285-293.

QI, W., LACHER, D. W., BUMBAUGH, A. C., HYMA, K. E., OUELLETTE, L. M., LARGE, T. M., TARR, C. L. & WHITTAM, T. S. EcMLST: an online database for multi locus sequence typing of pathogenic *Escherichia coli.* null, 2004. IEEE, 520-521.

QUAINOO, S., COOLEN, J. P., VAN HIJUM, S. A., HUYNEN, M. A., MELCHERS, W. J., VAN SCHAIK, W. & WERTHEIM, H. F. 2017. Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis. *Clinical microbiology reviews,* 30**,** 1015-1063.

RAESIDE, C., GAFFÉ, J., DEATHERAGE, D. E., TENAILLON, O., BRISKA, A. M., PTASHKIN, R. N., CRUVEILLER, S., MÉDIGUE, C., LENSKI, R. E. & BARRICK, J. E. 2014. Large chromosomal rearrangements during a long-term evolution experiment with *Escherichia coli*. *MBio,* 5**,** e01377-14.

RAKSHA, R., SRINIVASA, H. & MACADEN, R. 2003. Occurrence and characterisation of uropathogenic *Escherichia coli* in urinary tract infections. *Indian journal of medical microbiology,* 21**,** 102.

RASKO, D. A., ROSOVITZ, M., MYERS, G. S., MONGODIN, E. F., FRICKE, W. F., GAJER, P., CRABTREE, J., SEBAIHIA, M., THOMSON, N. R. & CHAUDHURI, R. 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *Journal of bacteriology,* 190**,** 6881-6893.

REIGSTAD, C. S., HULTGREN, S. J. & GORDON, J. I. 2007. Functional genomic studies of uropathogenic *Escherichia coli* and host urothelial cells when intracellular bacterial communities are assembled. *Journal of Biological Chemistry,* 282**,** 21259-21267.

REN, C.-P., CHAUDHURI, R. R., FIVIAN, A., BAILEY, C. M., ANTONIO, M., BARNES, W. M. & PALLEN, M. J. 2004. The ETT2 gene cluster, encoding a second type III secretion system from *Escherichia coli,* is present in the majority of strains but has undergone widespread mutational attrition. *Journal of bacteriology,* 186**,** 3547-3560.

REPPIN, F., COCHET, S., EL NEMER, W., FRITZ, G. & SCHMIDT, G. 2017. High affinity binding of *Escherichia coli* Cytotoxic Necrotizing Factor 1 (CNF1) to Lu/BCAM adhesion glycoprotein. *Toxins,* 10**,** 3.

RHOADS, A., & AU, K. F. (2015). PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5), 278-289.

RICHARDS, V. P., LEFÉBURE, T., BITAR, P. D. P., DOGAN, B., SIMPSON, K. W., SCHUKKEN, Y. H. & STANHOPE, M. J. 2015. Genome based phylogeny and comparative genomic analysis of intra-mammary pathogenic *Escherichia coli*. *PLoS One,* 10**,** e0119799.

RONALD, A. 2002. The etiology of urinary tract infection: traditional and emerging pathogens. *The American journal of medicine,* 113**,** 14-19.

RUBIN, B. E., WETMORE, K. M., PRICE, M. N., DIAMOND, S., SHULTZABERGER, R. K., LOWE, L. C., CURTIN, G., ARKIN, A. P., DEUTSCHBAUER, A. & GOLDEN, S. S. 2015. The essential gene set of a photosynthetic organism. *Proceedings of the National Academy of Sciences,* 112**,** E6634-E6643.

RUNCHAROEN, C., MORADIGARAVAND, D., BLANE, B., PAKSANONT, S., THAMMACHOTE, J., ANUN, S., PARKHILL, J., CHANTRATITA, N. & PEACOCK, S. J. 2017. Whole genome sequencing reveals high-resolution epidemiological links between clinical and environmental *Klebsiella pneumoniae*. *Genome medicine,* 9**,** 6.

RUSSO, T. A., STAPLETON, A., WENDEROTH, S., HOOTON, T. M. & STAMM, W. E. 1995. Chromosomal restriction fragment length polymorphism analysis of *Escherichia coli* strains causing recurrent urinary tract infections in young women. *Journal of Infectious Diseases,* 172**,** 440-445.

SAITOU, N. & NEI, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution,* 4**,** 406-425.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences,* 74**,** 5463-5467.

SANNER, M. F. 1999. Python: a programming language for software integration and development. *J Mol Graph Model,* 17**,** 57-61.

SAVAGEAU, M. A. 1983. *Escherichia coli* habitats, cell types, and molecular mechanisms of gene control. *The american naturalist,* 122**,** 732-744.

SCHEMBRI, M. A. & KLEMM, P. 2001. Coordinate gene regulation by fimbriae-induced signal transduction. *The EMBO journal,* 20**,** 3074-3081.

SCHICKLMAIER, P. & SCHMIEGER, H. 1995. Frequency of generalized transducing phages in natural isolates of the *Salmonella typhimurium* complex. *Applied and environmental microbiology,* 61**,** 1637-1640.

SCHILLING, J. D., MULVEY, M. A., VINCENT, C. D., LORENZ, R. G. & HULTGREN, S. J. 2001. Bacterial invasion augments epithelial cytokine responses to *Escherichia coli* through a lipopolysaccharide-dependent mechanism. *The Journal of Immunology,* 166**,** 1148-1155.

SEEMANN, T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics,* 30**,** 2068-2069.

SELANDER, R. K., CAUGANT, D. A., OCHMAN, H., MUSSER, J. M., GILMOUR, M. N., & WHITTAM, T. S. 1986. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Applied and environmental microbiology*, 51, 873.

SERVIN, A. L. 2005. Pathogenesis of Afa/Dr diffusely adhering *Escherichia coli*. *Clinical microbiology reviews,* 18**,** 264-292.

SHEIKH, J., DUDLEY, E. G., SUI, B., TAMBOURA, B., SULEMAN, A. & NATARO, J. P. 2006. EilA, a HilA-like regulator in enteroaggregative *Escherichia coli*. *Molecular microbiology,* 61**,** 338-350.

SIGUIER, P., FILÉE, J. & CHANDLER, M. 2006. Insertion sequences in prokaryotic genomes. *Current opinion in microbiology,* 9**,** 526-531.

SLAVCHEV, G., PISAREVA, E. & MARKOVA, N. 2009. Virulence of uropathogenic *Escherichia coli*. *Journal of culture collections,* 6**,** 3-9.

SMATI, M., CLERMONT, O., BLEIBTREU, A., FOURREAU, F., DAVID, A., DAUBIÉ, A. S., HIGNARD, C., LOISON, O., PICARD, B. & DENAMUR, E. 2015. Quantitative analysis of commensal *Escherichia coli* populations reveals host-specific enterotypes at the intra-species level. *Microbiologyopen,* 4**,** 604-615.

SMITH, Y. C., GRANDE, K. K., RASMUSSEN, S. B. & O'BRIEN, A. D. 2006. Novel three-dimensional organoid model for evaluation of the interaction of uropathogenic *Escherichia coli* with terminally differentiated human urothelial cells. *Infection and immunity,* 74**,** 750-757.

SMITH, Y. C., RASMUSSEN, S. B., GRANDE, K. K., CONRAN, R. M. & O'BRIEN, A. D. 2008. Hemolysin of uropathogenic *Escherichia coli* evokes extensive shedding of the uroepithelium and hemorrhage in bladder tissue within the first 24 hours after intraurethral inoculation of mice. *Infection and immunity,* 76**,** 2978-2990.

SRINIVASAN, U., FOXMAN, B. & MARRS, C. F. 2003. Identification of a gene encoding heat-resistant agglutinin in *Escherichia coli* as a putative virulence factor in urinary tract infection. *Journal of clinical microbiology,* 41**,** 285-289.

STAJICH, J. E., BLOCK, D., BOULEZ, K., BRENNER, S. E., CHERVITZ, S. A., DAGDIGIAN, C., FUELLEN, G., GILBERT, J. G., KORF, I. & LAPP, H. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome research,* 12**,** 1611-1618.

STAMATAKIS, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* 30**,** 1312-1313.

STAMM, W. E. & NORRBY, S. R. 2001. Urinary tract infections: disease panorama and challenges. *The Journal of infectious diseases,* 183**,** S1-S4.

SUBASHCHANDRABOSE, S. & MOBLEY, H. L. T. 2015. Virulence and fitness determinants of uropathogenic *Escherichia coli*. *Microbiology spectrum,* 3.

SUBASHCHANDRABOSE, S., SMITH, S. N., SPURBECK, R. R., KOLE, M. M. & MOBLEY, H. L. 2013. Genome-wide detection of fitness genes in uropathogenic *Escherichia coli* during systemic infection. *PLoS pathogens,* 9**,** e1003788.

SUSSMAN, M. 1997. *Escherichia coli: mechanisms of virulence,* Cambridge University Press.

SÄVE, S. & PERSSON, K. 2010. Extracellular ATP and P2Y receptor activation induce a proinflammatory host response in the human urinary tract. *Infection and immunity,* 78**,** 3609-3615.

TANGE, O. 2011. Gnu parallel-the command-line power tool. *The USENIX Magazine,* 36 42-47.

TARTOF, S. Y., SOLBERG, O. D., MANGES, A. R. & RILEY, L. W. 2005. Analysis of a uropathogenic *Escherichia coli* clonal group by multilocus sequence typing. *Journal of clinical microbiology,* 43**,** 5860-5864.

TATUSOV, R. L., KOONIN, E. V., LIPMAN, D. J. 1997. A genomic perspective on protein families. *Science,* 278, 631-7.

TEICHMANN, S. A. & BABU, M. M. 2004. Gene regulatory network growth by duplication. *Nature genetics,* 36**,** 492.

TENAILLON, O., SKURNIK, D., PICARD, B. & DENAMUR, E. 2010. The population genetics of commensal *Escherichia coli*. *Nature Reviews Microbiology,* 8**,** 207.

TENOVER FC, ARBEIT R, ARCHER G, BIDDLE J, BYRNE S, GOERING R, HANCOCK G, HÉBERT GA, HILL B, HOLLIS R. 1994. Comparison of traditional and molecular methods of typing isolates of Staphylococcus aureus. Journal of clinical microbiology. 32,407-15.

TERASHIMA, H., KOJIMA, S. & HOMMA, M. 2008. Flagellar motility in bacteria: structure and function of flagellar motor. *International review of cell and molecular biology,* 270**,** 39-85.

TETTELIN, H., MASIGNANI, V., CIESLEWICZ, M. J., DONATI, C., MEDINI, D., WARD, N. L., ANGIUOLI, S. V., CRABTREE, J., JONES, A. L. & DURKIN, A. S. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences,* 102**,** 13950-13955.

THOMAS, C. M. & NIELSEN, K. M. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews microbiology,* 3**,** 711.

THOMPSON, J., HIGGINS, D. & GIBSON, T. 1994. CLUSTALW: improving the sensitivity of progressive weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res,* 22**,** 4673-4680.

THOMSON, N., BAKER, S., PICKARD, D., FOOKES, M., ANJUM, M., HAMLIN, N., WAIN, J., HOUSE, D., BHUTTA, Z. & CHAN, K. 2004. The role of prophage-like elements in the diversity of *Salmonella enterica* serovars. *Journal of molecular biology,* 339**,** 279-300.

THORPE, H. A., BAYLISS, S. C., SHEPPARD, S. K. & FEIL, E. J. 2018. Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. *GigaScience,* 7**,** giy015.

TOUCHON, M., HOEDE, C., TENAILLON, O., BARBE, V., BAERISWYL, S., BIDET, P., BINGEN, E., BONACORSI, S., BOUCHIER, C. & BOUVET, O. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS genetics,* 5**,** e1000344.

TOVAL, F., KÖHLER, C.-D., VOGEL, U., WAGENLEHNER, F., MELLMANN, A., FRUTH, A., SCHMIDT, M. A., KARCH, H., BIELASZEWSKA, M. & DOBRINDT, U. 2014. Characterization of *Escherichia coli* isolates from hospital inpatients or outpatients with urinary tract infection. *Journal of clinical microbiology,* 52**,** 407-418.

TURRIENTES, M.-C., GONZÁLEZ-ALBA, J.-M., DEL CAMPO, R., BAQUERO, M.-R., CANTÓN, R., BAQUERO, F. & GALÁN, J. C. 2014. Recombination blurs phylogenetic groups routine assignment in *Escherichia coli*: setting the record straight. *PloS one,* 9**,** e105395.

UHLÉN, P., LAESTADIUS, Å., JAHNUKAINEN, T., SÖDERBLOM, T., BÄCKHED, F., CELSI, G., BRISMAR, H., NORMARK, S., APERIA, A. & RICHTER-DAHLFORS, A. 2000. α-Haemolysin of uropathogenic *E. coli* induces Ca 2+ oscillations in renal epithelial cells. *Nature,* 405**,** 694.

VAN OPIJNEN, T., BODI, K. L. & CAMILLI, A. 2009. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nature methods,* 6**,** 767.

VELASCO, E., WANG, S., SANET, M., FERNÁNDEZ-VÁZQUEZ, J., JOVÉ, D., GLARÍA, E., VALLEDOR, A. F., O'HALLORAN, T. V. & BALSALOBRE, C. 2018. A new role for Zinc limitation in bacterial pathogenicity: modulation of α-hemolysin from uropathogenic *Escherichia coli*. *Scientific reports,* 8.

VERNIKOS, G., MEDINI, D., RILEY, D. R. & TETTELIN, H. 2015. Ten years of pan-genome analyses. *Current opinion in microbiology,* 23**,** 148-154.

VISCIDI, R. P. & DEMMA, J. C. 2003. Genetic diversity of Neisseria gonorrhoeae housekeeping genes. *Journal of clinical microbiology,* 41**,** 197-204.

WALK, S. T., ALM, E. W., CALHOUN, L. M., MLADONICKY, J. M. & WHITTAM, T. S. 2007. Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. *Environmental microbiology,* 9**,** 2274-2288.

WALK, S. T., ALM, E. W., GORDON, D. M., RAM, J. L., TORANZOS, G. A., TIEDJE, J. M. & WHITTAM, T. S. 2009. Cryptic lineages of the genus *Escherichia*. *Applied and environmental microbiology,* 75**,** 6534-6544.

WALL, L. 1994. The Perl programming language. Prentice Hall Software Series.

WANG, L., BROCK, A., HERBERICH, B., & SCHULTZ, P. G. 2001. Expanding the genetic code of Escherichia coli. *Science*, *292*(5516), 498-500.

WANG, S., LIU, X., XU, X., ZHAO, Y., YANG, D., HAN, X., TIAN, M., DING, C., PENG, D. & YU, S. 2016. *Escherichia coli* type III secretion system 2 (ETT2) is widely distributed in avian pathogenic *Escherichia coli* isolates from Eastern China. *Epidemiology & Infection,* 144**,** 2824-2830.

WANG, S., XU, X., LIU, X., WANG, D., LIANG, H., WU, X., TIAN, M., DING, C., WANG, G. & YU, S. 2017. *Escherichia coli* type III secretion system 2 regulator EtrA promotes virulence of avian pathogenic *Escherichia coli*. *Microbiology,* 163**,** 1515-1524.

WARREN, J. W. 2001. Catheter-associated urinary tract infections. *International journal of antimicrobial agents,* 17**,** 299-303.

WATTS, R. E., TOTSIKA, M., CHALLINOR, V. L., MABBETT, A. N., ULETT, G. C., DE VOSS, J. J. & SCHEMBRI, M. A. 2012. Contribution of siderophore systems to growth and urinary tract colonization of asymptomatic bacteriuria *Escherichia coli*. *Infection and immunity,* 80**,** 333-344.

WELCH, R. A., BURLAND, V., PLUNKETT, G., REDFORD, P., ROESCH, P., RASKO, D., BUCKLES, E., LIOU, S.-R., BOUTIN, A. & HACKETT, J. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proceedings of the National Academy of Sciences,* 99**,** 17020-17024.

WHELAN, S. & GOLDMAN, N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution,* 18**,** 691-699.

WHITTAM, T. S., OCHMAN, H. & SELANDER, R. K. 1983. Multilocus genetic structure in natural populations of *Escherichia coli*. *Proceedings of the National Academy of Sciences,* 80**,** 1751-1755.

WIELER, L. H., MCDANIEL, T. K., WHITTAM, T. S. & KAPER, J. B. 1997. Insertion site of the locus of enterocyte effacement in enteropathogenic and enterohemorrhagic *Escherichia coli* differs in relation to the clonal phylogeny of the strains. *FEMS microbiology letters,* 156**,** 49-53.

WILES, T. J., KULESUS, R. R. & MULVEY, M. A. 2008. Origins and virulence mechanisms of uropathogenic *Escherichia coli*. *Experimental and molecular pathology,* 85**,** 11-19.

WILES, T. J., NORTON, J. P., RUSSELL, C. W., DALLEY, B. K., FISHER, K. F. & MULVEY, M. A. 2013. Combining quantitative genetic footprinting and trait enrichment analysis to identify fitness determinants of a bacterial pathogen. *PLoS genetics,* 9**,** e1003716.

WIRTH, T., FALUSH, D., LAN, R., COLLES, F., MENSA, P., WIELER, L. H., KARCH, H., REEVES, P. R., MAIDEN, M. C. & OCHMAN, H. 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Molecular microbiology,* 60**,** 1136-1151.

WOESE, C. R. 2000. Interpreting the universal phylogenetic tree. *Proceedings of the National Academy of Sciences,* 97**,** 8392-8396.

WOOLLEY, A. T., & MATHIES, R. A. 1995. Ultra-high-speed DNA sequencing using capillary electrophoresis chips. Analytical chemistry, 67, 3676-3680.

WRIGHT, K. J., SEED, P. C. & HULTGREN, S. J. 2005. Uropathogenic *Escherichia coli* flagella aid in efficient urinary tract colonization. *Infection and immunity,* 73**,** 7657-7668.

YAMAMOTO, S. 2007. Molecular epidemiology of uropathogenic *Escherichia coli*. *Journal of Infection and Chemotherapy,* 13**,** 68-73.

YAMANAKA, K., FANG, L. & INOUYE, M. 1998. The CspA family in *Escherichia coli*: multiple gene duplication for stress adaptation. *Molecular microbiology,* 27**,** 247-255.

YANG, Z. & RANNALA, B. 2012. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics,* 13**,** 303.

YOUNG, G. M., SCHMIEL, D. H. & MILLER, V. L. 1999. A new pathway for the secretion of virulence factors by bacteria: the flagellar export apparatus functions as a protein-secretion system. *Proceedings of the National Academy of Sciences,* 96**,** 6456-6461.

ZAHRA, R., JAVEED, S., MALALA, B., BABENKO, D. & TOLEMAN, M. A. 2018. Analysis of *Escherichia coli* STs and resistance mechanisms in sewage from Islamabad, Pakistan indicates a difference in *E. coli* carriage types between South Asia and Europe. *Journal of Antimicrobial Chemotherapy,* 73**,** 1781-1785.

ZHANG, L., PALLEN, M., FRANKEL, G., SHAW, R., KNUTTON, S. & STEVENS, M. 2004. Regulators encoded in the ETT2 gene cluster influence expression of genes within the locus of enetrocyre effacement in enterohaemorrhagic *Escherichia coli* O157: H7. *Infection and Immunity,* 72**,** 7282-7293.

ZHANG-AKIYAMA, Q.-M., MORINAGA, H., KIKUCHI, M., YONEKURA, S.-I., SUGIYAMA, H., YAMAMOTO, K. & YONEI, S. 2009. KsgA, a 16S rRNA adenine methyltransferase, has a novel DNA glycosylase/AP lyase activity to prevent mutations in *Escherichia coli*. *Nucleic acids research,* 37**,** 2116-2125.