Australian
National
University

A Genetic Analysis of *Escherichia coli* Using Bioinformatic Methods

Phataraporn Khumphai

May 2014

A thesis submitted for the degree of Doctor of Philosophy of The Australian National University

ANU College of Medicine, Biology & Environment

Declaration

I hereby certify that results presented in this thesis are my original work except where reference has been provided. Chapter 4 had a co-author but, in all cases, I am the primary contributor to the work. None of the work in this thesis has been submitted for the award of any other degree.

Phataraporn Khumphai

Phataraporn Khumphai

May 2014

Acknowledgments

First of all, I would like to express my sincere gratitude and deep appreciation to my wonderful supervisor, David Gordon, for his excellent and valuable advice, continuous guidance, constructive criticisms, understanding and especially kindness to me throughout the study which was enable me to carry out my work successfully. I am also very grateful to the member of my supervisory panel, Rod Peakall and John Truman for helpful suggestions for my thesis

I am indebted to Division of Evolution, Ecology & Genetics, Research School of Biology for all facilities and to Ministry of Science and Technology, Thailand for financial supports. My sincere thanks are extended to all members in Gordon lab for their helps with sincerity, kindness, encouragement and friendship.

Finally, I wish to express my deepest appreciation to my parents, Na'Na, P'Tong, all lovely friends, and Knt for their love, understanding and encouragement throughout my study.

Phataraporn Khumphai

May, 2014

ABSTRACT

Escherichia coli is the best-known species of Enterobacteriaceae. The species is genetically diverse and includes both commensal pathogenic strains and plays a significant role in veterinarian, environmental, and medical science. Despite the significance of *E. coli*, many aspects of the species biology, such as its genetic diversity and the pathogenesis are yet to be truly understood. To understand the diversity of *E. coli* as a whole, genome data of *E. coli* derived from variety of sources: humans, animals, and environment are essential. The aims of the research are to take the outstanding opportunity provided by the availability of many new *E. coli* genomes and to make use of a variety of bioinformatics tools to investigate the genetic diversity and reconstruct the evolutionary history of *E. coli* based on a genetic analysis and a comparative genomic approach. The thesis includes three main themes. Firstly, "Distribution of extra-intestinal virulence traits among *E. coli* isolated from native Australian vertebrates with those isolated from humans living in Australia". The frequency and distribution of extra-intestinal virulence traits in a collection of *E. coli* isolated from native Australian vertebrates as compared to *E. coli* isolated from humans living in Australia were reported. The result shows that the frequency and distribution of some traits varies with the source of isolation, human *versus* animal, and that there are traits typically associated with pathogenicity islands that are absent or very rare in animal isolates. The detected high rates of recombination in phylo-group B2 strains suggest that this is an important evolutionary adaptation for attaining virulence. Secondly, "Investigation of the evolution of conjugative plasmids in *E. coli* and their changing role in *E. coli* ecology". Conjugative plasmids: key agents in the adaptation of *E. coli* populations were investigated. Comparing between IncI1 and IncF plasmids, IncI1 plasmids were found to be more homogeneous and genetically conserved than IncF plasmids. These plasmids have changed their role as mediators of intra- and interspecies interactions to become associated with *E. coli* virulence. Lastly, "Genetic and metabolic characteristics of phylo-group F". In this study, phylo-group F: a recently described group of *E. coli* strains was investigated. Strains belonging to phylo-group F were found to be closely related to phylo-group D strains known to be responsible for extra-intestinal infection. Whilst a high degree of strain-specific genome differences were identified among F strains, some of genes shared by F strains (absent in D, B2, and H299) were also present in other phylo-groups (A, B1, and E). All together the outcomes of this project lead to significant advances in our understanding presented in *E. coli* species.

TABLE OF CONTENTS

TABLE OF CONTENTS (cont.)

FIGURES AND TABLES

FIGURES AND TABLES (cont.)

Page

CHAPTER 4 (cont.)

FIGURES AND TABLES (cont.)

Page

CHAPTER 5: Genetic and metabolic characteristics of phylo-group F strains

FIGURES AND TABLES (cont.)

Page

CHAPTER 5 (cont.)

CHAPTER 1

General Introduction and Research Significance

1.1 General Introduction

Escherichia coli is the best-known species of Enterobacteriaceae and represents one of the most important model organisms, especially *E. coli* K-12. *E. coli* is the most numerous facultative anaerobe presenting in the lower intestinal tract of birds and mammals. The species is very genetically diverse and includes both commensal strains with little ability to cause disease and pathogenic strains, which are able to cause intestinal or extra-intestinal infections. In addition to its significance as a pathogen, *E. coli* also plays a significant role in veterinarian, environmental, and medical science.

To understand the diversity of *E. coli* as a whole, the investigation of genome data of *E. coli* derived from variety of sources: humans, animals, and environment is essential. Thanks to high throughput sequencing technology, a number of complete and draft *E. coli* genomes are available at the GenBank database. However, most of the *E. coli* strains in the database are pathogenic strains isolated from humans. This limited information may lead to a biased assessment of the diversity to be found in *E. coli*. However, there is a growing body of genome sequence data for *E. coli*, for example, the *Escherichia* genome sequencing project recently undertaken by the Broad Institute of MIT and Harvard. Consequently, there is a more genome database representing the diversity of *E. coli* derived from the variety of sources. The availability of these genome sequence data will enable researchers to address a range of questions concerning studies of *E. coli*.

1.2 Research Significance

Strains of *E. coli* species play a significant role in several aspects: veterinarian, environmental, and medical science. Focusing on its ability to cause a variety of diseases, *E. coli* is considered to be a major cause of human morbidity and mortality around the world. Despite the significance of *E. coli*, however, many aspects of the species, such as its genetic diversity and the pathogenesis are yet to be truly understood.

Therefore the aims of the research are to take advantage of the outstanding opportunity provided by the availability of many *E. coli* genomes and to make use of a variety of bioinformatics tools to investigate the genetic diversity and reconstruct the evolutionary history of *E. coli.* To understand the diversity within the species, strains from a variety of sources: human, animal, and environment were investigated using a comparative genomic approach. The outcomes of this project will lead to significant advances in our understanding presented in *E. coli* species. Consequently, the thesis includes three main themes with a specific objective for each chapter as follows:

Chapter 3: Distribution of Extra-intestinal Virulence Traits among *E. coli* Isolated from Native Australian Vertebrates with Those Isolated from Humans Living in Australia

Objective: To examine the frequency and distribution of extra-intestinal virulence traits in a collection of *E. coli* isolated from native Australian vertebrates as compared to *E. coli* isolated from humans living in Australia.

Chapter 4: Investigation of the Evolution of Conjugative Plasmids in *E. coli* and Their Changing Role in *E. coli* Ecology

Objective: To investigate the evolution of two of the most common types of conjugative plasmids to be found in *E. coli*: IncF and IncI1, and their changing role in *E. coli* ecology.

Chapter 5: Genetic and Metabolic Characteristics of Phylo-group F Strains

Objective: To characterize the phylogenetic and metabolic diversity among the *E. coli* strains belonging to phylo-group F: a recently described group of *E. coli* strains.

CHAPTER 2

Literature Reviews

2.1 *Escherichia coli* and the genus *Escherichia*

E. coli is the best-known species of Enterobacteriaceae and represents one of the most important model organisms, especially *E. coli* K-12. *E. coli* is the most numerous facultative anaerobe presenting in the lower intestinal tract of birds and mammals. The species is very genetically diverse. *E. coli* strains include both commensal variants with little ability to cause disease and various pathogenic strains that are able to cause intestinal or extra-intestinal infections. In addition to its significance as a pathogen, *E. coli* also plays a significant role in veterinarian, environmental, and medical science.

Besides *E. coli*, there are other closely related species within the genus *Escherichia* such as *E. fergusonii*, *E. albertii*, and the 5 novel clades of the genus *Escherichia* including Clade I - Clade V (Walk et al., 2009). The *Escherichia* species and novel clades differed in their rates of evolution and *E. fergusonii* has evolved at an accelerated rate under selection conditions (Walk et al., 2009). Among 5 novel clades, Clade I was the most closely related clade to *E. coli,* whilst Clade V was the most distantly related clade. These 5 novel clades can be distinguished from typical *E. coli* using a gene sequence analysis, however they were not discriminated from *E. coli* by a traditional biochemical profiling with the exception for Clade III (Walk et al., 2009).

2.2 Typing methods for *E. coli*

The highly genetic diversity is found among *E.coli* isolates and is what enable the species to exhibit such a variety of life style (Gordon, 2010). Various typing methods have been developed to differentiate among isolates of *E. coli* species.

2.2.1 Phenotypic methods

Traditionally, phenotypic methods: serotyping and biochemical and antibiotic resistance profiling have been used for a typing method of bacteria.

Serotyping: Serotyping involves the determination of strains by the somatic (O), capsular (K), or flagellar (H) antigen present on a strain (Kauffmann, 1947). However the method cannot be used to discriminate the phylogenetic relationships among stains when genetically distinct *E. coli* genotype have been found to have the same serotype (Caugant et al., 1985).

Biochemical and antibiotic resistance profiling: Biochemical and antibiotic resistance profiling are inexpensive and provide a reasonable degree of discrimination among strains, as they are so phenotypically diverse. However, both techniques are very dependent on specific conditions: incubation time for biochemical profiling and the background level of resistance of a strain for antibiotic resistance profiling (Gordon, 2010). In addition, both techniques cannot be used to assess the genetic relatedness of *E. coli* strains.

2.2.2 *Multi-locus enzyme electrophoresis (MLEE)*

The technique analyses the genetic differences between *E. coli* isolates by discovering variants of a range of constitutively expressed enzymes in a species based on their electrophoretic mobility (Selander et al., 1986). However the method is no longer in common use as it is expensive and time consuming as compare with many of new DNA-based typing methods.

2.2.3 *DNA-based methods*

There is an abundance of DNA typing methods including random amplification of polymorphic DNA (RAPD), ribotyping, amplification fragment length polymorphism (AFLP), plused-field gel electrophoresis (PFGE), repetitive extragenic palindromic PCR (rep-PCR), and multi-locus sequence typing (MLST). The methods all have an advantage but also a disadvantage for each technique.

Of these techniques, PFGE is thought to be the sensitive discriminating method. However the technique is extremely labor intensive making it unsuitable for large scale studies (Gordon, 2010). Consequently, many studies currently employ rep-PCR as the method is simple and time efficient, reproducible, and inexpensive. The rep-PCR utilizes oligonucleotide primer complementary to particular repetitive sequences within the *E. coli* genome (Versalovic et al., 1991). However the degree of discrimination of rep-PCR obtained with the choice of primer used (Mohapatra et al., 2007).

The multi-locus sequence typing (MLST) uses the allele profile data of selected housekeeping genes (usually 7 genes), which the nucleotide sequence of a 300-700 bp region of each gene is determined, to assign a strain to a sequence type (ST) (Maiden et al., 1998, Gordon, 2010). The extensive collections of STs found in *E. coli* are available, for example the largest MLST database at the ERI, University College Cork available online at http://mlst.ucc.ie/mlst/dbs/Ecoli. MLST is also the method used to assign an unknown isolate to a phylo-group of *E. coli*. Although MLST is more discriminating as compare to PFGE, the method is expensive and time consuming. In addition, a PCR-based method is shown to be more sensitive and discriminating than MLST (Clermont et al., 2013).

2.2.4 Genome-based methods

The growing body of *E. coli* genome data has provided evidence of the extent of substructure in *E coli*. Based on whole-genome scale analysis, the genomic data are informative and also more reliably reflect the phylogenetic relationships among strains. In addition, the available *E. coli* genomes were used to develop primers of PCR-based methods resulting in the improvement of specificity and detection of a new phylo-group (Clermont et al., 2013).

2.3 The diversity and phylo-group structure of *E. coli*

E. coli has a largely clonality which enable *E. coli* strains to be classified into a number of distinct phylo-group. The existence of extensive substructure within the species has been demonstrated using various typing methods. Based on methods such as multi-locus enzyme electrophoresis (MLEE) (Selander et al., 1986) and multilocus sequence typing (MLST) (Gordon et al., 2008), *E. coli* can be subdivided into 5 main phylo-groups known as A, B1, B2, D, and E. Among these phylo-groups, D is described as a sister group to A and B1, which they are thought to be one clade, while B2 is considered to be the basal group of *E. coli* (Lecointre, 1998). *E. coli* strains belonging to phylo-group E are rare, and are largely enterohaemorrhagic *E. coli* (EHEC) (Gordon et al., 2008). However, due to the diversity of the strains and the growing body of multi-locus sequence data and genome data for *E. coli*, the additional phylo-groups have been recently delineated. Based on the recent method: the extended quadruplex PCR phylo-group assignment, *E. coli* are now assigned to 8 phylo-groups including A, B1, B2, C, D, E, F, and *Escherichia* cryptic clade I (Clermont et al., 2013). A phylo-group C is described as a group of strains closely related to the B1 group (Clermont et al., 2011). While a phylo-group F has been suggested for a sister group to phylo-group B2 (Jaureguy et al., 2008, Clermont et al., 2011).

E.coli strains have phylogenetic cohesiveness; however among strains of the various phylo-groups, they differ in their phenotypic characteristics (Gordon, 2004), genome size (Touchon et al., 2009), and propensity to cause intestinal or extra-intestinal infections. Strains of the different phylo-groups are also associated with certain ecological niches and life-history characteristics (Gordon and Cowling, 2003). The phenotypic differences of the phylo-groups of *E. coli* comprise differing growth rates, antibiotic resistance, and biochemical profiles (Gordon, 2004). *E.coli* strains are phenotypically heterogeneous (Touchon et al., 2009). However, evidence suggested that the environmentally adapted *E. coli* lineages were found to be phenotypically and taxonomically indistinguishable from typical *E. coli* based on traditional phenotypic tests (Luo et al., 2011).

2.3.1 <u>Genome size and genomic diversity of E. coli</u>

Genomes within the *E. coli* species can differ in size by more than 1 Mb. Among strains of the different phylo-groups, strains belonging to phylo-groups A and B1 have smaller genomes than B2 or D strains (Bergthorsson and Ochman, 1998). The investigation of 20 *E. coli* genomes from 5 main phylo-groups (A, B1, B2, D, and E) showed that the average size of an *E. coli* genome was around 5 Mb and represented about 4700 protein-coding genes (Touchon et al., 2009). However, only 46% were common to all genomes investigated. Of about 18,000 total genes for the pan-genome, there were about 11,000 genes with no strong relation of homology and about 10,000 unique genes after removing all transposable elements and prophage (Touchon et al., 2009). Among the 20 genomes investigated, about 62% of these genes were present in at least 18 genomes, whereas 26% were present in 4 or fewer genomes (Touchon et al., 2009). These indicated the enormous genetic diversity present in the species and they are what enable *E. coli* strains to exhibit such a variety of life styles.

2.3.2 *Ecological niches and life-history characteristics of E. coli*

The ecological niches of *E. coli* strains have been found to correlate with phylo-group membership. According to the genetic diversity of the species, strains belonging to phylo-groups B2 and D are less frequently isolated from the environment (Walk et al., 2007) or fish, frogs, and reptiles than A or B1 strains (Gordon and Cowling, 2003). However, B2 and D strains are more frequently recovered from extra-intestinal body sites than A or B1 strains (Gordon, 2004). In addition, B2 strains are rarely isolated from water samples; however, the strains are more frequently isolated from mammal hosts possessing hindgut modifications for microbial fermentation than other strains of *E. coli* (Gordon and Cowling, 2003). Moreover, B2 strains have been shown to persist for longer periods in infants than strains of the other phylo-groups (Nowrouzian et al., 2006).

Generally, *E. coli* depends on the presence of a vertebrate host population to persist as the primary habitat. However, *E. coli* can transit in water, sediment, and soil, where represent the species' secondary habitat (Savageau, 1983). There is evidence that some strains may be essentially free living in the environment independent of warm-blooded hosts (Power et al., 2005). For example, two environmental strains belonging to phylo-group F: *E. coli* SMS-3-5 (Fricke et al., 2008) and *E. coli* E1227 (this study). In addition, the strains belonging to novel five *Escherichia* clades (Clade I to Clade V) were isolated primarily from environmental sources (Luo et al., 2011).

2.3.3 *Propensity to cause diseases of E. coli*

Escherichia coli is genetically diverse and includes both commensal and various pathogenic strains. The ability of a strain to cause disease is due to the presence of a range of traits thought to enhance the ability to cause disease of a strain. Many of the most significant virulence genes: adhesins, extracellular protein secretion systems, and toxins, for several *E. coli* pathotypes are clustered together in pathogenicity-associated islands (PAIs) which are known to normally present in pathogens. Although the majority of traits associated with PAIs are located on the chromosome, many are encoded on plasmids (Hacker and Kaper, 2000) (Table 1).

Table 1. Major virulence factors (VFs) encoded by pathogenic strains

Category/ gene	Location*	Product/ function
Adhesins		
papA	PI (C)	Major structural subunit of pilus associated with pyelonephritis (P fimbriae); defines F antigen the P fimbrial structural subunit gene
sfa /focDE	PI (C)	Central region of *sfa* (S fimbriae) and *foc* (F1C fimbriae) operons
sfaS	PI (C)	Pilus tip adhesin, S fimbriae (sialic acid specific)
focG	PI (C)	Pilus tip molecule, F1C fimbriae (sialic acid specific)
iha	PI (C)	Novel nonhemagglutinin adhesin (from O157:H7 and CFT073)
eaeA	PI (C)	Intimin (mediates attaching/effacing lesions)
Toxins		
hlyA	PI (C)	α-Hemolysin
cnf1	PI (C)	Cytotoxic necrotizing factor 1
astA	PI (C, P)	heat-stable enterotoxin I (enterotoxin)
Protectins		
kpsMT II	PI (C)	Cell surface attribute: Group II capsular polysaccharide synthesis (e.g., K1, K5, and K12)
kpsMT K1	PI (C)	Cell surface attribute: Specific for K1 (group II) *kpsMT*
Siderophores		
iutA	C, P	Iron acquisition: Ferric aerobactin receptor (iron uptake: transport)
fyuA	PI (C)	Iron acquisition: *Yersinia* siderophore receptor (ferric yersiniabactin uptake)
iroN	PI (C, P)	Iron acquisition: Novel catecholate siderophore
ireA	PI (C, P)	Iron acquisition: iron-regulated outer membrane virulence protein (also found in plasmid pE11210p1 of *E. coli* O104:H4 str. E112/10)
Miscellaneous		
cvaC	C, P	Colicin V; conjugative plasmids (*traT, iss,* and antimicrobial resistance)
traT	P	Serum survival: Surface exclusion, serum survival (outer membrane protein)
iss	PI (P)	Serum survival: increased serum survival protein; involved in complement resistance
ompT	C	Cell surface attribute: Outer membrane protein T (protease)
H7	C	Cell surface attribute
ibeA	C	Cell invasion: Invasion of brain endothelium
she	C	Secreted protein
malX	PI (C)	PAI marker from strain CFT073
irp2	PI (C)	non-ribosomal peptide synthase (yersiniabactin siderophore biosynthetic protein)

*Most probable location of trait: PI, pathogenicity island; C, chromosome; and P, plasmid

The pathogenic *E. coli* can be divided into strains causing intestinal diseases and other strains causing extra-intestinal infections (Kaper et al., 2004). These lead to categories of the pathotypes of *E. coli* into intestinal and extra-intestinal pathotypes. The intestinal pathotypes consist of enterotoxigenic *E. coli* (ETEC), enteropathogenic *E. coli* (EPEC), enterohemorrhagic *E. coli* (EHEC), enteroaggregative *E. coli* (EAEC), enteroinvasive *E. coli* (EIEC), and diffusely adherent *E. coli* (DAEC) (Nataro and Kaper, 1998, Nataro, 2005, Kaper et al., 2004). The extra-intestinal pathotype includes extra-intestinal pathogenic *E. coli* (ExPEC) which can be subdivide into uropathogenic *E. coli* (UPEC), neonatal meningitis-associated *E. coli* (NMEC), necrotoxigenic *E. coli* (NTEC), and avian pathogenic *E. coli* (APEC) (Kaper et al., 2004).

The ability of a strain to cause diverse diseases varies among strains of various phylo-groups. As phylo-group B2 and to a lesser extent phylo-group D strains are frequently isolated from extra-intestinal body sites, they are most likely to be responsible for extra-intestinal infections. In contrast, phylo-group A strains are less likely and phlyo-group B1 strains the least likely to cause such infections (Johnson and Russo, 2002, Jaureguy et al., 2008). Strains belonging to phylo-groups A and B1 are typically only known to cause opportunistic infections in compromised hosts (Moreno et al., 2005).

The distribution of extra-intestinal disease associated traits varies among strains of the different phylo-groups (Johnson et al., 2001) and numerous studies have shown that most extra-intestinal virulence genes are concentrated in phylo-group B2 strains and to a lesser extent phylo-group D strains (Johnson et al., 2001, Gordon et al., 2005). Strains belonging to phylo-groups B2 and D express a number of putative virulence traits that are proposed to play roles in the process of causing disease (Bahrani-Mougeot et al., 2002). These traits include fimbriae, capsule, hemolysins, membrane-bound and secreted proteins, lipopolysaccharides, and iron- acquisition systems, which appear more frequently in uropathogenic *E. coli* (UPEC) (Johnson, 1991). Most UPEC strains belonging to phylo-groups B2 and D carry large blocks of genes encoding virulence associated traits, called pathogenicity-associated islands (PAIs), not found in fecal isolates (Bahrani-Mougeot et al., 2002, Johnson and Russo, 2002, Johnson, 1991). For example, *papA* gene encodes a structural subunit of P fimbriae which mediate attachment to intestinal and urinary epithelium, and *iutA* gene encodes an aerobactin receptor which required for an iron acquisition systems. These virulence genes are what enable B2 and D strains to cause extra-intestinal infections beyond fecal *E. coli* such as strains in phylo-groups A and B1 that are most likely to cause intestinal infections (Bahrani-Mougeot et al., 2002, Johnson et al., 2006). However available evidence has revealed that *E. coli* strains belonging to phylo-group B1 can also often cause extra-intestinal infections in birds (Barbieri et al., 2012). These virulence factors are considered to be one of the molecular factors that sustain the diversity of adaptive paths and complexity of *E. coli* niches (Tenaillon et al., 2010).

2.4 Conjugative plasmids in *E. coli*

Plasmids are extrachromosomal replicons that are prevalent symbionts of bacteria. Generally, plasmids are not essential for normal bacterial growth. However they often code for genes involved in antibiotic and heavy metal resistance, virulence, and ecological interactions. That is traits that encode for their host's adaptation to the environment. Plasmids can propagate themselves vertically via cell division and many can propagate themselves horizontally usually by infectious transfer via mobilization or conjugation (Summers, 1996, Bergstrom et al., 2000).

Based on mobility systems of plasmids, they can be categorized into three groups; conjugative (self-transmissable), mobilizable (transmissible), and nonmobilizable plasmids (Smillie et al., 2010). A mechanism of conjugation of conjugative plasmids involves four components of a conjugative machinery needed for self-transfer that includes an origin of transfer (*oriT*), a relaxase gene, a type IV coupling protein (T4CP), and a type IV secretion system (T4SS) (Smillie et al., 2010). All conjugative plasmids possess core components of a plasmid backbone including replication genes, stability genes, and a conjugative transfer (*tra*) region ensuring their self-transmissibility. Conjugation encoded by plasmids involves a donor bacterium carrying a conjugative plasmid and a recipient cell without a conjugative plasmid. Conjugation between donor and recipient cells of some conjugative plasmids can occur between different species, genera, or kingdoms (Amabile-Cuevas and Chicurel, 1992). Thanks to these properties of conjugative plasmids, they are known to be key agents playing an important role in horizontal gene transmission among bacteria.

In *E. coli*, many plasmids types comprising a number of plasmid incompatibility (Inc) groups (or plasmid replicon (Rep) groups which are used interchangeably) are known to occur among *E. coli* strains and they play an important role in the "adaptation" of bacterial populations (Frost et al., 2005). Of these Inc groups, almost conjugative plasmids belonging to IncF group carrying a fertility factor (F factor) transfer region for their self-transmissibility were associated with *E. coli* virulence (Johnson and Nolan, 2009). Although a number of other Inc (Rep) groups have been identified, only a few of them (i.e., IncB/O and IncI1) have been found to be associated with *E. coli* virulence (Johnson and Nolan, 2009).

However these *E. coli* plasmids originally encoded traits thought to mediate competitive interactions among strains, traits known as bacteriocins: key agents mediating a strain's adaptation to local environmental niches (Eberhard, 1990). Bacteriocins are defined as antimicrobial proteins with a narrow killing range, that are toxic only to bacteria closely related to the producing strain (Riley and Wertz, 2002). These substances play a significant role in maintaining microbial biodiversity by acting as important mediators of intra- and interspecies interactions (Kirkup and Riley, 2004, Czaran et al., 2002). In *E. coli*, the production of multiple bacteriocins by a single strain is a common phenomenon (Gordon and O'Brien, 2006). Several bacteriocins have been found to be encoded on the same large conjugative plasmid in *E. coli* strains more often than expected by chance (Gordon and O'Brien, 2006, Gordon et al., 2007).

2.5 Applications of bioinformatics in the genomic research of *E. coli*

Advances in innovative high-throughput biotechnologies, especially in a high throughput sequencing technology, are resulting in the exponential growth of high-dimensional data. Consequently, the huge amounts of biological data have created an enormous challenge to optimize and make effective use of accumulated information. Thanks to modern computational tools in bioinformatics, they have revolutionized biological research by providing powerful approaches for managing biological databases and investigating biological data systematically.

The management of biological databases includes an archive of information, a logical organization or structure of that information, and tools to gain access to it (Lesk, 2005).

The National Center for Biotechnology Information (NCBI) in the United States (http://www.ncbi.nlm.nih.gov/) and the European Bioinformatics Institute (EBI) in England (http://www.ebi.ac.uk/) are two classic life science servers maintaining databases and software tools in the life sciences. The usefulness of these and other centralized databases allow scientific community to make effective use of the databanks as free resources of the structure and function of genes and proteins in biology and medical research.

Furthermore, advances in computational methodologies have generated a variety of bioinformatic tools for investigating biological data ranging from the individual nucleic acid sequence to systems biology of the living cell. These include tools for nucleotide and protein sequence analysis, phylogenetic analysis, genomics, proteomics, transcriptomics, metabolomics, systems biology, etc. Several bioinformatic tools are freely-accessed on-line for scientists to make use of them as a set of scientific tools to address a range of research.

Thanks to a high throughput sequencing technology, a number of *E. coli* genomes are being sequenced to serve as resources of the diversity within *E. coli* and the genus *Escherichia.* The power of the sequencing technology, i.e., the Genome Sequencer FLX System (454 Life Sciences, Roche) has also been applied to generate enormous sequence information for studying the genomic content in a complex mixture of microorganisms so called Metagenomics (Harkins and Jarvie, 2007). To manage *E. coli* genome sequences systematically, many special databases of *E. coli* have been established by several institutes. The database, such as the *Escherichia coli* Antibiotic Resistance Database hosted by the Broad Institute of MIT and Harvard brings together the *E. coli* genome sequences and their annotation with bioinformatic tools providing users with BLAST search and several comparative analysis tools. The other database, ASAP, the database hosted by University of Wisconsin-Madison also brings together the *E. coli* genome annotation and bioinformatic tools with other bacterial strains. In addition to the database for *E. coli* genomes, there are databases for Multi-locus sequence typing (MLST) of *E. coli* (*Ec*MLST, *Escherichia coli* MLST Database, etc.) that allow scientists to make effective use of the databases.

CHAPTER 3

Distribution of extra-intestinal virulence traits among *E. coli* isolated from native Australian vertebrates with those isolated from humans living in Australia

3.1 Introduction

Escherichia coli is the most numerous facultative anaerobe present in the lower gastrointestinal tract of humans and other warm-blooded species. Although normally a commensal, strains of the species can cause a variety of intestinal and extra-intestinal infections. Methods such as multi-locus enzyme electrophoresis (MLEE) (Ochman and Selander, 1984) and multi-locus sequence typing (MLST) (Gordon et al., 2008) have shown that *E. coli* can be subdivided into five main phylo-groups known as A, B1, B2, D and E. Besides these phylo-groups, an additional phylo-group called F has been delineated (Jaureguy et al., 2008). Strains of the various phylo-groups may differ in many aspects, including their phenotypic characteristics (Gordon, 2004), genome size (Touchon et al., 2009), ecological niches, and life-history characteristics (Gordon and Cowling, 2003). Strains of the different phylo-groups also vary in their propensity to cause extra-intestinal infections. Phylo-group B2 and to a lesser extent phylo-group D strains are most likely to be responsible for extra-intestinal infections in humans, whilst phylo-group A are less likely, and phylo-group B1 strains the least likely to cause such infections (Jaureguy *et al.*, 2008; Johnson, 2002). However, available evidence has revealed that *E. coli* strains belonging to phylo-group B1 can also often cause extra-intestinal infections in birds (Barbieri et al., 2012).

According to MLST and several PCR-based assays for the identification of highly virulent clonal group, these methods have suggested that some particular sequence types (STs) and clonal groups may be clinically significance as they are found to be common and responsible for extra-intestinal infection in *E. coli* virulent strains (Bonacorsi et al., 2009, Bidet et al., 2007, Mora et al., 2009, Johnson et al., 2004). For example, many phylo-group B2 strains with the serotypes O1:K1, O2:K1 and O18:K1 are members of the MLST ST95 group of strains (Mora et al., 2009, Clermont et.al., 2011). Much effort has been devoted to identifying those traits that enhance an *E. coli* strain's ability to cause extra-intestinal disease and a great many proven and putative virulence traits have been identified (Johnson and Russo, 2002). It is well known that the distribution of these extra-intestinal disease associated traits varies among strains of the four main phylo-groups (A, B1, B2 and D) (Johnson et al., 2001) and numerous studies have shown that many of these traits are concentrated in phylo-group B2 strains and to a lesser extent phylo-group D strains (Johnson et al., 2001, Gordon et al., 2005). Many of the most significant virulence genes (adhesins, extracellular protein secretion systems, and toxins) for several *E. coli* pathotypes are clustered together in pathogenicity-associated islands (PAIs). For example, the High Pathogenicity Island (HPI) that encodes the virulence genes *fyuA* and *irp2* (Schubert et al., 2002). This and other *E. coli* pathogenicity-associated islands (PAI) are concentrated in phylo-groups B2 and D, and these islands are only occasionally found in phylo-groups A and B1 (Schubert et al., 2009, Clermont et al., 2001, Boyd and Hartl, 1998). Conjugative transfer and homologous DNA recombination are reported to play a major role in horizontal transfer of PAI within *E. coli* (Schubert et al., 2009).

Previous study has demonstrated that phylo-group B2 strains known to be responsible for extra-intestinal infections in humans are more frequently isolated from mammal host possessing hindgut modifications for microbial fermentation than other strains of *E. coli* (Gordon and Cowling, 2003). Phylo-group B2 strains have been also capable of persisting in infants than strains of the other phylo-group (Nowrouzian et al., 2006). There is some controversy as to whether the propensity of phylo-group B2 strains to cause disease is a consequence of the concentration of known and suspected virulence factors among phylo-group B2 strains (Johnson et al., 2006) or an ancestral characteristic of this phylo-group (Le Gall et al., 2007). Johnson *et al.* (2006) showed that it is the presence of specific virulence factors in a strain that are the best predictor of a strain's virulence in a mouse lethality model rather than its phylo-group membership. By contrast, Le Gall *et al.* (2007) argued that the virulence of phylo-group B2 strains in the mouse lethality assay is an ancestral trait, and extra-intestinal virulence is a coincidental by-product of these strains' enhanced ability to persist in the gastro-intestinal tract compared to strains of the other phylo-groups. However the exact answer for this controversy is still unclear. Additional investigations might be required to deal with the fact that B2 strains are competitively dominate in the gut leading to the numberical abundant of the strains. B2 strains also show the ability of persisting for longer period in the gut. Therefore, both these traits of B2 strains (competitively dominate and persisting for longer period in the gut) should be taken into account as these make it likely that it will be B2 strains responsible for the contamination event.

The majority of studies examining the distribution of extra-intestinal virulence traits among strains of the different *E. coli* phylo-groups have been clinical isolates or isolates from humans. Few studies have examined the distribution of the traits in strains isolated from native animals. Here we report on the frequency and distribution of extra-intestinal virulence traits in a collection of *E. coli* isolated from native Australian vertebrates (Gordon & Cowling, 2003) as compared to *E. coli* isolated from humans living in Australia (Gordon *et al.*, 2005). We show that the frequency and distribution of some traits varies with the source of isolation, human *versus* animal, and that there are traits typically associated with pathogenicity-associated islands (PAIs) that are absent or very rare in animal isolates.

3.2 Materials and Methods

3.2.1 Strain collection

The strains used in this study were 266 *E. coli* isolated from faecal samples taken from people living in the Canberra region of Australia and 690 faecal isolates from native non-human vertebrates living in Australia. Further details of the collection and identification of the strains were described in Gordon *et al.* (2005) for the isolates from humans and in Gordon & Cowling (2003) for the isolates from animals. All isolates were previously assigned to one of the four main *E. coli* phylo-groups (A, B1, B2, and D) (Ochman & Selander, 1984; Herzer *et al.*, 1990) using a PCR based technique (Clermont et al., 2000). Sample sizes are as follows: isolates from humans; A (n = 52), B1 (n = 33), B2 (n = 120), D (n = 61): animals; A (n = 93), B1 (n = 272), B2 (n = 213), D (n = 120). Strain are labeled with a letter prefix indicating host group: H, human; M/TA, mammal (non-human); B, bird; R, reptile; and E, environmental.

3.2.2 Virulence gene detection

The 956 *E. coli* isolates (266 human and 690 animal isolates) were previously screened for the presence of 29 virulence genes associated with intestinal and extra-intestinal disease including *malX*, *irp2*, *afa/draBC*, *eaeA*, *fimH*, *focG*, *gafD*, *papAH*, *sfa/focDE*, *ibeA*, H7 *fliC*, *kpsMT*.II, *kpsmMT*.K1, *ompT*, *fyuA*, *iha*, *ireA*, *iroN*, *iutA*, *iss*, *traT*, *astA*, *cnf1*, *hylD*, *stx1*, *stx2*, *she*, *eaag*, and *13fb*. All 29 virulence genes were detected using a PCR method as described by Gordon *et al.* (2005). The information of detected virulence factors (VFs) were used for the investigation of the distribution of VFs in this collection of *E. coli* strains.

3.2.3 MLST analysis of *E. coli* strains belonging to phylo-groups B2 and D

Strain selection for MLST: Due to their clinical significance, a subset of 171 phylo-group B2 and D strains were selected from David Gordon's *E. coli* strain collection for MLST characterization. Included in the study were clinical and faecal human isolates (B2 = 33, D = 19, and F = 1), faecal isolates from native non-human vertebrates (B2 = 65, D = 30, and F = 1), and isolates from soil and water (B2 = 16, D = 6) collected in Australia. The environmental isolates were PCR-screened for the 29 virulence factors described in the previous section. The MLST data subsequently revealed that 2 of the D strains were actually members of phylo-group F.

MLST Sequencing: The 171 *E. coli* isolates (114 phylo-group B2, 55 phylo-group D, and 2 phylo-group F strains) were characterized using the MLST scheme (German) described by Wirth *et al.*, (2006). This scheme examines the 7 housekeeping genes: *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA*. Sequence data for the 7 genes were concatenated (3,422 bp) and analyzed to construct phylogenetic relationships among strains for each phylo-group using MEGA5 program (neighbor joining (NJ) method; p-distance model) (Tamura et al., 2007). Sequence data for the 7 genes were also submitted to the MLST database (http://mlst.ucc.ie/mlst/dbs/Ecoli) in order to obtain a sequence type (ST) for each of the strains.

3.2.4 Analysis of recombination rates in phylo-group B2 strains

The concatenated sequences (3,422 bp) of 114 phylo-group B2 strains were partitioned into 4 input sample sets based on their source (humans, mammals, birds, and environment). The duplicate STs in each input sample set were removed. The number of strains analyzed varied for each sample set as shown in Table 3. These datasets were used for the investigation of putative recombination rates using three methods to compare the consistency of recombination rates derived from each method.

The first method used was the PHI statistic (ϕ_w) (Bruen et al., 2006), as implemented in SplitsTree4 (Huson and Bryant, 2006). The second method, was Clonal Frame, a model-based approach for determining bacterial microevolution (Didelot and Falush, 2007). The last method was, a coalescent-based likelihood permutation test, as implemented in LDhat 2.1 (McVean et al., 2002).

3.2.5 Statistical analyses

Statistical analyses were performed using the package JMP® (SAS Institute). The Fisher's Exact test was used to determine if there are nonrandom associations between the distribution of virulence-associated traits and the phylo-group membership of the strains. Effects were not considered to be significant unless the probability values were less than 0.05.

3.3. Results

3.3.1 Distribution of virulence factors (VFs) among *E. coli* strains belonging to 4 main phylo-groups

The distribution of 29 VFs for human- and animal-source isolates from the 4 main phylo-groups, A, B1, B2, and D was determined. Among the 956 isolates, the virulence-associated trait *13fb* was not detected, whilst *stx1*, *stx2*, *eaag*, and *gafD* were detected 2, 3, 4, and 6 times, respectively (data not shown).

Many of the traits are non-randomly distributed among strains of the 4 phylo-groups (Table 1) and, as this is a well-established observation (Johnson et al., 2001), these results will not be explicitly presented. Of the 29 investigated traits, 24 traits were more frequently detected, and 18 traits (out of 24 traits) were significant for the source of isolation. In addition, 2 traits (*traT* and *ompT*) were significant for the interaction term (Source*Genetic group) (Table 1). Comparing between human-source and animal-source isolates indicated that some traits which were common in the human isolates

Table 1. The distribution of 24 frequently detected traits among strains of the 4 genetic groups

Trait	Source	A	B1	B2	D	Source	Genetic Group	Interaction
Sample Size	**Human**	**52**	**33**	**120**	**61**			
	Animal	**93**	**272**	**213**	**120**			
iutA	Human	26.9	18.2	40.0	29.5	0.0000	0.0002	ns
	Animal	5.4	3.4	9.9	7.1			
fyuA	Human	42.3	21.2	95.8	39.3	0.0000	0.0000	ns
	Animal	11.8	6.3	87.8	17.9			
irp2	Human	42.3	21.2	95.8	39.3	0.0000	0.0000	ns
	Animal	14.0	8.1	85.9	22.3			
iroN	Human	9.6	18.2	50.8	14.7	0.0000	0.0000	ns
	Animal	2.1	6.3	24.9	5.4			
astA	Human	5.8	6.1	1.7	6.6	0.0000	ns	ns
	Animal	30.1	22.8	25.4	24.1			
iss	Human	9.6	15.2	11.7	14.8	0.0004	ns	ns
	Animal	5.4	6.3	11.3	4.5			
*kpsMT*II	Human	21.5	3.0	79.2	57.4	0.0046	0.0000	ns
	Animal	6.4	1.1	44.1	24.1			
fimH	Human	80.8	97.0	99.2	96.7	ns	0.0000	ns
	Animal	88.2	99.3	99.1	96.4			
traT	Human	32.7	39.4	64.2	49.2	ns	0.0083	0.0071
	Animal	35.5	43.4	38.0	35.5			
eaeA	Human	3.8	15.1	3.3	1.6	ns	ns	ns
	Animal	6.5	6.3	6.6	3.6			
ompT	Human	13.5	18.2	88.3	59.0	ns	0.0000	0.0002
	Animal	24.7	35.7	78.4	35.7			
*kpsMT*K1	Human	11.5	3.0	37.5	22.9	0.0000	0.0000	-
	Animal	0	0.4	17.8	3.6			
iha	Human	21.5	9.1	35.8	26.2	0.0000	0.0000	-
	Animal	2.1	0	0.5	0			
hylD	Human	1.9	0	40.8	1.6	0.0000	0.0000	-
	Animal	0	0.4	1.4	0			
papAH	Human	1.9	0	41.7	22.9	0.0000	0.0000	-
	Animal	0	0	1.4	0			
afa/dra	Human	11.5	0	5.0	3.3	0.0000	ns	-
	Animal	0	0	0	0			
H7	Human	0	9.1	24.2	6.6	0.0015	0.0000	-
	Animal	10.7	14.3	8.0	2.7			
ibe	Human	0	0	37.5	8.2	0.0000	0.0000	-
	Animal	4.3	0.7	64.3	8.9			
malX	Human	0	0	93.3	18.0	ns	0.0000	-
	Animal	1.1	1.5	89.2	16.1			
she	Human	0	6.1	36.7	1.6	ns	0.0000	ns
	Animal	6.5	7.4	27.2	0.9			
sfa/foc	Human	0	3.0	42.5	0	0.0000	0.0000	-
	Animal	0	1.5	11.3	0			
focG	Human	0	3.0	20.0	0	0.0000	0.0000	-
	Animal	0	0	6.1	0			
ireA	Human	0	3.0	18.3	8.2	0.0000	0.0000	-
	Animal	0	0.7	3.8	4.5			
cnf1	Human	0	0	35.8	0	0.0000	0.0000	-
	Animal	0	0.4	1.4	0			

ns, not significant; -, not determined

were absent or virtually absent in the animal isolates, and these traits include *iha*, *hylD*, *papAH*, *afa/draBC*, *sfa/foc*, *focG*, and *cnf1* (Table 1). Other traits, *iutA*, *fyuA*, *irp2*, *iroN*, *iss*, *kpsMT*.II, *kpsMT*.K1, and *ireA* were less common in the animal isolates compared to the human isolates. Two traits, *ibe* and *astA* were detected more frequently in animal isolates than in human isolates. Although there was no overall difference in the frequency of *traT* positive strains between isolates from humans and from animals, for the isolates from animals, the frequency of *traT* did not differ among strain of the 4 genetic groups. However, in isolates from humans, *traT* was more prevalent in strains belonging to groups B2 and D as compared to strains from the other groups. The gene *ompT* was more frequently detected in isolates from animals belonging to genetic groups A and B1 compared to A and B1 isolates from humans, whilst *ompT* was more frequently detected in D strains from humans compared to D strains isolated from animals. Overall, the virulence associated traits were more prevalent in human-source and animal-source isolates belonging to phylo-group B2 as compared to phylo-group D strains.

3.3.2 MLST and distribution of virulence traits (VFs) with phylogeny

The results of the MLST analysis revealed that the 114 phylo-group B2 strains represented 83 sequence types (STs), whilst the 55 phylo-group D and 2 phylo-group F strains represented 38 STs (Fig. 1 and 2, respectively). Some strains isolated from humans, animals, and environmental sources were found to represent the same ST, such as ST127 (B2) and ST38 (D). Although other STs, such as ST 95 (B2) and ST69 (D) were predominately observed in humans (Fig. 1 and 2, respectively).

The virulence profiles of the 114 phylo-group B2 strains revealed that all contained at least one of the 24 traits studied (Fig. 1). The human B2 isolate, H411 ST73, contained 16 VF traits, whilst only a single trait was detected in the reptile isolate (R138 ST1909) and environmental isolate (E211 ST1910); *ompT* and *malX*, respectively. Of the 24 VFs examined, the traits *malX*, *fyuA*, *irp2*, and *ompT* showed the substantial prevalence among 114 B2 strains. Some traits (*sfa/foc*, *focG*, *hlyA*, *cnf1*, H47, H7, and *ibe*) presenting in some of phylo-group B2 strains were absent in all 55 phylo-group D and 2 phylo-group F strains (Fig. 2). Of 55 phylo-group D strains, the human isolate (H699 ST69) contained 9 traits, whereas 2 animal isolates including

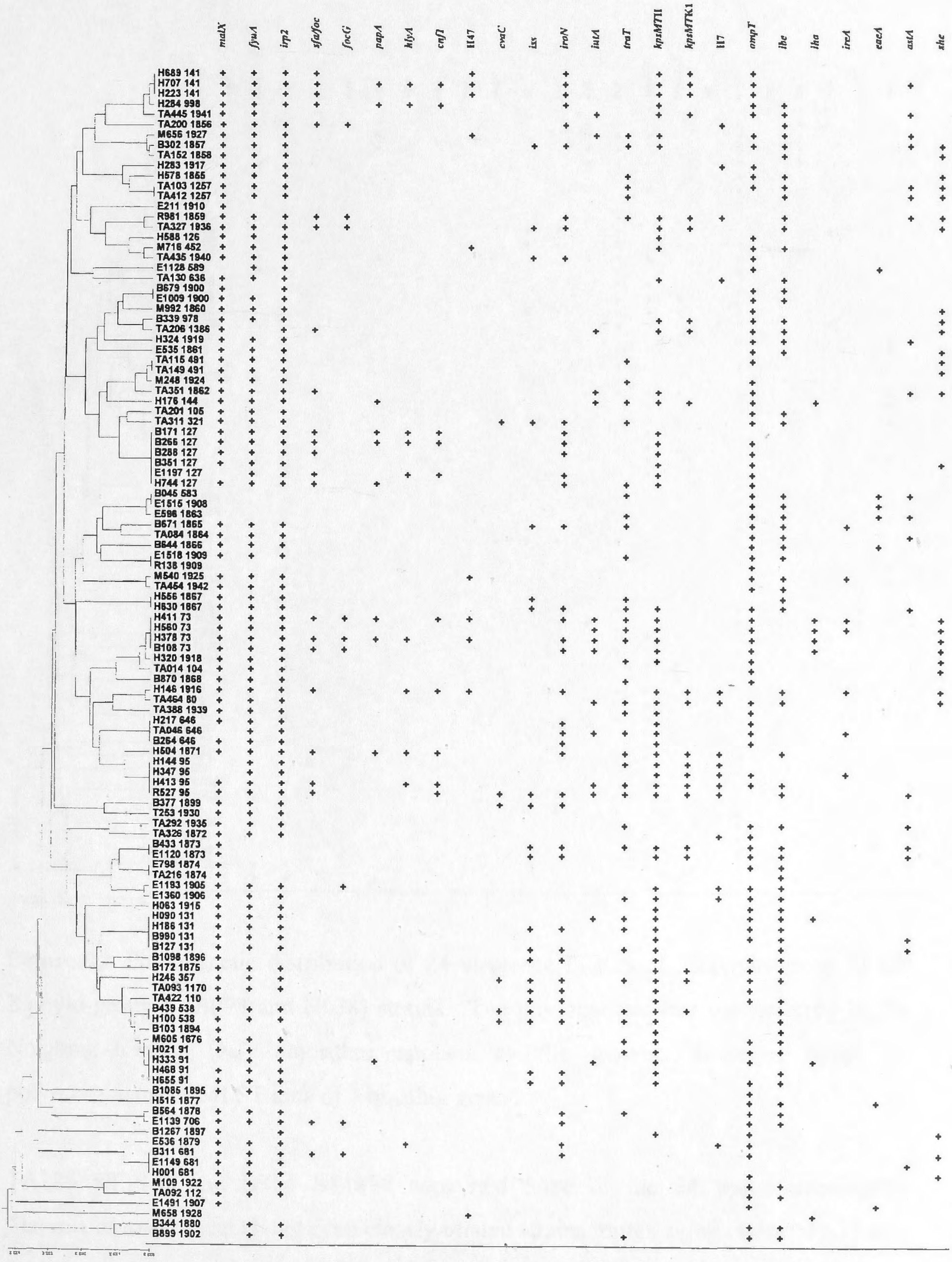


Figure 1. Phylogenetic distribution of 24 virulence factors of 114 phylo-group B2 strains. The phylogenetic tree was inferred by the Neighbour-Joining (NJ) algorithm applied to the genetic distances based on polymorphisms of MLST data of 7 housing genes.

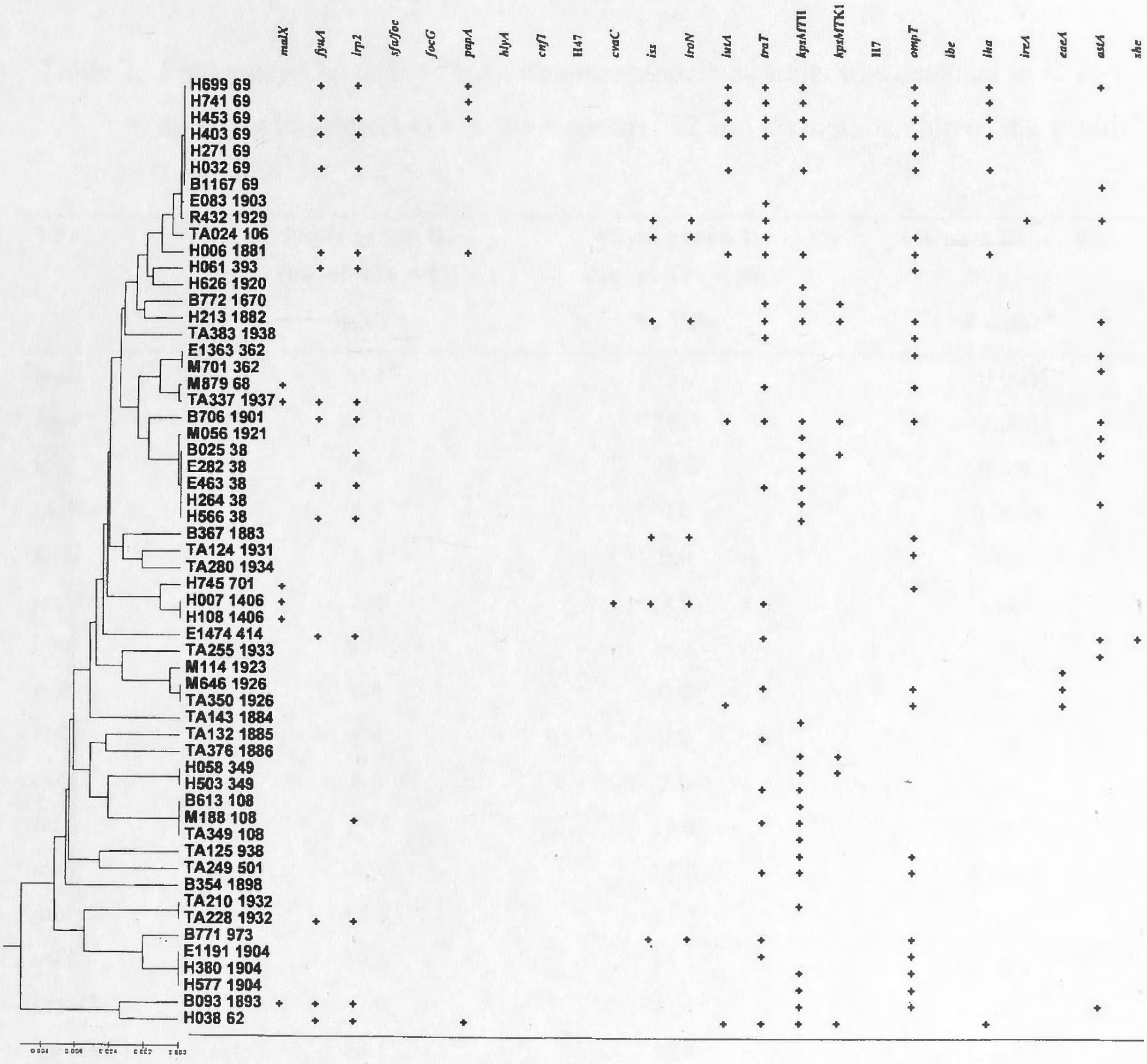


Figure 2. Phylogenetic distribution of 24 virulence factors of 55 phylo-group D and 2 phylo-group F (B093 and H038) strains. The phylogenetic tree was inferred by the Neigbour-Joining (NJ) algorithm applied to the genetic distances based on polymorphisms of MLST data of 7 housing genes.

TA280 ST1934 and B354 ST1898 contained none of the 24 traits investigated. The results also revealed that even closely related strains representing identical STs may have a variable pattern of VFs. These suggest that phylogenetic signal appears to have a little influence on the distribution of VFs to be found among *E. coli* strains. Other traits: *papA*, *iha*, *cvaC*, *ireA*, and *iutA* were absent in the environmental isolates in both phylo-groups B2 and D. The prevalence of each virulence factor was also calculated at the sequence type level (Table 2). In this analysis if any example of a particular ST was positive for the trait, all members of the ST were scored as positive for the trait. This analysis reveals that the frequency of many traits is significantly less

Table 2. Frequency (%) with which virulence-associated traits was detected in *E. coli* strains with respect to the phylo-group (B2 and D) membership of the strains[a].

VFs	**Phylo-group B2**	**Phylo-group D**	**Fisher's Exact Test**
	(no. of STs = 83)	**(no. of STs = 38)**	
	% VF+	**% VF+**	***P* value** [c]
malX	90.4 [b]	13.2	<0.0001
fyuA	83.1 [b]	26.3	<0.0001
irp2	84.3 [b]	29.0	<0.0001
sfa/foc	16.9 [b]	0.0	0.0048
focG	8.4	0.0	ns
papAH	8.4	7.9	ns
hlyA	9.6 [b]	0.0	ns
cnf1	8.4	0.0	ns
H47	8.4	0.0	ns
cvaC	6.0	2.6	ns
iss	19.3	15.8	ns
iroN	34.9 [b]	15.8	0.0330
iutA	13.3	13.2	ns
traT	39.8	55.3	ns
*kpsMT*II	47.0	47.4	ns
*kpsMT*K1	18.1	18.4	ns
H7	14.5 [b]	0.0	0.0174
ompT	84.3 [b]	47.4	<0.0001
ibe	66.3 [b]	0.0	<0.0001
iha	6.0	7.9	ns
ireA	7.2	2.6	ns
eaeA	8.4	5.3	ns
astA	20.5	26.3	ns
she	24.1 [b]	2.6	0.0035

[a] Produced/grouped at least one example of duplicate STs (e.g. ST95) was positive for a trait the ST was scored as +.

[b] Significant *P* values for Fisher's Exact Test

[c] Two-tailed Fisher's Exact Test *P* values for frequency differences across phylo-groups

ns, not significant

at the ST level than at the isolate level. For example, *papAH* was detected in 16% of the B2 isolates, but only in 8% of B2 STs. Similarly, *papAH* is detected in about 42% on phylo-group B2 isolates and 23% of phylo-group D isolates (Table 1), but at the ST level is equally frequent in B2 and D STs (Table 2).

3.3.3 Recombination rates in phylo-group B2 strains

Several different techniques were used to assess the extent of recombination in phylo-group B2 sequence types (STs) isolated from different sources (humans, mammals, birds, and environment) using the concatenated sequence data collected for the MLST characterization of the isolates.

The PHI statistic indicates significant levels of recombination among phylo-group B2 strains regardless of their source (Table 3). The LDhat and ClonalFrame estimates of the relative importance of recombination *versus* mutation are highly correlated (Table 3). The magnitude of the estimates also suggests that the relative importance of recombination *versus* mutation is greatest in B2 STs from humans and least important in B2 strains from the environment. However, the 95 % confidence intervals estimated by ClonalFrame reveal that estimates for each of the populations overlap. Further with the exception of the estimates for the environmentally derived STs the confidence intervals provide little evidence that recombination is a more important force than mutation in lineage diversification. That is, neither the the *r/m* or the ρ / θ ratios are significantly greater than 1.

Table 3. Recombination in phylo-group B2 strains using the PHI test, LDhat and ClonalFrame

Methods	**Humans** **(no. of STs = 22)**	**Mammals** **(no. of STs = 32)**	**Birds** **(no. of STs = 23)**	**Environment** **(no. of STs = 16)**
PHI test (ϕ_w)				
P value	<0.0001	<0.0001	<0.0001	<0.0001
LDhat				
θ, Watterson [a]	0.00641	0.00668	0.00728	0.00652
$2N_er$	97	123	96	54
$\frac{2N_er\ /\ \#\ \text{sites}}{\theta}$	4.42	5.38	3.85	2.42
$2N_er$ Moment [b]	164.204	132.901	129.875	75.909
ClonalFrame				
Mean θ (95% CI)	461.53 (29.94 – 1259.12)	857.05 (54.03 - 2,069.15)	1,991.18 (110.06 - 4,375.22)	1,723.35 (26.63 - 6,628.53)
Mean *R* (95% CI)	353.52 (15.69 - 669.83)	438.42 (95.76 - 677.85)	455.97 (47.43 - 675.25)	142.89 (0.06 - 644.90)
Mean *r/m* (95% CI)	1.79 (0.43 – 5.79)	1.37 (0.27 – 5.38)	0.64 (0.15 – 1.49)	0.11 (0.00 – 0.48)
Mean ρ/θ (95% CI)	1.18 (0.24 – 4.32)	0.91 (0.10 – 4.55)	0.30 (0.06 – 0.84)	0.05 (0.00 – 0.20)

[a] The theta per site based on a finite-sites approximation to Watterson's estimate

$2N_er$ = the population-scaled recombination rate for the whole region

[b] $2N_er$ estimated by moment method (Wakeley, 1997)

θ, the mutation rate

R, the recombination rate

r/m, the ratio of probabilities that a given site is altered through recombination and mutation.

ρ/θ, the ratio of rates at which recombination and mutation occur.

95% CI, 95 percent confidence interval

3.4 Discussion and Conclusion

In the present study, the frequency and distribution of extra-intestinal virulence traits in a collection of *E. coli* isolated from native Australian vertebrates (Gordon & Cowling, 2003) as compared to *E. coli* isolated from humans living in Australia (Gordon *et al.*, 2005) were reported. The frequency and distribution of virulence traits of studied *E. coli* isolates belonging to phylo-groups A, B1, B2, and D could be categorized into i) traits common in the human isolates but absent or virtually absent in the animal isolates (*papAH, iha, hlyD*, *cnf1*, and *afa/dra*), ii) traits more common in the human isolates than in the animal isolates (*iutA, fyuA, irp2, iroN, iss, kpsMTK1, ireA, she, sfa/foc*, and *focG*), iii) traits more common in the animal isolates than in the human isolates (*eaeA, astA, ibe*, and H7), and iv) traits common in both human and animal isolates (*fimH, malX, traT*, and *ompT*). Notably there was no trait that was common in the animal isolates but absent in the human isolates. As is well established the virulence-associated traits are not evenly distributed across all *E. coli* phylo-groups but predominantly found in *E. coli* belonging to phylo-groups B2 and D (Schubert et al., 2009, Clermont et al., 2001, Boyd and Hartl, 1998).

Pathogenicity-associated islands (PAIs), genetic elements encoding various virulence factors, are known to be normally present in pathogenic strains. Many of the most significant virulence genes: adhesins, extracellular protein secretion systems, and toxins, for several *E. coli* pathotypes are clustered together in PAIs. In this study, most of the traits investigated were traits associated with PAIs except for H7, *ompT, ibe,* and *she.* The PAI associated traits were found to be more significantly correlated with and concentrated in phylo-group B2 strains known to be responsible for extra-intestinal infections as compare to phylo-group D strains. However the distribution of virulence factors (VFs) among *E. coli* B2 strains was found to be diverse. Phylogenetic signal appears to have a little influence on the distribution of VFs to be found among *E. coli* B2 strains as even closely related strains representing identical sequence types (STs) may have a variable pattern of virulence traits. These suggest that the frequency of many traits is significantly less at the ST level than at the isolate level.

Evidences from this study reveal that the frequency and distribution of many traits varies with the source of isolation (human *versus* animal). Certain traits typically associated with PAIs are absent or very rare in animal isolates. Of these virulence factors, traits including *papAH*, *hlyA*, and *cnf1* were common in the human B2 isolates but were virtually absent in the animal B2 isolates. While *sfa/foc* and *iroN* were more common in the human B2 isolates than in the animal B2 isolates. The virulence traits associated with PAIs were most concentrated in human B2 isolates represented to ST95 and ST73. These *E. coli* virulence traits which are more common in human isolates might provide a selective advantage to those *E. coli* strains in term of their intestinal colonization in human hosts or hitchhiking alongside *E. coli* strains to confer a significant fitness advantage in particular extra-intestinal infections in humans. Therefore, traits which are more common in human isolates might be either very rare or virtually absent in animal isolates, as the distribution of virulence traits might be dependent on conferring a selective advantage to specific host organisms (Gordon, 1992). However the result in this study indicates that bird B2 isolates represented to ST127 seem to be an exception as they were found to carry virulence traits typically associated with PAIs.

Although the ST diversity in *E. coli* is extensive, some STs are very common (Gordon, 2010). According to the frequency of VFs in this study, evidences clearly suggest that the observed frequency of particular VFs associated with PAIs may be greatly determined by the relative abundance of particular common STs found in *E. coli* strains. Of these STs, STs such as ST95 and ST73 were predominately observed in human B2 strains. Therefore, that some particular VFs associated with PAIs are more common in human isolates compare to that animal isolates can be explained by the fact that *E. coli* B2 strains carrying PAIs represented to particular STs, e.g. ST95 and ST73 are very common in and restricted to human isolates. This might suggest that these STs abundant in human represent host specific STs. Therefore that the frequency of particular VFs associated with PAIs found to be common in human isolates results from successful STs carrying these PAIs in humans. For example, many phylo-group B2 strains with the serotypes O1:K1, O2:K1 and O18:K1 are members of the MLST ST95 group of strains (Mora et al., 2009, Clermont et al., 2011). In this study, the set of *papA-hlyA-cnf1* genes (P-fimbrial genes: adhesins, α-hemolysin gene: toxin, and cytotoxic necrotizing factor 1 gene: toxin, respectively) known to be encoded by UPEC-specific PAIs (Hacker and Kaper, 2000) were found to be common in the human B2 isolates but were virtually absent in the animal B2 isolates. This suggests that with these successful STs becoming abundant for there are many opportunities for these STs to spread among humans, but limited opportunity for them to spread into animals. However bird B2 isolates represented to ST127 is being a possible exception as they found to carry a number of virulence traits.

As conjugative transfer and homologous DNA recombination were reported to play a major role in horizontal transfer of PAIs within *E. coli* strains (Schubert et al., 2009). The result in this study also shows that the relative importance of recombination *versus* mutation is greatest in B2 STs from humans. This reveals the evidence of intraspecies recombination influencing the spread of PAIs within human B2 strains. On the other hand, the opportunity for STs to succeed in spreading into animals is infrequent. This might because there is not a huge amount of contact between *E. coli* animal strains to allow the transmission of PAIs to the populations. In addition, the virtual absence of VFs associated with PAIs in animal isolates might suggest that these particular VFs never absent in animal isolates (except for bird B2 isolates represented to ST127). However, instead of losing of particular VFs, for example *papA* (and/or very rare *sfa/foc*) as a set of adhesins in animal isolates, the intestinal adherence factor called intimin encoded by *eaeA* was found to be more common in animal strains. Previous study has demonstrated that genes coding for adhesins as well as combinations of single nucleotide polymorphisms (SNPs) in coding or regulatory regions could participate to the host specificity (Clermont et al., 2011). Therefore, these might suggest that the difference of sets of genes encode adhesins presented in human and animal strains might be dependent on conferring a selective advantage to specific host organisms. The evolution of independent pathogenic types of *E. coli* is also likely linked to the concomitant evolution of different mammalian hosts (Welch, 2006).

This present study, therefore, suggests that *E. coli* B2 strains isolated from animals might be a subset diversity of *E. coli* B2 strains. These human B2 and animal B2 strains are very closely related strains representing identical STs, however, they have a different pattern of virulence traits. Moreover, *E. coli* B2 strains isolated from animal hosts might be no potential pathogenic strains as traits which typically associated with PAIs are absent or very rare. However, a number of pathogenic strains might be more frequently detected in some animal species than other animal species (Ishii et al., 2007) as presented by bird B2 isolates represented to ST127 in this study.

In conclusion, the frequency and distribution of some traits associated with PAIs were found to be significantly correlated with and concentrated in phylo-group B2 strains. However, the phylogenetic signal appears to have a little influence on the distribution of VFs as even closely related strains representing identical STs may have different VF profiles. The difference of VF profiles among *E. coli* B2 strains varies with the source of isolation, humans *versus* animals. Among B2 strains, traits typically associated with PAIs are absent or very rare in animal isolates. The frequency observed may be greatly determined by the relative abundance of particular STs which are very common in human strains.

CHAPTER 4

Investigation of the evolution of conjugative plasmids in *E. coli* and their changing role in *E. coli* ecology

2 Poster presentations:

Khumphai, P. & Gordon, D. M. (2011). Evolution of Colicin Ib and Ia Plasmids in the RepI1 Group in *Escherichia coli*. The Australasian Genomic Technologies Association (AMATA) Conference, Canberra, ACT.

Khumphai, P. & Gordon, D. M. (2012). Evolution of Bacteriocin Plasmids in the RepFIB and RepFIIA Groups in *Escherichia coli*. The Asia Pacific Bioinformatics Conference, Melbourne, QLD.

4.1 Introduction

Plasmids are extrachromosomal replicons that are prevalent symbionts of bacteria. Generally, plasmids are not essential for normal bacterial growth. However they often code for genes involved in antibiotic and heavy metal resistance, virulence, and ecological interactions. That is traits that encode for their host's adaptation to the environment. Plasmids can propagate themselves vertically via cell division and many can propagate themselves horizontally usually by infectious transfer via mobilization or conjugation (Summers, 1996, Bergstrom et al., 2000).

The conjugative plasmid is self-transmissible, and ordinarily possesses all the genes necessary for transmission (transfer (*tra*) and origin of transfer (*oriT*) genes). Many conjugative plasmids harbor genes involved in antibiotic resistance or virulence that can spread between commensal and pathogenic bacteria (Martinez and Baquero, 2002). Conjugation between donor and recipient cells of some conjugative plasmids can occur between different species, genera, or kingdoms (Amabile-Cuevas and Chicurel, 1992). Thanks to these properties of conjugative plasmids, they are known to be key agents playing an important role in horizontal gene transmission among bacteria.

To be established and persist in bacterial population, in theory plasmids required the mechanisms responsible for the plasmid maintenance (Bergstrom et al., 2000).

Although plasmids can propagate themselves vertically via cell division, they cannot be maintained through vertical transfer alone due to a finite rate of segragation. Also evidence suggests that plasmids cannot be maintained as parasites through infectious transfer as observed infectious transfer rates are too low (Gordon, 1992, Bergstrom et al., 2000). In addition, there is nothing preventing plasmid borne genes, which confer a selective advantage from transferring to the chromosome in which case plasmid carriage become a cost. However plasmids have been reported to confer some sort of selective advantages to their hosts to be maintained by carrying beneficial genes to their host bacteria (Gordon, 1992). Therefore the possible mechanism responsible for the maintenance of these plasmids is that plasmids are probably hitchhiking alongside host bacteria to confer the significant fitness advantage in order to prominence in their host populations.

In *E. coli*, many plasmids types comprising a number of plasmid incompatibility (Inc) groups are known to occur among *E. coli* strains and they play an important role in the "adaptation" of bacterial populations (Frost et al., 2005). It is well known that *E. coli* includes both commensal strains with little ability to cause disease and pathogenic strains that are able to cause intestinal or extra-intestinal infections. The ability of *E. coli* to cause disease is due to the presence of a range of traits thought to enhance the ability of a strain to cause disease. Although these virulence-associated traits of *E. coli* can be located on the chromosome, many important virulence traits are encoded on large conjugative plasmids. Almost plasmids associated with *E. coli* virulence belong to the IncF group carrying the F transfer region for their self-transmissibility (Johnson and Nolan, 2009).

A previous study has suggested that bacterial plasmids are maintained within bacterial populations as they are key agents mediating a strain's adaptation to local environmental niches (Eberhard, 1990). These plasmids originally encoded traits thought to mediate competitive interactions among strains, traits known as bacteriocins. Bacteriocins are defined as antimicrobial proteins with a narrow killing range, that are toxic only to bacteria closely related to the producing strain (Riley and Wertz, 2002). These substances play a significant role in maintaining microbial biodiversity by acting as important mediators of intra- and interspecies interactions (Kirkup and Riley, 2004, Czaran et al., 2002). Bacteriocins produced by *E.coli* can be divided into two classes:

the colicins (Ia, Ib, B, and M etc.) and the microcins (V or ColV and H47 etc.). The colicin gene cluster includes an activity gene encoding the toxin, an immunity gene encoding an immunity protein to protect the cell from its own toxin, and a lysis gene, and these genes are plasmid encoded (Riley and Gordon, 1999). While the microcin gene cluster consists of an activity and an immunity genes which may be plasmid or chromosomally encoded (Gordon et al., 2007). In *E. coli*, the production of multiple bacteriocins by a single strain is a common phenomenon (Gordon and O'Brien, 2006). Several bacteriocins have been found to be encoded on the same large conjugative plasmid in *E. coli* strains more often than expected by chance (Gordon and O'Brien, 2006, Gordon et al., 2007).

However, many virulence traits have been recently acquired by *E. coli* and have become associated with *E. coli* plasmids, especially conjugative plasmids which encoded bacteriocins as important mediators of intra- and interspecies interactions (Riley and Gordon, 1992, Johnson and Nolan, 2009). As we have a relatively poor understanding of these conjugative plasmids and also the evolution of conjugative plasmids in *E. coli* because, until recently, genome data for a representative set of *E .coli* plasmids was unavailable. Due to a recent initiative of the Broad Institute of MIT and Harvard genome data is now available for over 100 strains of *E. coli*. This genome resource enabled us to determine that the majority of the conjugative plasmids in *E. coli* belong to the RepFIB or RepFIIA (IncF) backbone types, while some of them belong to the RepI1 (IncI1) group. By using a comparative genomics approach, the purpose of this study was to investigate the evolution of conjugative plasmids in *E. coli* belonging to the RepFIB /RepFIIA (IncF) and RepI1 (IncI1) groups in *E. coli*. The outcome of the project will lead to significant advances in our understanding of the evolution of conjugative plasmids. In addition, plasmids derived from this study will be broadly representative of the ancestral bacteriocin plasmids and the coassociation of bacteriocins to be found in *E. coli*.

4.2 Materials and Methods

4.2.1 Data set

Draft genomes of the conjugative plasmids investigated in this study were from the *E. coli* Antibiotic Resistance Database at the Broad Institute (http://www.broadinstitute.org/). Reference plasmid genomes were also retrieved from the National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/) to include in the study. Conjugative plasmids were divided into 2 groups based on their plasmid replication backbones: RepI1 (IncI1) and RepFIB/FIIA (IncF). *Salmonella enterica* serovar Typhimurium plasmid R64 (a prototypical RepI1 MDR-encoding plasmid) was used as a reference plasmid for RepI1 plasmid backbone (Kaper et al., 2004). A prototypical ColIb plasmid (pColIb-P9) of *Shigella*, belived to be in the same species division as *E. coli*, was included in this study. The main characteristics of these plasmids and their hosts are presented in Table 1.

The draft genome sequences of conjugative plasmids were extracted from draft genome sequences of *E. coli* strains in the *E. coli* Antibiotic Resistance Database (Broad Institute). For each host strain, sequence assemblies (supercontigs: a scaffold constructed by successively linking pairs of contigs sharing at least two forward-reverse links) from the draft genomes corresponding to conjugative plasmids were identified as representing plasmids if they contained: 1) plasmid backbone genes; 2) bacteriocin genes known to be associated with conjugative plasmids; and 3) plasmid-associated virulence genes.

I). Searching genes of the basic RepFIB/FIIA and RepI1 plasmid backbones: The presence of three core components of the basic RepFIB/FIIA and RepI1 plasmid backbones were determined: 1) replication genes (RepFIB/FIIA: *repA/repA1*; RepI1: *repZ*), 2) stability genes (*psiAB* and *sopAB*), and 3) a conjugative transfer (Tra) region were considered (Johnson and Nolan, 2009). Gene search and BLAST similarity search options via the database were performed together using gene names and nucleotide sequences of each region as queries, respectively. Gene names for the replication-associated region: *repA* and *repA1* (for RepFIB/FIIA) and *repZ* (for RepI1) were used as queries. Gene names for the stability region: *psiAB* and *sopAB* (for both plasmid

Table 1. Plasmids used in this study

A) RepI1 (IncI1) group (19 plasmids: 12 plasmids from Broad Institute and 7 plasmids from NCBI)

Plasmid/ Isolate	Replicon(s)	Source	Phylogroup	Size (bp)	Colicin	Reference
H397	I1	Human	B2	78.239	-	Broad Institute
H296	I1	Human	B2	84.776	-	Broad Institute
H660	I1	Human	B2	85.848	-	Broad Institute
TA255	I1	Mammal	D	113,874	-	Broad Institute
H383	I1	Human	A	122,889	-	Broad Institute
T426	I1	Animal	B1	135.928	-	Broad Institute
R64	I1		*Salmonella typhimurium*	120.826	-	NCBI
R424	I1	Reptile	A	90.725	Ib	Broad Institute
H223	I1	Human	B2	92.236	Ib	Broad Institute
TA024	I1	Mammal	D	95,622	Ib	Broad Institute
H489	I1	Human	A	118,898	Ib	Broad Institute
B367	I1, FIB	Bird	D	176.956	Ib, MBimm	Broad Institute
pColIb-P9	I1		*Shigella sonnei* P9	93,399	Ib	NCBI
pCVM29188_101	I1	Poultry	*Salmonella Kentucky* CVM29188	101,461	Ib	NCBI
pND11_107	I1	Porcine	unknown	107,138	Ib	NCBI
pO113	I1, FIB	Human	unknown	165.548	Ib	NCBI
TA447	I1	Mammal	E	118,840	Ia	Broad Institute
pUMNK88_91	I1	Porcine	A	90.868	Ia	NCBI
pND12_96	I1	Porcine	unknown	92,290	Ia	NCBI

B) RepFIB/RepFIIA (IncF) Group (55 plasmids: 43 plasmids from Broad Institute and 12 plasmids from NCBI)

Plasmid or Isolate	Replicon(s)	Source	Phylogroup	Size (bp)	Colicin	Reference
FVEC1302	FIB, FIIA	Human	D	77.108	-	Broad Institute
H454	FIB, FIIA	Human	A	102.291	-	Broad Institute
E1167	FIB, FIIA	Environment	B1	109,960	-	Broad Institute
T408	FIB, FIIA	Animal	B1	119,233	-	Broad Institute
M863	FIB, FIIA	Mammal	Clade I	161.535	-	Broad Institute
H120	FIB, FIIA	Human	B1	172.159	-	Broad Institute
B088	FIB, FIIA	Bird	B1	173.792	-	Broad Institute
TA271	FIB, FIIA	Mammal	B1	190.011	-	Broad Institute
TA464	FIIA	Mammal	B2	63.769	-	Broad Institute
H489	FIIA	Human	A	64.203	-	Broad Institute
TA280	FIIA	Mammal	D	66.203	-	Broad Institute
H736	FIIA	Human	A	69.306	-	Broad Institute
T426	FIIA	Animal	B1	89.620	-	Broad Institute
TW10509-2	FIIA		Clade I	103,880	-	Broad Institute
PUTI459	FIIA		D	109.592	-	Broad Institute
TW10509-1	FIIA		Clade I	148.785	-	Broad Institute
H378	FIIA	Human	B2	154.870	-	Broad Institute

Table 1. (cont.)

B) RepFIB/RepFIIA (IncF) Group (55 plasmids: 43 plasmids from Broad Institute and 12 plasmids from NCBI) (cont.)

Plasmid or Isolate	Replicon(s)	Source	Phylogroup	Size (bp)	Colicin	Reference
pEC-B24	FIIA	Human	unknown	73,801	BM	NCBI
B706	FIB, FIIA	Avian (Bird)	D	124,755	BMIa	Broad Institute
TA206	FIB, FIIA	Mammal	B2	168,728	BMIa	Broad Institute
TA435	FIB, FIIA	Mammal	B2	99,144	Ia	Broad Institute
TA141	FIB, FIIA	Mammal	B1	108,887	Ia	Broad Institute
R529	FIB, FIIA	Reptile	B1	110,030	Ia	Broad Institute
TA103	FIB, FIIA	Mammal	B2	110,881	Ia	Broad Institute
TA024	FIB, FIIA	Mammal	D	118,704	Ia	Broad Institute
H591	FIIA	Human	B1	93,003	Ia	Broad Institute
pO86A1	FIB, FIIA		unknown	120,730	IaIb	NCBI
H299	FIIA	Human	unknown	128,709	IaIb	Broad Institute
H660	FIB, FIIA	Human	B2	113,012	Iaimm	Broad Institute
pEC14-114	FIB, FIIA	Human	unknown	114,222	Iaimm	NCBI
H504	FIB, FIIA	Human	B2	115,061	Iaimm	Broad Institute
H263	FIB, FIIA	Human	B2	115,789	Iaimm	Broad Institute
p1ESCUM	FIB, FIIA	Human	D	122,301	Iaimm	NCBI
H413	FIB, FIIA	Human	B2	127,473	Iaimm	Broad Institute
FVEC1412	FIB, FIIA	Mammal	D	135,029	Iaimm	Broad Institute
FVEC1465	FIB, FIIA	Human	D	157,725	Iaimm	Broad Institute
H305	FIB, FIIA	Human	B2	127,401	Iaimm	Broad Institute
H296	FIB, FIIA	Human	B2	147,191	Iaimm	Broad Institute
H252	FIB, FIIA	Human	B2	124,318	IaV	Broad Institute
pECOS88	FIB, FIIA	Human	B2	133,853	IaV	NCBI
H461	FIB, FIIA	Human	B2	146,733	IaV	Broad Institute
pCVM29188-146	FIB, FIIA	Avian (Bird)	*Salmonella*	146,811	IaV	NCBI
pAPEC-O103-ColBM	FIB, FIIA	Avian (Chicken)	B2	124,705	MBimm	NCBI
pSMS35-130	FIB, FIIA	Environment	F	130,440	MBimm	NCBI
pVM01	FIB, FIIA	Avian	B2	151,002	MBimm	NCBI
M718	FIB, FIIA	Mammal	E	157,967	MBimm	Broad Institute
pAPEC-O1-ColBM	FIB, FIIA	Avian (Chicken)	B2	174,241	MBimm	NCBI
H420	FIB, FIIA	Human	B1	109,677	MBimm	Broad Institute
H299	FIB, FIIA	Human	unknown	148,344	MBimm	Broad Institute
B921	FIB, FIIA	Avian (Bird)	A	139,594	VMBimmIaimm	Broad Institute
pAPEC-1	FIB, FIIA	Avian (Chicken)	B2	103,275	V	NCBI
R527	FIB, FIIA	Reptile	B2	149,708	V	Broad Institute
B671	FIB, FIIA	Avian (Bird)	B2	153,934	V	Broad Institute
H218	FIB, FIIA	Human	B1	157,229	V	Broad Institute
pAPEC-O2-ColV	FIB, FIIA	Avian	B2	184,501	V	NCBI

backbones) were used queries. Nucleotide sequences of these genes were from reference plasmids: pAPEC-O2-ColV (accession no. AY545598) for RepFIB/FIIA and *Salmonella enterica* serovar Typhimurium plasmid R64 (accession no. AP005147) for RepI1 backbones, respectively. For the conjugative transfer (Tra) regions, nucleotide sequences of whole *tra* operons from *E. coli* F sex factor (accession no. AF112469) and plasmid R64 (accession no. AP005147) were used as reference sequences for RepFIB/FIIA and RepI1 plasmid backbones, respectively.

II). Extraction of E. coli isolates harboring interested bacteriocin genes: In this study, 4 colicins including Ia, Ib, B, and M and microcin V were considered. Gene search and BLAST similarity search options via the database were performed together using gene names and nucleotide sequences of bacteriocin operons as queries, respectively. Nucleotide sequences used as queries were from bacteriocin-plasmid reference strains retrieved from NCBI (Table 1).

III). Searching genes of putative virulence genes/clusters and antibiotic resistance genes known to be plasmid associated: A number of putative virulence genes/clusters known to be plasmid associated including *eitABCD*, *etsABC*, *hlyF*, *iroBCDEN*, *iucABC*, *iutA*, *iss/bor*, *sitABCD*, and *tsh* were considered. In addition, the antibiotic resistance profiles (ampicillin, kanamycin, streptomycin, trimethoprim, tetracycline, and chloramphenicol) for each plasmid were investigated. Gene search and BLAST similarity search options via the database were performed using gene names and nucleotide sequences of each gene/cluster retrieved from NCBI as queries, respectively

4.2.2 Plasmid assembly

Three core components of a conjugative plasmid: 1) plasmid backbone genes (RepFIB/FIIA and RepI1 plasmid backbones), 2) bacteriocin genes, and 3) virulence-associated genes were assembled in order to get the draft genome sequence of conjugative plasmids. To visualize plasmid genomes, Argo Genome Browser (http://www.broadinstitute.org/annotation/argo/) was used to determine the location of particular genes on the supercontig of each plasmid.

The draft genomes of the conjugative plasmids in the Broad Institute database can be divided into two groups. i) A complete conjugative plasmid where one or more of the three core components are located on the same supercontig (Table 2A). ii) An incomplete conjugative plasmid which three core components are not located on the same supercontig or which resulted from an assembly error (Table 2B). While, *E. coli* strains harboring chromosomally bacteriocin-encoding genes (ColV, *cvaA*, and *cvaB*) were also detected.

The supercontig was identified as a representing complete conjugative plasmid when it contained one or more of the three core components and only one conjugative transfer region was detected in that strain. However, in a strain, more than one conjugative transfer regions could be detected on different supercontigs in which each supercontig carrying one or more of the three core components (1. plasmid backbone genes: replication genes, stability genes, and a conjugative transfer region; 2. bacteriocin genes known to be associated with conjugative plasmids; 3. plasmid-associated virulence genes). In this case, each supercontig was identified as the representing complete conjugative plasmid when all plasmid backbone genes (replication genes, stability genes, and a conjugative transfer region) were found to be located together on each supercontig.

An incomplete conjugative plasmid that its three core components are not located on the same supercontig was investigated. All supercontigs (and/or contigs) harboring plasmid's core component were aligned and reordered against the closely bacteriocin reference plasmids harboring the same bacteriocin type using the progressiveMauve (Darling et al., 2010) and the Mauve Contig Mover (MCM) (Rissman et al., 2009) of the MAUVE program, respectively. In addition, because some genes may have more than one copy and may be plasmid or chromosomally encoded. Therefore, an assembly error may occur as an assembly program try to place reads (contigs) of that kind of genes, derived from genome sequencing, on the same supercontig. This led to the detection of a large supercontig that was expected to be an unusually large size of a plasmid, for example, a plasmid extracted from strain M718. All three core components of this plasmid were located on supercontig 19. This supercontig consists of contig 55 to contig 73 with a total length was around 231 kb. However, three core components of a conjugative plasmid (plasmid backbone genes, bacteriocin genes, and

Table 2. Conjugative plasmids extracted from the *E. coli* Antibiotic Resistance Database (http://www.broadinstitute.org/).

A) Forty-one complete conjugative plasmid which all 3 core components are located on the same supercontig.

Strain	**Bacteriocin**	**Plasmid backbone**				**Other bacteriocins**
(No. of conjugative		**Replication**	**Stability**		**Conjugative**	
plasmids in strain)			***psiAB***	***sopAB***	**transfer***	
H397 (1)	-	RepI1	+	-	R	
H296 (2)	-	RepI1	+	-	R	
	Iaimm	RepFIB, RepFIIA	+	+	F	Colicin-E2 immunity protein
H660 (2)	-	RepI1	+	-	R	
	Iaimm	RepFIB, RepFIIA	+	-	F	-
TA255 (1)	-	RepI1	+	-	R	
H383 (1)	-	RepI1	+	-	R	
T426 (2)	-	RepI1	+	-	R	
	-	RepFIIA	+	-	F	
H223 (1)	Ib	RepI1	+	-	R	*cvaA* and *cvaB* on chromosome
TA024 (2)	Ib	RepI1	+	-	R	-
	Ia	RepFIB, RepFIIA	+	+	F	-
H489 (2)	Ib	RepI1	+	-	R	ColV on chromosome
	-	RepFIIA	+	-	F	
B367 (1)	Ib, MBimm	RepI1, RepFIB	+	+	R	Remnant of ColV operon
FVEC1302 (1)	-	RepFIB, RepFIIA	+	-	F	
H454 (1)	-	RepFIB, RepFIIA	+	+	F	
E1167 (1)	-	RepFIB, RepFIIA	+	+	F	
T408 (1)	-	RepFIB, RepFIIA	+	+	F	
M863 (1)	-	RepFIB, RepFIIA	+	-	F	
H120 (1)	-	RepFIB, RepFIIA	+	-	F	
B088 (1)	-	RepFIB, RepFIIA	+	+	F	
TA271 (1)	-	RepFIB, RepFIIA	+	-	F	
TA464 (1)	-	RepFIIA	+	-	F	
TA280 (1)	-	RepFIIA	+	-	F	
H736 (1)	-	RepFIIA	+	-	F	
TW10509 (2)	-	RepFIIA	+	-	F	
	-	RepFIIA	+	-	R	
PUTI459 (1)	-	RepFIIA	+	-	R	
H378 (1)	-	RepFIIA	+	+	F	
TA206 (1)	BMIa	RepFIB, RepFIIA	+	-	R	ColA, ColE7, and ColK
TA435 (1)	Ia	RepFIB, RepFIIA	+	+	F	ColY
TA141 (1)	Ia	RepFIB, RepFIIA	+	+	F	-
TA103 (1)	Ia	RepFIB, RepFIIA	+	+	F	ColY
H591 (1)	Ia	RepFIIA	+	-	F	-

* F, *E. coli* F sex factor transfer region; R, the transfer region of *Salmonella enteric* serovar Typhimurium plasmid R64

Table 2. (cont.)

A) Forty-one complete conjugative plasmid which all 3 core components are located on the same supercontig. (cont.)

Strain	Bacteriocin	Plasmid backbone				Other bacteriocins
(No. of conjugative		Replication	Stability		Conjugative	
plasmids in strain)			*psiAB*	*sopAB*	transfer*	
H263 (1)	Iaimm	RcpFIB, RepFIIA	+	-	F	ColV on chromosome
FVEC1412 (1)	Iaimm	RepFIB, RcpFIIA	+	-	F	-
FVEC1465 (1)	Iaimm	RcpFIB, RepFIIA	+	-	F	*cvaA* and *cvaB* on chromosome
B921 (1)	VMBimmIaimm	RcpFIB, RepFIIA	+	+	F	-
B671 (1)	V	RcpFIB, RepFIIA	+	+	F	-
H281 (1)	V	RcpFIB, RcpFIIA	+	+	F	-

* F, *E. coli* F sex factor transfer region; R, the transfer region of *Salmonella enteric* serovar Typhimurium plasmid R64

B) Fourteen incomplete conjugative plasmid which three core components are not located on the same supercontig or which resulted from an assembly error.

Strain	Bacteriocin	Feature	Plasmid backbone				Other
(No. of			Replication	Stability		Conjugative	bacteriocins
conjugative plasmids in strain)				*psiAB*	*sopAB*	transfer*	
R424 (1)	Ib	Assembly error: A conjugative plasmid and another non-conjugative plasmid were assembled on the same large supercontig.	RepI1	+	-	R	-
TA447 (1)	Ia	Assembly error: A conjugative plasmid and another non-conjugative plasmid were assembled on the same large supercontig.	RepI1	+	-	R	-
B706 (1)	BMIa	Core plasmid components were not located on the same supercontig	RcpFIB, RepFIIA	+	+	F	Col E1
R529 (1)	Ia	Core plasmid components were not located on the same supercontig	RepFIB, RepFIIA	+	+	F	Col E1

* F, *E. coli* F sex factor transfer region; R, the transfer region of *Salmonella enteric* serovar Typhimurium plasmid R64

Table 2. (cont.)

B) Fourteen incomplete conjugative plasmid which three core components are not located on the same supercontig or which resulted from an assembly error. (cont.)

Strain (No. of conjugative plasmids in strain)	Bacteriocin	Feature	Plasmid backbone				Other bacteriocins
			Replication	Stability		Conjugative transfer*	
				psiAB	*sopAB*		
H299 (2)	Ialb	Core plasmid components were not located on the same supercontig	RepFIIA	+	-	R	Col E1
	MBimm		RepFIB, RepFIIA	+	+	F	-
H504 (1)	Iaimm	Core plasmid components were not located on the same supercontig	RepFIB, RepFIIA	+	+	F	-
H413 (1)	Iaimm	Core plasmid components were not located on the same supercontig	RepFIB, RepFIIA	+	+	F	-
H305 (1)	Iaimm	Core plasmid components were not located on the same supercontig	RepFIB, RepFIIA	+	+	F	Colicin-E2 immunity protein
H252 (1)	IaV	Core plasmid components were not located on the same supercontig	RepFIB, RepFIIA	+	+	F	-
H461 (1)	IaV	Core plasmid components were not located on the same supercontig	RepFIB, RepFIIA	+	+	F	ColE7
M718 (1)	MBimm	Assembly error: A conjugative plasmid and a part of the chromosome were assembled on the same large supercontig.	RepFIB, RepFIIA	+	+	F	-
H420 (1)	MBimm	Core plasmid components were not located on the same supercontig	RepFIB, RepFIIA	+	+	F	-
R527 (1)	V	Assembly error: Two copies of the *sit* cluster gene were assembled on the same large supercontig.	RepFIB, RepFIIA	+	+	F	-

* F, *E. coli* F sex factor transfer region; R, the transfer region of *Salmonella enteric* serovar Typhimurium plasmid R64

virulence-associated genes) were located on the first 158 kb started from contig 55 to contig 71. While nucleotide sequences of the rest contig 72 and contig 73 were similar to sequences of *E. coli* genome reference strains. Therefore, PCR reactions were performed to check the connections between contig 71 and contig 72, and contig 72 and contig 73 in order to clarify whether this is an assembly error or contig 72 and contig 73 are parts of the plasmid.

The results revealed that an assembly mistake occurred between contig 71 and contig 72 as there was no PCR product from this region. In contrast, a PCR product generated between contig 72 and contig 73 indicated that these two contigs were put together correctly. The result also demonstrated that the last gene on contig 71 and the first gene on contig 72 were genes of an iron acquisition cluster (iron transport protein, periplasmic-binding protein) which are encoded both on a plasmid and the chromosome in this isolate.

After treating incomplete conjugative plasmids by using a combination of bioinformatic tools, the total number of 55 extracted conjugative plasmids (12 RepI1 (IncI1) plasmids and 43 RepFIB/FIIA (IncF) plasmids) was then used for the comparative genome analysis.

4.2.3 Inferring the patterns of genome content evolution using whole-genome sequences

Conjugative plasmids in this study were divided into 2 groups based on their basic replication plasmid backbones, RepFIB/FIIA (IncF) and RepI1 (IncI1) groups (Table 1). Reference plasmid genomes for each group were also retrieved from the GenBank database for inclusion in this study (Table 1). RepFIB/FIIA (IncF) group includes 17 non-bacteriocin encoding and 38 bacteriocin encoding plasmids (Ia, Ib, B, M, and V). RepI1 (IncI1) group includes 7 non-bacteriocin encoding and 12 bacteriocin encoding plasmids (Ia and Ib). The phylogenetic diversity analysis based on whole-genome sequences was used to investigate the evolution of conjugative plasmids by performing the analysis separately for each plasmid group based on their replication backbone.

To construct the phylogeny of plasmids the core genome and the gene content data were used. The phylogenetic tree inferred from the core genome was constructed. The draft genomes for each plasmid retrieved from the Broad Institute as previously described were used for multiple genome alignments with GenBank reference plasmids using the progressiveMauve of the MAUVE program (Darling et al., 2010). The output files derived from progressiveMauve: the .xmfa (the eXtended Multi-FastA file format) file contains the complete genome alignment, the .bbcols file contains a region of the alignment where one or more genomes have a sequence element that one or more others lack, and the .backbone file contains regions conserved among subsets of the genomes under study were then used as input files for the next steps. The core genome was extracted from the .xmfa and .bbcols files using stripSubsetLCBs script (available at http://gel.ahabs.wisc.edu/mauve/snapshots/) to generate .xmfa files of the core genome sequences (core alignment blocks) greater than 500 nt for each group. The derived .xmfa files of the core genome sequences for each plasmid group were used as input files for constructing trees inferred from the core genome data using ClonalFrame V1.1 (Didelot and Falush, 2007).

The phylogeny inferred from the variable genome data (the dispensable genome), that is genes not found in all plasmids, was also constructed for comparative purposes. The gene content data were extracted from the .backbone file (derived from the progressiveMauve) using bbFilter script (available at http://gel.ahabs.wisc.edu /mauve/snapshots/) to generate the .bin file containing binary features (presence/absence) of blocks in a particular plasmid. The .bin files of gene content data for each group were converted to Nexus format (.nex) using GenAlEX (Peakall and Smouse, 2006) and then used to construct the phylogenetic tree using SplitTree4 (Parameters: GeneContentDistance>UPGMA>ConsensusTree>Phylogram) (Huson and Bryant, 2006).

The distribution of the variable genome among the different conjugative plasmids was visualized using GenoPlast (Didelot et al., 2008). Trees inferred from either the core or the variable genome together with the file tabulating the presence/absence of each DNA block in each of the plasmids were used as input files for GenoPlast.

The particular blocks of the variable genome present in one and more plasmids were also identified from the .backbone file generated from the progressiveMauve. The nucleotide sequences of these DNA blocks were investigated using Blast similarity search via the Board Institute and NCBI blast to identify the genes present the regions of the variable genome.

4.3 Results

4.3.1 Conjugative plasmids from the *E. coli* Antibiotic Resistance Database

For this study, a number of conjugative plasmids were extracted from all 92 *E.coli* strains in the *E. coli* Antibiotic Resistance Database (http://www.broadinstitute.org/). A total of 55 conjugative plasmids were detected: 12 RepI1 (IncI1) plasmids and 43 RepFIIA/RepFIB (IncF) plasmids. This result shows that 60% (55 out of 92) of these *E. coli* strains harbor conjugative plasmids. Of these *E. coli* plasmids, 56% (31 out of 55) encodes at least one bacteriocin.

4.3.2 Phylogenetic Diversity of Plasmids in the RepI1 (IncI1) Group in *E. coli*

4.3.2.1 Conjugative plasmids in the RepI1 (IncI1) group

In this study, 19 plasmids belonging to the RepI1 group (RepI1 and RepI1/FIB plasmids) including 17 plasmids containing only RepI1 (so-called RepI1 plasmids) and 2 plasmids containing RepI1 and RepFIB (so-called RepI1/RepFIB plasmids) were investigated (Table 1A). Although, 2 plasmids (a plasmid of strain B367 and pO113), also contain the RepFIB replication backbone, they were grouped into the RepI1 group as their transfer regions were found to share similarities with the transfer region of plasmid R64 which was used as reference transfer region for the IncI1 group.

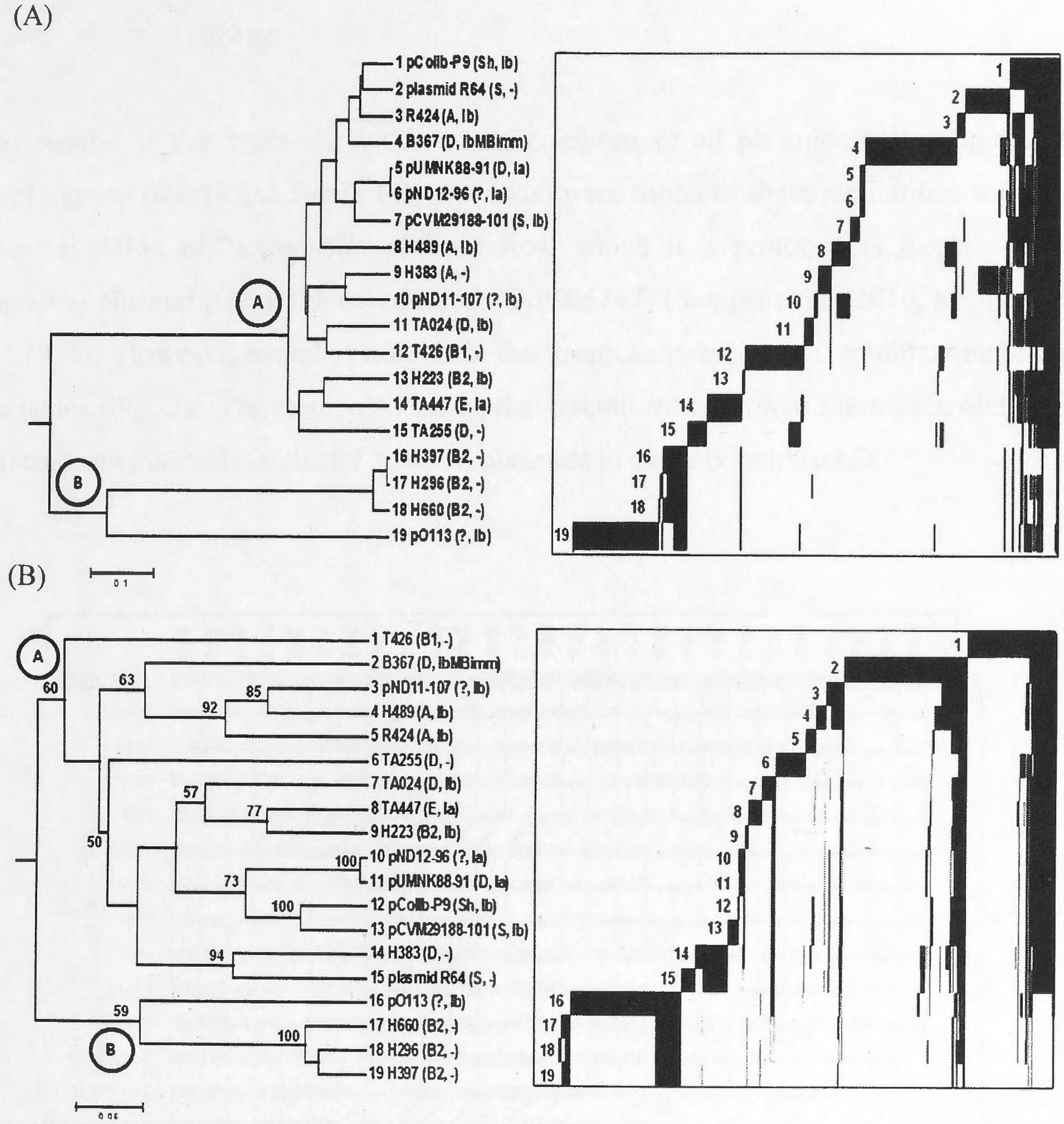
Conjugative bacteriocin plasmids belonging to RepI1 group in this study included (i) plasmids encoding the colicin Ib activity and immunity genes (ColIb plasmids), (ii) plasmids encoding the colicin Ia activity and immunity genes (ColIa plasmids), and (iii) a plasmid encoding the colicin Ib genes (activity and immunity genes), the colicin M genes (activity and immunity genes), and the colicin B immunity gene (with a truncated colicin B activity gene) which was referred as ColIbMBimm plasmid (a plasmid of strain B367) (Table 1A). However, the majority of bacteriocin-encoding plasmids belonging to RepI1 group are ColIb plasmids (67%: 8 out of 12), while ColIa plasmids (25%: 3 out of 12) and the ColIbMBimm plasmid (8%: 1 out of 12) were also detected.

4.3.2.2 Phylogenetic relationships of 19 plasmids in the RepI1 (IncI1) group

The average genome size of these 19 plasmids (Broad: 12 and NCBI: 7) is 110 kb (range from 78.24 to 176.96 kb) and, on average, 36% (39.11 kb) represents the core genome. The phylogenetic relationships of 19 plasmids belonging to RepI1 (IncI1) group inferred from the core genome (39.11 kb) shows that these plasmids spilt into 2 major clusters, cluster A and cluster B (Fig. 1A). Plasmids in cluster A include plasmids of *E. coli*, *Salmonella*, and *Shigella*. Most of plasmids in cluster A harbour either colicin Ia or Ib and they are plasmids of *E. coli* belonging to phylo-groups A, B1, B2, D, and E. ColIbMBimm plasmid of strain B367 (phylo-group D) was part of cluster A together with pColIb-P9 (*Shigella sonnei* P9: a prototypical ColIb plasmid), plasmid R64 (*Salmonella typhimurium*: a prototypical RepI1 MDR-encoding plasmid) (Krčméry et al., 1971, Kaper et al., 2004), and ColIb plasmid of strain R424 (phylogroup A). All plasmids in cluster B are non-bacteriocin encoding plasmids of *E. coli* belonging to phylo-group B2 except ColIb pO113. Based on the tree inferred using the core genome data, plasmids belonging to each cluster are monophyletic (Fig. 1A).

The phylogeny for 19 plasmids belonging to RepI1 (IncI1) group was also reconstructed using the gene content (Fig. 1B) in order to compare with the tree inferred from the core genome (Fig. 1A). These gene content data are binary features (block presence/absence data) which are present in a subset of the 19 plasmids belonging to RepI1 (IncI1) group. The phylogeny based on the gene content data reveals that these plasmids also fall into 2 major clusters, cluster A and cluster B (Fig. 1B), which is congruent with the phylogeny inferred from the core genome (Fig. 1A).

The variable genome pattern of each plasmid was shown on the right of the phylogenies inferred from the core genome and gene content data (Fig. 1A and 1B, respectively). The result shows that there are regions of the variable genome common to all cluster A strains but which are absent from strians in Cluster B and vise versa (Fig. 1).

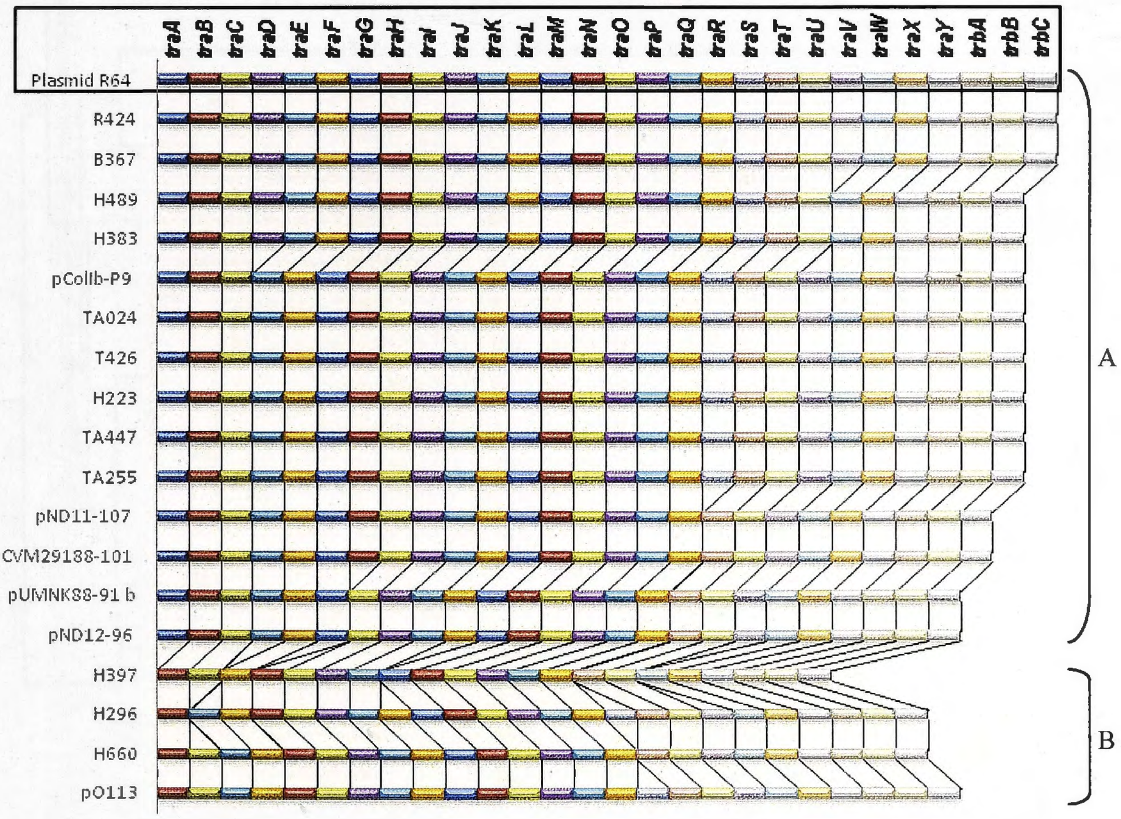


S; *Salmonella*, Sh; Shigella, ?; a genome sequence of the strain is not available, -; no colicin.

Figure 1. Phylogenetic relationships of the 19 plasmids in the RepI1 group. (A) Clonal genealogy of the plasmids inferred from the core genome using ClonalFrame V1.1 (100% consensus tree) and drawn with MEGA5. (B) UPGMA tree of the plasmids inferred from the gene content using SplitTree4 and drawn with MEGA5. Bootstrap values are based on 500 replicates and bootstrap confidence values greater than 50% are listed to the left of the nodes. The figure to the right of the phylogeny depicts the variable genome of each of the RepI1 plasmids. Each row of black bars corresponds to a plasmid in phylogeny. The horizontal width of the black bars corresponds to the size of the variable genome for each plasmid, and overlapping black bars corresponds to variable genome content shared by two or more plasmids.

4.3.2.3 Transfer region of conjugative plasmids in the RepI1 (IncI1) group

The results of this study show that transfer regions of all plasmids belonging to the RepI1 group (RepI1 and RepI1/FIB plasmids) were found to share similarities with the transfer region of *Sallmonella* plasmid R64, which is a prototypical RepI1 MDR-encoding plasmid (GenBank accession no. AP005147) (Sampei et al., 2010, Krčméry et al., 1971). However, overall, plasmids in this group carry a different combination of the *tra* genes (Fig. 2). The result also shows that overall *tra* operon of plasmid in cluster B differs from plasmids in cluster A, as all plasmids in clade B lack *traADG.*

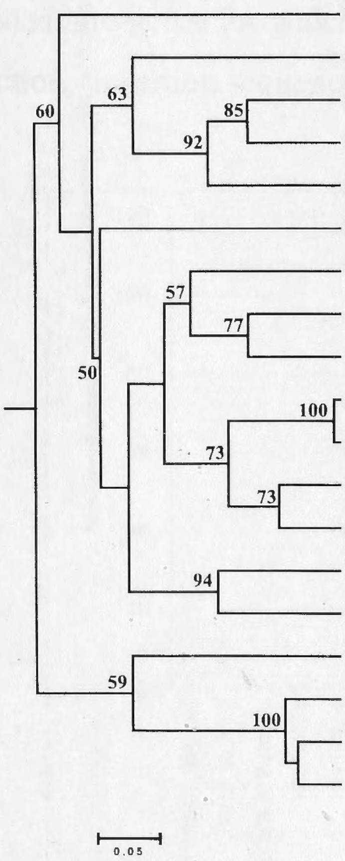


Note: The transfer gene region between *trbC* to *traA* of *Salmonella enterica* serovar Typhimurium plasmid R64 (GenBank accession no. AP005147) was used as the reference.

Figure 2. The transfer genes of 19 plasmids belonging to the RepI1 (IncI1) group. The top map is the *tra* operon of *Salmonella* plasmid R64. The black lines follow each colored block throughout the diagram to illustrate gene presence of each plasmid.

4.3.2.4 Antibiotic resistance and virulence-associated gene profiles

Further, the antibiotic resistance genes of 19 plasmids belonging to RepI1 group (RepI1 and RepI1/FIB plasmids) were investigated (Fig. 3). Overall, the antibiotic resistance profiles of plasmids were found to be diverse. Based on 6 antimicrobials investigated in this study, plasmids carrying antibiotic resistance genes were detected in 47% of the strains (9 out of 19 plasmids). Of these plasmids, 3 plasmids including ColIbMBimm



	Phylo-group	Bacteriocin	Inc group	Resistance profile[a]	eitABCD	estABC	hylF	iroBCDEN	iucABCD	iutA	iss	sitABCD	tsh
T426	B1		I1	C(c[b])	-	+	-	-	-	-	c	-	-
B367	D	IbMBimm	I1	S, Tm, T	-	-	+	+	-	-	+	+	-
pND11-107	?	Ib	I1	S, C	-	-	-	-	-	-	-	-	-
H489	A	Ib	I1	A, S, T, C	-	-	-	-	-	-	c	c	-
R424	A	Ib	I1	S*, T*, C(c)	-	-	-	-	-	-	-	-	-
TA255	D		I1	C(c)	-	-	-	-	-	-	-	-	-
TA024	D	Ib	I1		-	-	-	-	-	-	c	c	-
TA447	E	Ia	I1	C(c)	-	-	-	-	-	-	c	-	-
H223	B2	Ib	I1	C(c)	-	-	-	c	-	-	c	c	-
pND12-96	?	Ia	I1		-	-	-	-	-	-	-	-	-
pUMNK88-91	A	Ia	I1		-	-	-	-	-	-	-	-	-
pColIb-P9	*Sh*	Ib	I1		-	-	-	-	-	-	-	-	-
pCVM29188-101	*S*	Ib	I1	A	-	-	-	-	-	-	-	-	-
H383	A		I1	A, S, Tm, T, C(c)	-	-	-	-	-	-	c	-	-
plasmid R64	*S*		I1	S, T	-	-	-	-	-	-	-	-	-
pO113	?	Ib	I1		-	-	EHEC	-	-	-	-	-	-
H660	B2		I1	A	-	-	-	-	-	-	c	c	-
H296	B2		I1	A*, S*, C(c)	-	-	-	-	-	-	c	c	-
H397	B2		I1		-	-	-	-	-	-	c	c	-

?, unknown phylogroup; *S*; *Salmonella*, *Sh*; *Shigella sonnei* P9

[a] A, ampicillin; K, kanamycin; S, streptomycin; Tm, trimethoprim; T, tetracycline; C, chloramphenicol.

[b] c, chromosome

* located on another plasmid of the strain

EHEC, Enterohemorrhagic *Escherichia coli* hemolysin (Leyton et al., 2003)

Figure 3. Genotypic characteristics of 19 plasmids belonging to the RepI1 group.

plasmid of strain B367, ColIb plasmid of strain H489, and a plasmid of strain H383 (no colicin) represent plasmids capable of carrying multidrug resistance (MDR) functions (resistant to 3 out of 6 antimicrobials studied). Among these 3 plasmids carrying MDR functions, ColIbMBimm plasmid of strain B367 (phylogroup D) also encoded certain virulence factors including *hlyF, iroBCDEN, iss*, and *sitABCD* (Fig. 3). However, only 16% (3 out of 19) of plasmids belonging to the RepI1 group encoded putative virulence genes known to be plasmid associated.

4.3.2.5 Genes of conjugative plasmids in the RepI1 (IncI1) group

Genes within unique blocks of the variable genome were also investigated (Fig. 4 and Table 3). The result shows that genes within unique blocks of each strain can be divided into genes for which a function can be assigned, genes of unknown/ unclassified function, insertion sequence (IS)-like elements and transposase, and non-coding

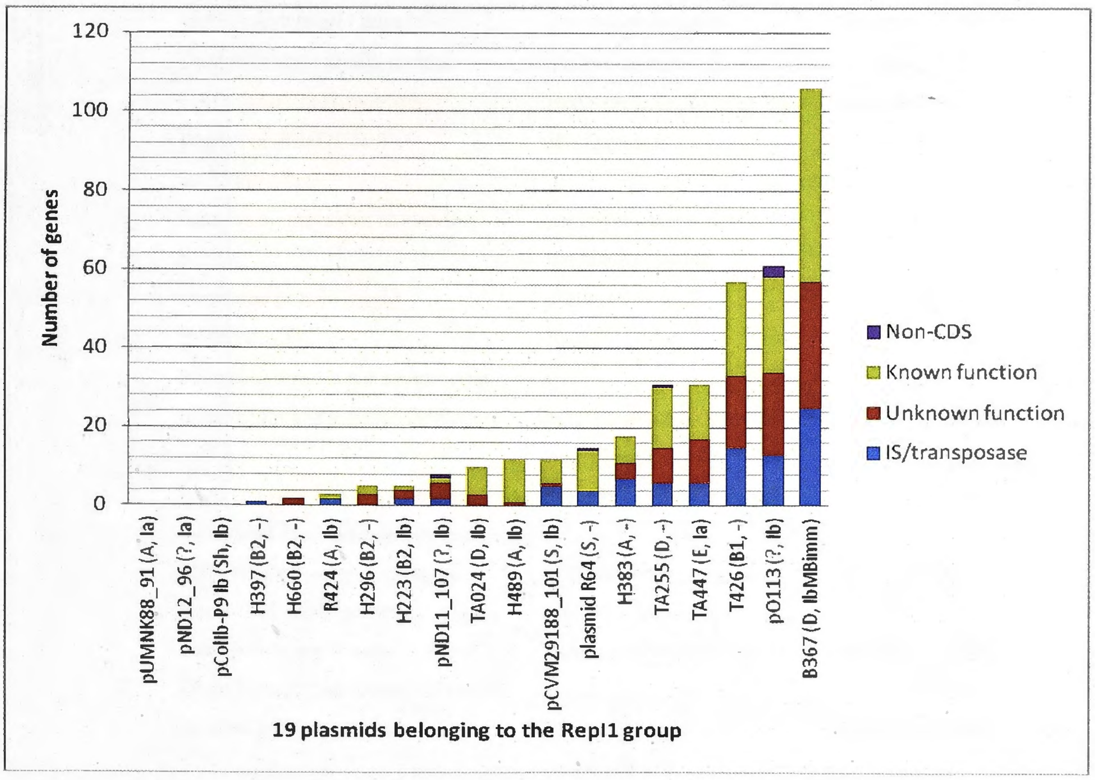


Figure 4. Genes present in each unique genomic region of conjugative plasmids in the RepI1 group. The x-axis is the genes present in a single plasmid. Coloured rectangles represent the proportion of non-coding sequences (purple), genes for which a function can be assigned (green), genes of unknown/unclassified function (red), and insertion sequence (IS)-like elements/transposase.

sequences (Table 3). Of these plasmids, CoIIa pND12-96 was found to share all genes with other plasmids belonging to the RepI1 (IncI1) group as no any unique gene was detected in this plasmid.

Table 3. List of unique genes within unique genomic regions of each conjugative plasmid in the RepI1 group. (segments smaller than 100nt were not considered).

Plasmids extracted from Broad Institute			
Isolate	**Gene product or feature**	**Gene Family**	**Partial or complete gene (nt)**
1. H397	transposase		107
2. H296	ychA-1	YchA-1	partial (121 out of 255)
	Putative cytoplasmic protein	yubE protein	partial (142 out of 222)
3. H660	Hypothetical protein		405
	Hypothetical protein		596
4. R424	RNA-directed DNA polymerase	group II intron	complete (1,902)
5. H223	Anti-restriction protein	hypothetical protein	partial (398 out of 699)
6. TA024	pyridine nucleotide-disulfide oxidoreductase	pyridine nucleotide-disulfide oxidoreductase	complete (1,533)
	MtN3/saliva family protein	glutathione synthase	complete (270)
	Oxidoreductase family protein	oxidoreductase	complete (1,017)
	HAD-superfamily hydrolase	HAD-superfamily hydrolase	complete (924)
	DegT/DnrJ/EryC1/StrS aminotransferase	WblQ protein	complete (1,239)
	hemolysin expression modulating protein	hemolysin expression modulating protein	complete (210)
7. TA255	chaperone protein AggD	chaperone AggD	complete (309)
	protein AggB	AggB	complete (429)
	outer membrane usher protein AfaC	-	complete (252)
	outer membrane usher protein AfaC	outer membrane usher protein AfaC	complete (2,028)
	ProQ activator of osmoprotectant transporter ProP superfamily protein	ProQ/FINO family protein	partial (245 out of 594)
	HTH-type transcriptional regulator AppY (M5 polypeptide)	HTH-type transcriptional regulator AppY	complete (765)
	outer membrane protein C (Porin OmpC) (Outer membraneprotein 1B)	outer membrane protein C	complete (1,110)
	putative prophage protein	fels-1 Prophage proteinprotein	partial (157 out of 361)
	F1845 adhesin operon regulatory protein	adhesin biosynthesis transcription regulatory protein	complete (270)
	putative histidyl-tRNA synthetase (Histidine--tRNA ligase)(HisRS)	conserved hypothetical protein	complete (504)

Table 3. (cont.)

Plasmids extracted from Broad Institute			
Isolate	**Gene product or feature**	**Gene Family**	**Partial or complete gene (nt)**
7. TA255 (cont.)	3-methyl-2-oxobutanoate hydroxymethyltransferase	3-methyl-2-oxobutanoate hydroxymethyltransferase	complete (798)
	pantoate--beta-alanine ligase	pantoate--beta-alanine ligase	complete (852)
8. H489	shufflon protein A	shufflon protein A	partial (143 out of 392)
	shufflon protein	shufflon protein	complete (193)
	Hg(II)-responsive transcriptional regulator	Hg(II)-responsive transcriptional regulator	complete (435)
	MerT mercuric transporter	mercuric transporter	complete (351)
	mercuric transporter periplasmic component protein	mercuric transporter periplasmic component	complete (276)
	MerC mercury resistance protein	MerC mercury resistance protein	partial (363 out of 462)
	mercuric reductase	mercury(II) reductase	partial (1,070 out of 1,695)
	mercuric resistence transcriptional repressor protein MerD	mercuric resistence transcriptional repressor protein MerD	complete (363)
	MerE protein	MerE protein	complete (237)
	EAL domain-containing protein	Urf2	complete (708)
	shufflon protein C	shufflon protein C	complete (156)
9. TA447	YgeA-1	YgeA-1	complete (252)
	DNA (cytosine-5-)-methyltransferase	DNA (cytosine-5-)-methyltransferase	partial (201 out of 639)
	YchA-1	YchA-1	complete (282)
	YdiA-1	YdiA-1	partial (333 out of 633)
	type VI secretion protein, VC_A0107 family	type VI secretion protein	complete (477)
	type VI secretion protein, EvpB/VC_A0108 family	EvpB/family type VI secretion protein	complete (1,485)
	type VI secretion protein, VC_A0110 family	type VI secretion protein	complete (1,803)
	type VI secretion protein, VC_A0111 family	type VI secretion protein	complete (933)
	Rhs element Vgr protein	rhs element Vgr protein	complete (2,043)
	paar	PAAR domain-containing protein	complete (255)
	type VI secretion-associated protein, ImpA family	ImpA family type VI secretion-associated protein	complete (1,077)
	type VI secretion protein, VC_A0114 family	type VI secretion protein	complete (1,380)
	type IV / VI secretion system protein, DotU family	DotU family protein type IV/VI secretion system protein	complete (699)
	putative chaperone-associated ATPase	ClpV1 family protein type VI secretion ATPase	complete (1,623)

Table 3. (cont.)

Plasmids extracted from Broad Institute

Isolate	Gene product or feature	Gene Family	Partial or complete gene (nt)
10. H383	dihydrofolate reductase	dihydrofolate reductase	complete (279)
	beta-lactamase	beta-lactamase	complete (861)
	initiator RepB protein	replication initiation protein	complete (612)
	glutathione synthase	glutathione synthase	complete (1,437)
	IPR001584: integrase catalytic domain protein	-	complete (354)
	invasion plasmid antigen, fragment	conserved hypothetical protein	complete (387)
	protein PsiB	PsiB	partial (116 out of 585)
11. T426	resolvase, N domain protein	resolvase	complete (573)
	outer membrane efflux family protein	outer membrane efflux protein	complete (1,371)
	macrolide export ATP-binding/permease protein MacB 2	macrolide export ATP-binding/permease MacB 2	partial (1,839 ot of 1,941)
	secretion protein, HlyD family	efflux transporter	complete (1,188)
	protease 7 (Protease VII) (Omptin) (Outermembrane protein 3B) (Protease A)	-	partial (502 out of 630)
	protease 7 (Protease VII) (Omptin) (Outermembrane protein 3B) (Protease A)	-	complete (192)
	toxin-antitoxin system, toxin component, RelE family	RelE family toxin-antitoxin system	partial (101 out of 204)
	lipid A biosynthesis (KDO)2-(lauroyl)-lipid IVA acyltransferase	-	complete (873)
	required for proper localization of IcsA (VirG) at the surface of bacteria	VirK domain-containing protein	complete (933)
	UDP-sugar diphosphatase	-	complete (663)
	putative carbohydrate transport protein	polysaccharide deacetylase	complete (822)
	outer membrane protease E (E protein)	omptin family protein	complete (921)
	adhesin/virulence factor Hek	adhesin/virulence factor Hek	complete (759)
	putative cytoplasmic protein	yubE protein	partial (168 out of 366)
	related to plasmid partition protein ParA	parA protein	complete (633)
	HTH-type transcriptional regulator TdcA (Tdc operon transcriptionalactivator)	HTH-type transcriptional regulator TdcA	partial (713 out of 930)
	hypothetical metal-dependent amidase/aminoacylase/carboxypeptidase	peptidase M20D	complete (1,179)
	transcriptional activator AarP	-	complete (204)

Table 3. (cont.)

Plasmids extracted from Broad Institute

Isolate	Gene product or feature	Gene Family	Partial or complete gene (nt)
11. T426 (cont.)	lipid A biosynthesis (KDO)2-(lauroyl)-lipid iva acyltransferase	-	complete (438)
	putative glycosyl transferase, group 1 family	glycosyl transferase group 1	complete (1,152)
	putative carbohydrate transport protein	-	complete (822)
	protein PerC (Protein bfpW)	-	complete (189)
12. B367	putative cobalamin synthesis protein	cobalamin synthesis protein cobW domain	complete (1,194)
	lipoprotein bor	bor protein	complete (309)
	endopeptidase (Lysis protein Rz)	endopeptidase	complete (278)
	putative glucosyltransferase	iroB protein	complete (1,164)
	ABC transporter, permease/ATP-binding protein	ABC transporter	complete (3,660)
	ferric enterochelin esterase	ferric enterochelin esterase	complete (1,230)
	IroE protein	IroE protein	complete (957)
	ferric enterobactin receptor	ferric enterobactin receptor	complete (2,178)
	colicin-V (Microcin-V bacteriocin)	colicin-V	complete (312)
	Colicin-V immunity protein	98 %Identity with Colicin-V immunity protein	partial (192 out of 234)
	colicin-B	colicin-B	partial (150 out of 1536)
	colicin-B immunity protein (Microcin-B immunity protein)	colicin-B immunity protein	complete (528)
	colicin-M	colicin-M	complete (468)
	colicin-M immunity protein (Microcin-M immunity protein)	colicin-M immunity protein	complete (339)
	toxin-antitoxin system, antitoxin component, AbrB famil	SpoVT/AbrB domain-containing protein	complete (231)
	toxin-antitoxin system, toxin component, PIN family	mvpA protein	complete (417)
	ProQ activator of osmoprotectant transporter ProP superfamily protein	ProQ/FINO family protein	complete (618)
	putative permease	hypothetical protein	partial (755 out of 1,080)
	transcriptional regulator, ArsR family	arsR family regulatory protein	complete (324)
	Hg(II)-responsive transcriptional regulator	Hg(II)-responsive transcriptional regulator	complete (435)
	mercuric transport protein (Mercury ion transport protein)	mercuric transporter	complete (351)
	mercuric transport protein periplasmic component	mercuric transporter periplasmic component	complete (276)
	mercuric resistance protein MerC	MerC mercury resistance protein	complete (462)
	mercury(II) reductase	mercury(II) reductase	complete (1,695)

Table 3. (cont.)

Plasmids extracted from Broad Institute			
Isolate	**Gene product or feature**	**Gene Family**	**Partial or complete gene (nt)**
12. B367 (cont.)	mercuric resistence transcriptional repressor protein MerD	mercuric resistence transcriptional repressor protein MerD	complete (231)
	IPR001633: EAL domain protein	MerE protein	complete (237)
	IPR001633: EAL domain protein	Urf2	complete (798)
	acetyltransferase, GNAT family	GNAT family acetyltransferase	complete (501)
	dihydropteroate synthase	dihydropteroate synthase	complete (867)
	quaternary ammonium compound-resistance protein QacE (Quaternaryammonium determinant E)	multidrug transporter EmrE	complete (348)
	dihydrofolate reductase	dihydrofolate reductase	partial (279 out of 474)
	abortive infection protein	abortive infection protein	complete (345)
	colicin-Ib immunity protein	colicin-Ib immunity protein	partial (132 out of 348)
	colicin-Ib	colicin pore forming domain-containing protein	partial (123 out of 1,881)
	type IV pilus biogenesis protein PilV	shufflon protein	partial (219 out of 1,338)
	shufflon protein B	-	partial (219 out of 237)
	anti-restriction protein	hypothetical protein	partial (158 out of 348)
	protein SopB (Plasmid partition protein B)	SopB	complete (972)
	protein SopA (Plasmid partition protein A)	SopA	complete (1,119)
	phosphopyruvate hydratase	phosphopyruvate hydratase	complete (183)
	iron chelate ABC transporter, permease protein	ABC 3 transporter	complete (858)
	iron chelate ABC transporter, permease protein AfeC	ABC 3 transporter	complete (858)
	iron chelate ABC transporter, ATP-binding protein	ABC transporter	complete (828)
	iron chelate ABC transporter, periplasmic iron chelate-binding protein	iron transporter	complete (915)
	replication protein RepA	initiator Replication protein	complete (978)
	probable site-specific recombinase	phage integrase	complete (741)
	putative transcriptional regulator	mig-14	complete (903)
	nucleoside-diphosphate-sugar epimerases	nucleoside-diphosphate-sugar epimerases	partial (938 out of 1,110)
	protease 7 (Protease VII) (Omptin) (Outermembrane protein 3B) (Protease A)	protease 7	complete (954)

Table 3. (cont.)

Reference plasmids from NCBI			
Plasmid	**Gene or annotation**	**Product**	**Partial or complete gene (nt)**
13. pND11_107	*excA*	surface exclusion protein ExcA	partial (109 out of 615)
14. pCVM29188_101	SeKA_C0014	TnpA	complete (1,263)
	SeKA_C0016	beta-lactamase	partial (1,090 out of 1,230)
	SeKA_C0017	outer membrane lipoprotein blc	complete (534)
	SeKA_C0018	quaternary ammonium compound-resistance protein SugE2	partial (348 out of 435)
	ccdA1	post-segregation antitoxin CcdA	complete (219)
	ccdB2	cytotoxic protein CcdB	complete (306)
15. plasmid R64	*tetD*	transcriptional regulator of tet operon	complete (417)
	tetC	transcriptional regulator of tet operon	complete (594)
	tetA	tetracycline resistance protein	complete (1,206)
	tetR	tetracycline resistance protein TetR	partial (421 out of 624)
	ybdA	transcriptional regulator of tet operon	partial (598 out of 687)
	ybeA	ybeA	complete (489)
	ybeB	ybeB	partial (198 out of 321)
	ybfA	sodium/glutamate symporter	partial (152 out of 1,206)
	R64_p118	PilVD' C-terminal segment (Partial start)	complete (288)
	R64_p119 (Partial start)	PilVA' C-terminal segment (Partial start)	partial (102 out of 225)
16. pO113	LH0049	Membrane-associated, metal-dependent hydrolase	complete (1,545)
	epeA	Autotransporter protease	complete (4,080)
	LH0059	Neurotensin receptor R8	complete (312)
	LH0111	Outermembrane receptor	partial (1,825 out of 1,974)
	LH0112	Component of an ABC transporter system	complete (792)
	traT	Complement resistance protein	complete (732)
	repA	Initiator replication protein	partial (916 out of 978)
	LH0141	Transcriptional regulator	complete (549)
	ehxD	Hemolysin D	complete (1,440)
	ehxB	Hemolysin B	complete (2,118)

Table 3. (cont.)

Reference plasmids from NCBI			
Plasmid	**Gene or annotation**	**Product**	**Partial or complete gene (nt)**
	ehxA	Hemolysin A	complete (2,997)
	ehxC	Hemolysin C	partial (271 out of 516)
	LH0146	OmpA-family lipoprotein	complete (804)
	LH0147	Adhesin	complete (4,296)
	LH0148	Regulatory protein	complete (285)
	lldP	L-lactate permease	complete (1,557)
	LH0159	Dehydrogenase subunit	complete (720)
	LH0162	Neurotensin receptor R8	complete (312)
	espP	Extracellular serine protease	complete (3,903)
	LH0168	Nuclease	complete (414)
	Saa	STEC autoagglutinating adhesin	complete (1,605)
	lha	Outer membrane receptor for ferrienterochelin and colicins	partial (1,987 out of 2,088)
	sub	Subtilase cytotoxin, subunit B	complete (426)
	subA	Subtilase cytotoxin, subunit A	complete (1,044)
17. pColIb-P9	segments smaller than 100nt		
18. pUMNK88-91	segments smaller than 100nt		
19. pND12-96	None		

The particular blocks that share by all plasmids in the RepI1 group based on the results of the phylogeny inferred from core genome data (Fig. 1A) were investigated (Table 4). The particular blocks that share by plasmids in cluster A and cluster B were also investigated as shown in Table 5 and 6, respectively. All plasmids in the RepI1 group share similarity based upon the replication genes (*repZ* and *repY*), stability gene (*psiA*), and genes involved in conjugation (Table 4). However, some genes shared by plasmids in cluster A and by plasmids in cluster B were also found to be genes involved in conjugation (*trb*, *tra*, and *pil*). This can be explained by the fact that nucleotide sequences of genes involved in conjugation of plasmids in cluster A differ from cluster B. However, those genes share similarity based on amino acid sequences. The previous study has demonstrated that the transfer region of pO113 shares a high degree of amino acid sequence similarity with plasmid R64 *tra* gene products (Leyton et al., 2003). Therefore, plasmids in the RepI1 group were clustered separately based on similarity of genes involved in conjugation into cluster A (based on plasmid R64) and cluster B (based on pO113). However overall, for most of the plasmids, the balance of the genome consists of genes that are largely unique to each plasmid (Fig. 1A and 1B).

Table 4. List of genes with known functions on conserved segments sharing by all conjugative plasmids in the RepI1 group based on the core genome data.

Gene	Function or product	Gene	Function or product
repY	Regulator of *repZ* expression	*traT*	TraT transfer protein
repZ	Replication initiation protein	*traR*	TraR transfer protein
psiA	Plasmid SOS inhibition protein A	*traQ*	TraQ transfer protein
ygbA	DUF 1472 family protein	*traP*	TraP transfer protein
ard	Antirestriction protein	*traO*	TraO transfer protein
ygcA	Hypothetical protein	*traN*	TraN transfer protein
ygdA	Hypothetical protein	*traM*	IcmL family protein
ygdB	Hypothetical protein	*sogL*	SogL DNA primase
ygeA	Hypothetical protein	*sogS*	SogS transfer protein
ygfA	Hypothetical protein	*traK*	TraK transfer protein
ygfB	Hypothetical protein	*traJ*	Nucleotide-binding protein
yggA	Hypothetical protein	*traI*	Lipoprotein
nikA	NikA *oriT*-specific DNA binding protein	*traH*	Lipoprotein
nikB	NikB relaxase	*traF*	Hypothetical protein
trbC	TrbC transfer protein	*traE*	Hypothetical protein
traB	Protein disulfide isomerase	*pilR*	Integral membrane protein
trbA	conjugal transfer protein TrbA	*pilQ*	Nucleotide-binding protein
traY	Integral membrane protein	*pilN*	Lipoprotein
traX	TraX transfer protein	*pilL*	lipoprotein
traW	Lipoprotein	*traB*	Transcription termination factor NusG
traU	Nucleotide-binding protein		

Note: Genes of plasmid R64 (GenBank: accession no. AP005147) was used as a reference

Table 5. List of genes with known functions on conserved segments shared by conjugative plasmids in cluster A (absent from cluster B).

Length of alignment (nt)	**Gene**	**Function or product**
1,057	*trbB*	conjugal transfer protein TrbB
691	*traP*	IncI1 conjugal transfer protein TraP
3,582	*sogL*	DNA primase SogL
893	*traK*	IncI1 conjugal transfer protein TraK
577	*traG*	IncI1 conjugal transfer protein TraG
1,141	*rci*	IncI1 shufflon-specific DNA recombinase Rci
553	*pilV*	typeIV prepilin, IncI1 shufflon protein
626	*pilU*	IncI1 conjugal transfer prepilin peptidase PilU
529	*pilT*	IncI1 conjugal transfer lytic transglycosylase PilT
561	*pilS*	IncI1 conjugal transfer prepilin protein PilS
992	*pilR*	IncI1 conjugal transfer pilus biogenesis protein PilR
1,461	*pilQ*	IncI1 conjugal transfer protein PilQ
453	*pilP*	IncI1 conjugal transfer pilus biogenesis protein PilP
1,296	partial sequence of *pilO*	IncI1 conjugal transfer protein PilO
1,022	*pilL*	IncI1 conjugal transfer protein PilL
431	*pilM*	IncI1 conjugal transfer protein PilM
1,639	*pilN*	IncI1 conjugal transfer secretion protein PilN
591	*pilK*	IncI1 conjugal transfer protein PilK
593	Transcription termination protein nusG	Transcription termination protein nusG
684	*traC*	IncI1 conjugal transfer protein TraC
288	*traA*	IncI1 conjugal transfer protein TraA

Note: Genes of pUMNK88_91 (GenBank: accession no. CP002731) was used as a reference

Table 6. List of genes with known functions on conserved segments shared by conjugative plasmids in cluster B (absent from cluster A).

Length of alignment (nt)	Gene	Function or product
634	*traB*	Transcription antitermination factor
600	*traC*	Transfer protein C
606	*yqiJ*	YQIJ Inner membrane protein yqiJ
1,673	*yqiK*	SPFH domain/SPFH family protein
240	*pilI*	plasmid conjugative transfer protein PilI
976	*pilL*	Lipoprotein
224	*pilM*	Exported protein
1,422	*pilN*	Pilus assembly protein
518	*pilO*	Pilin accessory protein
1,006	*pilQ*	ATPase
1,019	*pilV*	Minor pilin subunit
3,495	*sogL*	DNA primase competative inhibitor
645	*traP*	TraP protein
1,632	hypothetical protein	hypothetical protein
1,150	*trbA*	Conjugal transfer protein
740	*trbB*	Conjugal transfer protein
2,167	*trbC*	Conjugal transfer protein
961	hypothetical protein	hypothetical protein

Note: Genes of pO113 (GenBank: accession no. AY258503) was used as a reference

4.3.3 Phylogenetic Diversity of Plasmids in the RepFIB/RepFIIA (IncF) Group in *E. coli*

4.3.3.1 Conjugative plasmids in the RepFIB/RepFIIA (IncF) group

In this study, 55 plasmids belonging to RepFIB/RepFIIA (IncF) group (RepFIIA and RepFIB/RepFIIA plasmids), including 12 plasmids containing only RepFIIA (so-called RepFIIA plasmids) and 43 plasmids containing RepFIB and RepFIIA (so-called RepFIB/RepFIIA plasmids), were investigated (Table 1B).

Conjugative bacteriocin plasmids belonging to RepFIB/RepFIIA (IncF) group included 1) a ColBM plasmid (3%: 1 out of 38), 2) ColBMIa plasmids (5%: 2 out of 38), 3) ColMBimm plasmids (18%: 7 out of 38), 4) ColIa plasmids (16%: 6 out of 38),

5) ColIaIb plasmids (5%: 2 out of 38), 6) ColIaimm plasmids; plasmids encoding only the colicin Ia immunity gene (26%: 10 out of 38), 7) ColIaV plasmids (11%: 4 out of 38), 8) a ColVMBimmIaimm plasmid encoding the colicin V operon, the colicin M genes (activity and immunity genes), the colicin B immunity gene (with a truncated colicin B activity gene), and the colicin Ia immunity gene (with a truncated colicin Ia activity gene) (3%: 1 out of 38), and 9) ColV plasmids (13%: 5 out of 38) (Table 1B). Overall, of these plasmids, 69% (38 out of 55) encode bacteriocins.

4.3.3.2 Phylogenetic relationships of 55 plasmids in the RepFIB/RepFIIA (IncF) group

The average genome size of these 55 plasmids (Broad: 43 and NCBI: 12) is 126 kb (range from 63.8 to 190.0 kb). However the size of core genome of these 55 plasmids is less than 0.5% (577 bp) that is a partial sequence of RepFIIA *repA1* replication gene. Consequently, there was too little data available to determine the phylogenetic relationships of the plasmids based on their core genome (Fig. 5A). Although the phylogeny of these plasmids was constructed from the small core genome data, overall, they were fall into 2 major clusters: cluster A and cluster B, in which plasmids encoding the same bacteriocin type (*e.g.* Iaimm) and plasmids having coassociation of bacteriocin (*e.g.* IaV, MBimm) were clustered together for each group as expected (Fig. 5A).

Consequently the relationships among these plasmids were investigated using their variable genome content. The gene content tree for the 55 plasmids belonging to RepFIB/RepFIIA (IncF) group reveals that these plasmids fall into 2 major clusters, cluster A and cluster B (Fig. 5B). The cluster A includes plasmids encoding ColIaimm, ColIa, ColV, IaIb, BMIa, BM, and non-bacteriocin encoding plasmids. The cluster B includes ColV, coassociated-bacteriocin encoding (MBimm, IaV, and VMBimmIaimm), and non-bacteriocin encoding plasmids. Overall, there is only small fraction of gene content share by IncF plasmids belonging to each cluster. The result also shows that, for most IncF plasmids, the balance of entire genome consists of genes unique to each plasmid.

However overall the results derived from these 2 trees: Figures 5A and 5B were congruent. The phylogeny based on the partial sequence of RepFIIA *repA1* replication

(A) Based on core genome data

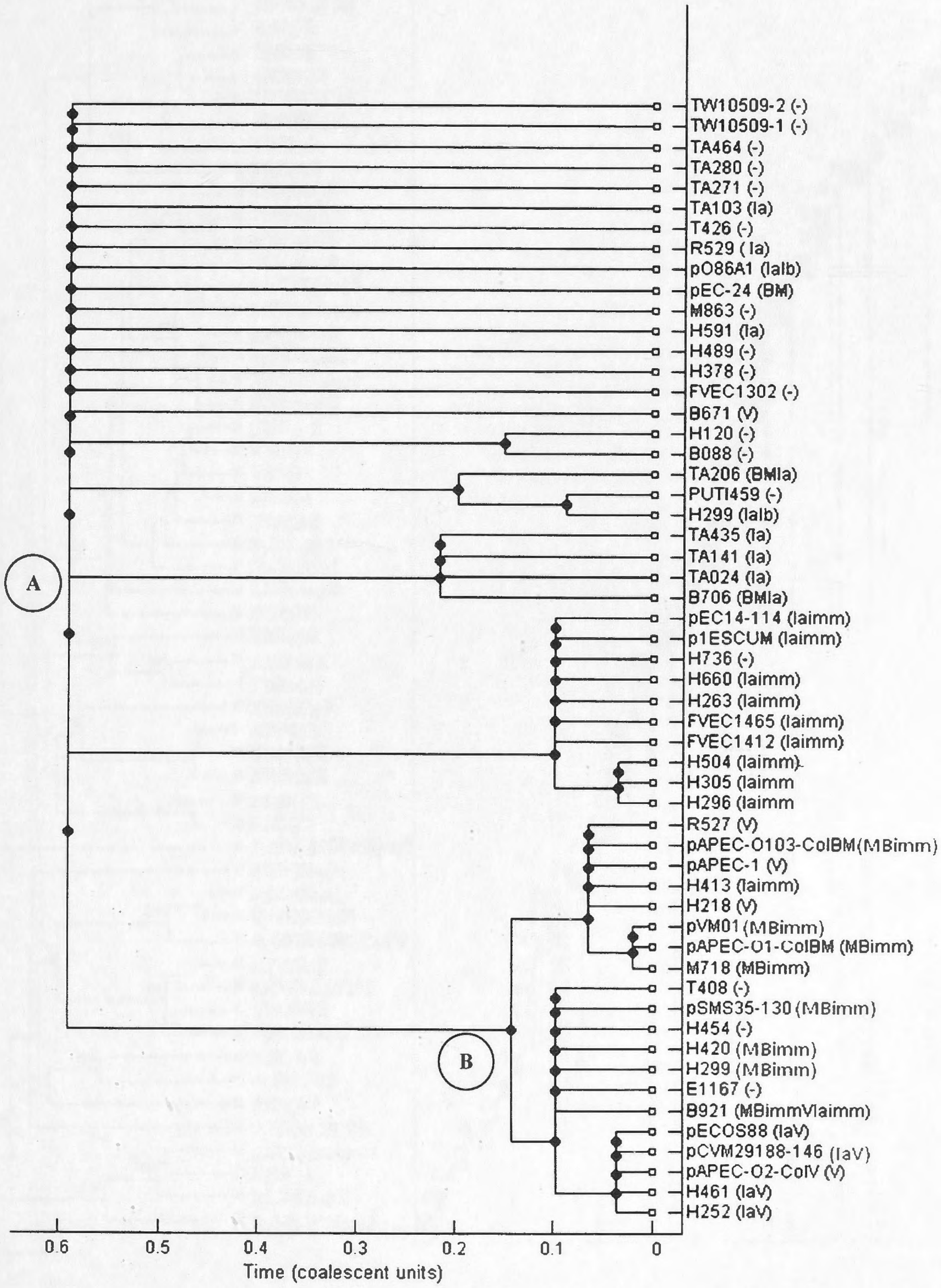


Figure 5. Phylogenetic relationships of the 55 plasmids belonging to RepFIB/RepFIIA group. (A) Clonal genealogy of the plasmids inferred from the core genome using ClonalFrame V1.1 (75% consensus tree). The core genome used to represent the genealogy of plasmids is a partial sequence of RepFIIA *repA1* replication gene; 577 nt in length.

B) Based on gene content data

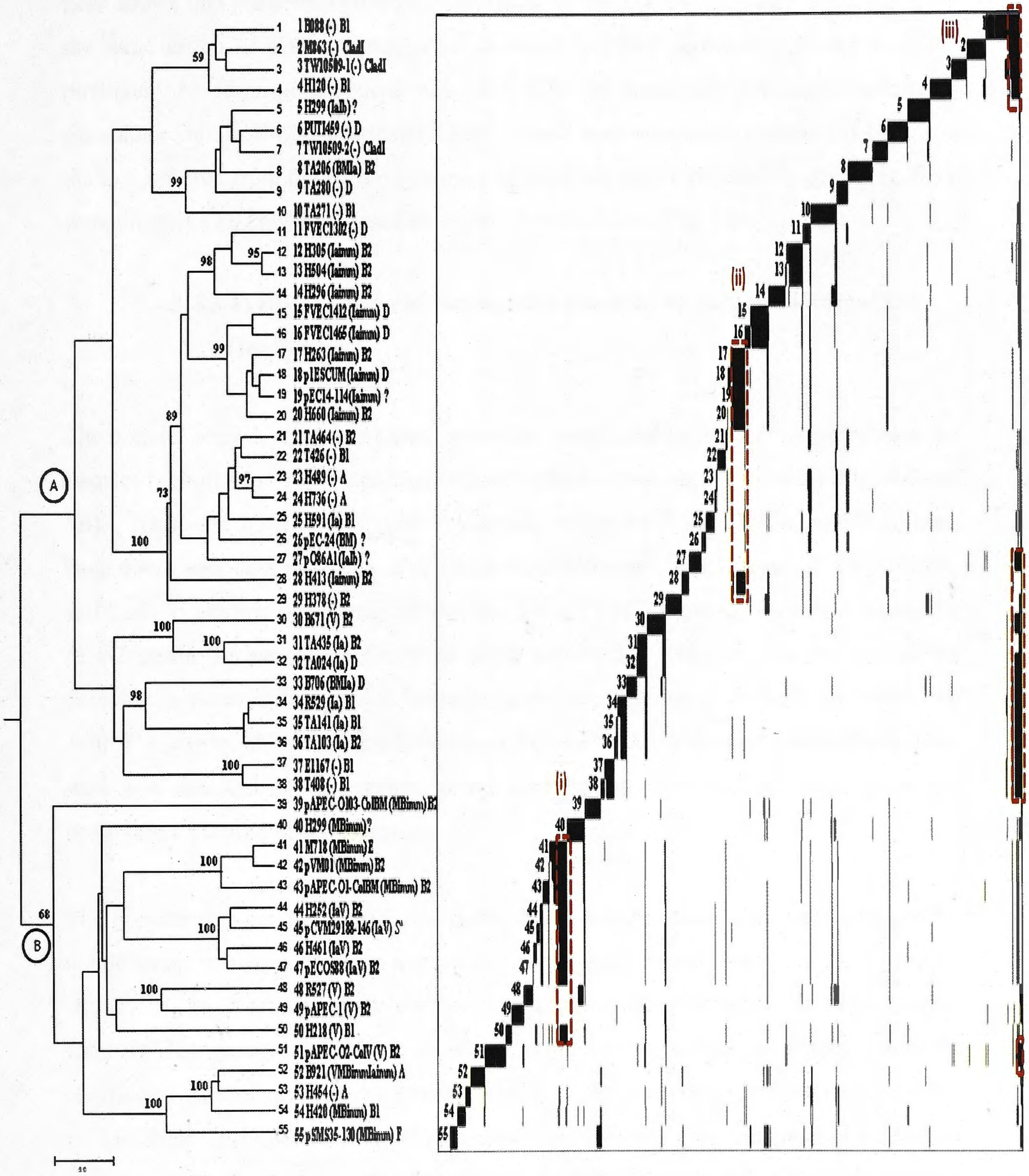


Figure 5. (cont.) (B) UPGMA tree of the plasmids inferred from the gene content using SplitTree4 and drawn with MEGA5. Bootstrap values are based on 500 replicates and bootstrap confidence values greater than 50% are listed to the left of the nodes. The figure to the right of the phylogeny depicts the variable genome of each plasmid. Each row of black bars corresponds to a plasmid in phylogeny. The horizontal width of the black bars corresponds to the size of the variable genome for each plasmid, and overlapping black bars corresponds to variable genome content shared by two or more plasmids. The red dashed lines indicate 3 groups of plasmids based on their vaiable genome content that relatively overlap among them (see Tables 7, 8, and 9).

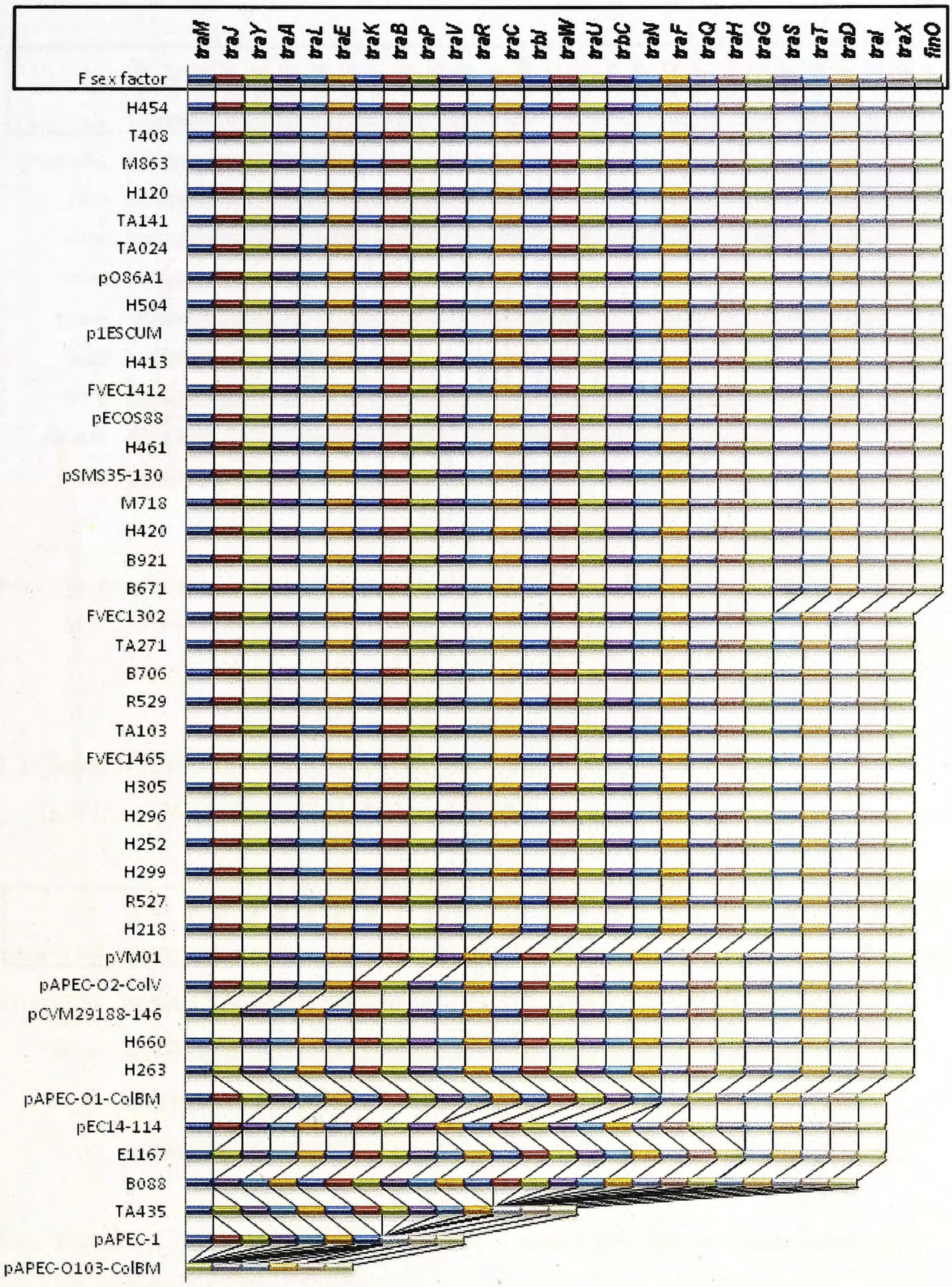
gene shows that plasmids that were clustered together on clusters A and B (Fig. 5A) are the same group of plasmids that were clustered together on clusters A and B of the phylogeny based on gene content data (Fig. 5B). Of these plasmids, only 3 plasmids: plasmids of strain H413, T408, and E1167, which were clustered on cluster B based on the tree inferred from the partial sequence of RepFIIA *repA1* replication gene (Fig. 5A), were clustered on cluster A based on the gene content tree (Fig. 5B).

4.3.3.3 Transfer region of conjugative plasmids in the RepFIB/RepFIIA (IncF) group

The transfer regions of most of these plasmids were found to share similarities with the transfer regions of *E. coli* F sex factor (GenBank accession no. AF112469) (Fig. 6A and 6B). While the transfer region of 4 plasmids including 3 RepFIIA plasmids (2 non-bacteriocin encoding plasmids of D strain PUTI459 and clade I strain of TW10509-2, and ColIaIb plasmid of D strain H299), and 1 RepFIB/FIIA plasmid (ColBMIa plasmid of B2 strain TA206) were found to share similarities with the transfer regions of *Salmonella* plasmid R64 which belongs to the IncI1 group (GenBank accession no. AP005147) (Fig. 6C). This result reveals the reason why genes involved in conjugation were not detected as core genes in the IncF group even though these genes are conjugative plasmid backbone genes.

The transfer regions of *E. coli* F sex factor (IncF) and *Salmonella* plasmid R64 (IncI1) are different in their gene order and nucleotide sequence of *tra* genes (e.g. TraP, -M, -L, -K, -J, -E, and -A). Overall, the *tra* gene clusters of some plasmids belonging to RepFIB/FIIA group are similar. However, some other plasmids were found to carry a different combination of *tra* genes. For example, the transfer regions of pAPEC-O103-ColBM, pAPEC-1, and ColIa plasmid of strain TA435 are truncated in a different manner in each plasmid. pAPEC-O103-ColBM contains *traGSTDIX*, pAPEC-1 contains *traMJYALEKIX* and pseudo *traB*, and ColIa plasmid of strain TA435 contains *traMYALEKBPVRCIX*, respectively (Fig. 6A).

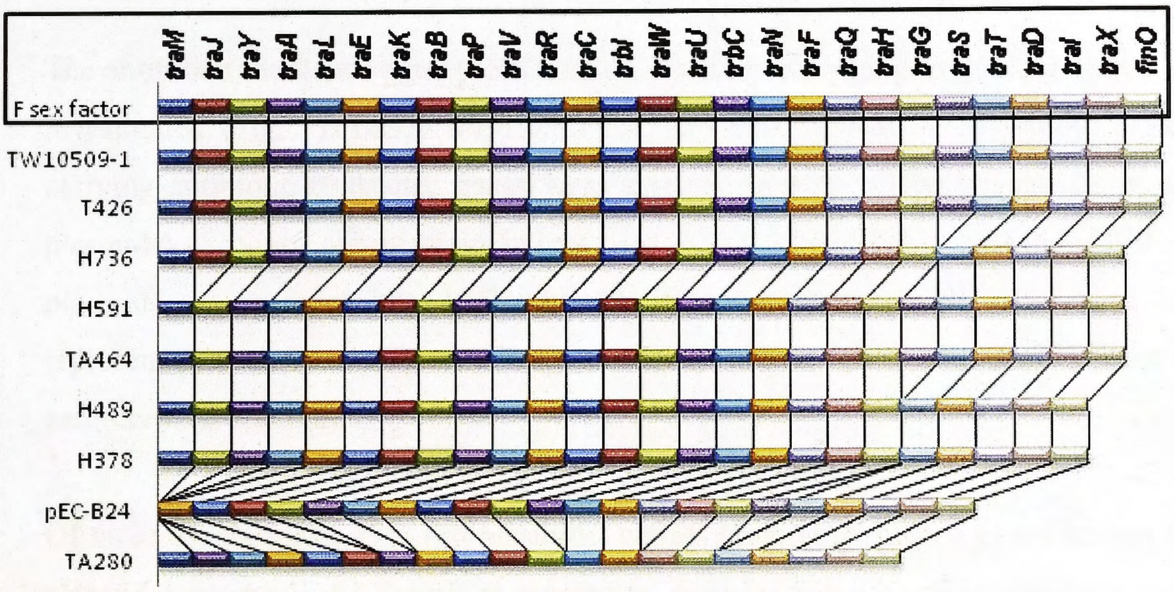
(A) RepFIB/RepFIIA plasmids



Note: The transfer gene region between *traM* to *finO* of *E. coli* F sex factor (GenBank accession no. AF112469) was used as the reference.

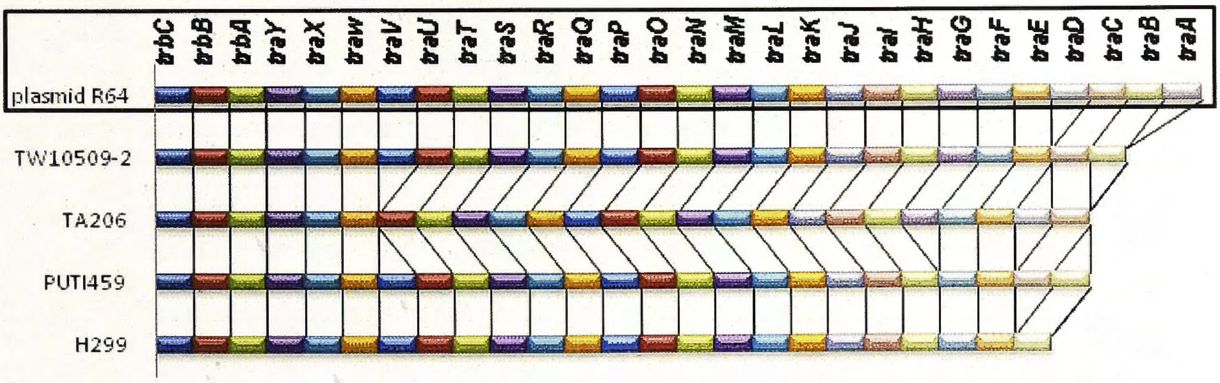
Figure 6. The transfer genes of plasmids belonging to RepFIB/RepFIIA (IncF) group. The top map is the *tra* operon of *E. coli* F sex factor transfer region. The black lines follow each colored block throughout the diagram to illustrate gene presence of each plasmid.

(B) RepFIIA plasmids



Note: The transfer gene region between *traM* to *finO* of *E. coli* F sex factor (GenBank accession no. AF112469) was used as the reference.

(C) Plasmids harboring plasmid R64 transfer region (RepFIB/RepFIIA: TA206, RepFIIA: TW10509-2, PUTI459, and H299)



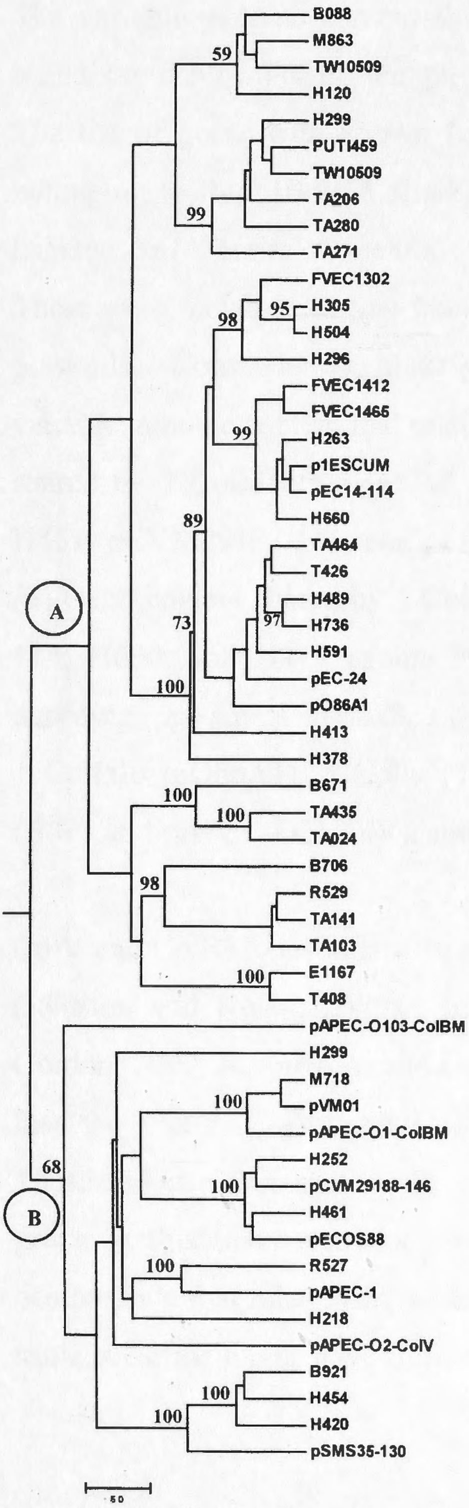
Note: The transfer gene region between *trbC* to *traA* of plasmid R64 (GenBank accession no. AP005147) was used as the reference sequences.

Figure 6. (cont.)

4.3.3.4 Antibiotic resistance and virulence-associated gene profiles

The antibiotic resistance gene profiles of 55 plasmids belonging to the IncF group were investigated (Fig. 7). Based on 6 antimicrobials investigated in this study, plasmids carrying antibiotic resistance genes were detected in 40% of the strains (22 ot of 55 plasmids). Nine plasmids including pAPEC-O103-ColBM, pSMS35_130, and plasmids of strains H299, PUTI459, FVEC1465, H489, H378, B706, and H218 represent plasmids carrying multidrug resistance (MDR) functions (resist to 3 out of 6 antimicrobial studied).

Of these 55 plasmids, 47% (26 plasmids) carries putative virulence genes known to be plasmid associated. As described previously, the phylogeny based on the gene content data of 55 plasmids belonging to RepFIB/RepFIIA (IncF) group separates all plasmids into 2 major clusters (Fig. 7). The result shows that almost plasmids of cluster A lack putative virulence genes known to be plasmid associated overall. In contrast, the putative virulence genes known to be plasmid associated were more common in plasmids of cluster B (Fig. 7). All plasmids which were clustered together on cluster B contain *hlyF* and *sitABCD.* Moreover, the result shows that the virulence gene profiles of ColV plasmids are similar except a plasmid of strain B671 (B2) which was clustered with plasmids belonging to cluster A.



	Phylo-group[a]	Bacteriocin	Transfer region (F/R)[e]	Resistance profile[b]	eitABCD	estABC	hylF	iroBCDEN	iucABCD	iutA	iss	sitABCD	tsh
B088	B1	-	F	C(c[c])	-	-	-	-	-	-	c	-	-
M863	CladeI	-	F		-	-	-	-	-	-	c	-	-
TW10509	CladeI	-	F		-	-	-	-	-	-	-	-	-
H120	B1	-	F		-	-	-	-	-	-	-	-	-
H299	?	IaIb	R	A, S, Tm, T, C(c)	-	-	-	-	-	-	c	-	-
PUTI459	D	-	R	S, Tm, T	-	-	-	-	c	c	c	c	-
TW10509	CladI	-	R	Tm	-	-	-	-	-	-	-	-	-
TA206	B2	BMIa	R	C(c)	-	-	-	-	c	c	-	-	-
TA280	D	-	F	C(c)	-	-	-	-	-	-	-	c	-
TA271	B1	-	F	C(c)	-	-	-	-	+	+	-	-	-
FVEC1302	D	-	F		-	-	-	-	c	c	c	c	-
H305	B2	Iaimm	F	C(c)	+	-	-	c	-	-	c	c	c
H504	B2	Iaimm	F		+	-	-	c	-	-	c	c	-
H296	B2	Iaimm	F	A, S, C(c)	+	-	-	-	-	-	c	c	-
FVEC1412	D	Iaimm	F	S, Tm	-	-	-	-	c	c	c	c	-
FVEC1465	D	Iaimm	F	A, S, Tm, T, C (p[d]/c)	-	-	-	-	-	-	-	c	-
H263	B2	Iaimm	F		-	-	-	-	-	-	c	c	-
p1ESCUM	D	Iaimm	F	T, C(remnant)	-	-	-	-	c	c	c	c	-
pEC14-114	?	Iaimm	F		-	-	-	-	-	-	-	-	-
H660	B2	Iaimm	F	A*	-	-	-	-	-	-	c	c	-
TA464	B2	-	F	C(c)	-	-	-	-	c	c	-	c	-
T426	B1	-	F	C(c)	-	-	-	-	-	-	c	-	-
H489	A	-	F	C(c)	-	-	-	-	-	-	-	c	-
H736	A	-	F	A, S, T	-	-	-	-	-	-	c	c	-
H591	B1	Ia	F	C(c)	-	-	-	-	-	-	-	-	-
pEC-24	?	BM	F	A	-	-	-	-	-	-	-	-	-
pO86A1	?	IaIb	F	K	-	-	-	-	-	-	-	-	-
H413	B2	Iaimm	F	C(c)	-	-	-	-	+	+	c	+/c	c
H378	B2	-	F	A, T, C(p/c)	-	-	-	c	c	c	c	c	-
B671	B2	V	F		-	+	-	+	-	-	+	c	+
TA435	B2	Ia	F		-	+	-	+	-	-	+/c	c	-
TA024	D	Ia	F		-	+	-	+	-	-	+/c	c	-
B706	D	BMIa	F	S, Tm, T	+	+	-	-	c	c	c	c	-
R529	B1	Ia	F		-	-	-	-	-	-	c	c	-
TA141	B1	Ia	F		-	-	-	-	-	-	-	-	-
TA103	B2	Ia	F		-	-	-	-	-	-	c	c	-
E1167	BI	-	F		-	-	-	-	-	-	c	-	-
T408	B1	-	F		-	-	-	-	-	-	c	-	-
pAPEC-O103-ColBM	B2	MBimm	F	S, T, Tm	-	-	+	+	+	+	+	+	-
H299	?	MBimm	F	T, C(c)	-	+	+	+	+	+	+/c	+	-
M718	E	MBimm	F	T, C(c)	-	+	+	+	+	+	+/c	+	-
pVM01	B2	MBimm	F		-	+	+	+	+	+	+	+	+
pAPEC-O1-ColBM	B2	MBimm	F		+	+	+	+	+	+	+/c	+/c	+
H252	B2	IaV	F		-	+	+	+/c	+	+	+/c	+/c	-
pCVM29188-146	*S*	IaV	F	S, T	-	+	+	+	+	+	+	+	-
H461	B2	IaV	F	A, T, C(c)	-	+	+	+	+	+	+/c	+/c	-
pECOS88	B2	IaV	F		-	+	+	+	+	+	+/c	+/c	-
R527	B2	V	F	C(c)	-	+	+	+	+	+	+/c	+/c	-
pAPEC-1	B2	V	F		-	+	+	+	+	+	+	+	+
H218	B1	V	F	A, S, T	-	+	+	+	+	+	+/c	+	-
pAPEC-O2-ColV	B2	V	F		+	+	+	+	+	+	+	+	+
B921	A	VMBimmIaimm	F	Tm	-	-	+	+	-	-	+	+	-
H454	A	-	F	C(c)	-	-	+	+	-	-	+/c	+	-
H420	B1	MBimm	F	A, S, C(c)	-	-	+	+	c	c	+/c	c	-
pSMS35-130	F	MBimm	F	A, K, S, Tm, T, C	-	-	+	-	c	c	-	+/c	-

[a] ?, unknown phylogroup; *S*, *Salmonella enteric* serovar Kentucky

[b] A, ampicillin; K, kanamycin; S, streptomycin; Tm, trimethoprim; T, tetracycline; C, chloramphenicol.

[c] c, chromosome; [d] p, plasmid

[e] F, *E. coli* F sex factor transfer region; R, *Salmonella enteric* serovar Typhimurium plasmid R64

* located on another plasmid of the strain; +, present on the plasmid; -, totally absent; +/c, present on the plasmid and chromosome

Figure 7. Genotypic characteristics of 55 plasmids in the RepFIB/RepFIIA (IncF) group in *E. coli*.

4.3.3.5 Genes of conjugative plasmids in the RepFIB/RepFIIA (IncF) group

The variable genome content shared by two or more plasmids was also investigated based on the results of the phylogeny inferred from gene content data (Fig. 5B). The list of genes with known functions on conserved segments sharing by plasmids belonging to RepFIB/FIIA (IncF) group was identified. For most IncF plasmids, the balance of entire genome consists of genes unique to each plasmid. There were, however, a few fractions of variable genomes share among multiple IncF plasmids. Consequently, these plasmids were divided into 3 groups based on their variable genome content that relatively overlap among them (Fig. 5B). (i) Gene content shared by 3 ColMBimm (M718, pVM01, and pAPEC-O1-ColBM), 4 ColIaV (H252, H461, pCVM29188-146, and pECOS88), and ColV plasmid of strain H218 (Table 7). (ii) Gene content shared by 5 ColIaimm plasmids including H263, p1ESCUM, pEC14-114, H660, and H413 (Table 8). (iii) Gene content shared by 6 non-bacteriocin encoding plasmids (B088, M863, TW10509-1, H120, E1167, and T408), 1 ColIaIb (pO86A1), 5 ColIa (TA435, TA024, R529, TA141, and TA103), 2 ColV (B671 and pAPEC-O2-ColV), and 1 ColBMIa (B706) (Table 9).

ColV and ColBM plasmids were reported to have evolve from a RepFIB/FIIA ancestral (Johnson and Nolan, 2009). In agreement with previous studies (Christenson and Gordon, 2009, Jeziorowski and Gordon, 2007), the results in this study also demonstrate that the ColIa, ColV, and ColBM plasmids were members of the IncFII group. These indicate that bacteriocin encoding plasmids belonging to RepFIB/FIIA (IncF) group in this study shared a common RepFIIA ancestor. The results of this study demonstrate that relationships based on gene content data indicate that plasmids of the same bacteriocin type have similar gene content.

Table 7. List of genes with known functions on conserved segments sharing by conjugative plasmids in the group (i): M718, pVM01, pAPEC-O1-ColBM, H252, pCVM29188- 146, H461, pECOS88, and H218 (Fig. 5B).

Length of segments (nt)	Gene or Function (total length; nt)
1858	ABC transporter (1,941)
312	efflux transporter (1,182)
1371	outer membrane efflux protein (1,371)
648	*iroB* (1,116)
253	ABC transporter (3,570)
793	IroE protein (957)
2152	TonB-dependent siderophore receptor protein (2,169)
278	endopeptidase (278)
294	bor protein (294)
198	YacA protein (198)
123	YacB protein (183)
294	YbaA protein (294)
303	YdeA protein (303)
129	colicin-V (129)
1304	colicin V secretion/processing ATP-binding protein CvaB (2,097)
774	colicin V secretion protein cvaA
318	colicin V secretion protein cvaA
477	hypothetical protein (Gene family: yubP protein) (822)
246	hypothetical protein (Gene family: yubO protein) (288)
210	hypothetical protein (Gene family: YchA-1) (255)
635	hypothetical protein (635)
327	transposase (327)
191	transposase (228)
186	transposase (378)
317	transposase (336)
699	integrase core domain-containing protein (699)
404	omptin family protein (954)
359	CobQ/CobB/MinD/ParA nucleotide binding domain-containing protein (Gene family: SopA) (1,167)
179	dksA/traR C4-type zinc finger (222)
168	sok (225)
121	hok/gef family protein (159)

Table 8. List of genes with known functions on conserved segments sharing by conjugative plasmids in the group (ii): 5 ColIaimm plasmids, including H263, p1ESCUM, pEC14-114, H660, and H413 (Fig. 5B).

Length of segments (nt)	Gene or Function (total length; nt)
357	hypothetical protein (Gene family: YdeA protein) (510)
173	hypothetical protein (Gene family: YbaA protein) (348)
1284	hypothetical protein (Gene family: ABC transporter) (1,284)
1131	hypothetical protein (Gene family: cell division protein FtsX) (1,131)
735	hypothetical protein; Gene family: conserved hypothetical protein (792)
486	hypothetical protein; Gene family: conserved hypothetical protein (486)
243	hypothetical protein; Gene family: conserved hypothetical protein (243)
866	hypothetical protein; Gene family: conserved hypothetical protein (1,170)
267	transposase (366)
267	transposase (267)
163	transposase (327)
228	transposase (327)
250	transposase (348)
347	transposase (405)
519	IS66 family protein element (519)
258	IS66 family protein element (258)
211	TonB family protein domain-containing protein (777)
2262	TonB dependent receptor protein (2,262)
1971	TonB-dependent heme/hemoglobin receptor family protein (1,971)
1176	senB protein (Gene family: enterotoxin TieB protein) (1,176)
201	IstB ATP binding protein (330)
696	ABC transporter (696)
486	redoxin (486)
1380	S domain-containing protein (1,380)
450	glucose-1-phosphatase (450)
321	glucose-1-phosphatase (321)
1548	iron permease FTR1 family protein (1,887)
591	transglycosylase SLT domain-containing protein (591)
1053	reverse transcriptase (1,458)
357	integrase core domain-containing protein (357)
300	hok/gef family protein (300)
219	post-segregation antitoxin protein CcdA (219)
327	CcdB protein (327)
184	alpha/beta hydrolase (861)
261	replication regulatory protein RepB (261)
377	phage integrase (741)
978	initiator Replication protein (978)

Table 9. List of genes with known functions on conserved segments sharing by conjugative plasmids in the group (iii): B088, M863, TW10509-1, H120, pO86A1, B671, TA435, TA024, B706, R529, TA141, TA103, E1167, T408, and pAPEC-O2-ColV (Fig. 5B).

Length of segments (nt)	Gene or Function (total length; nt)
1,939	hypothetical protein; Gene family: conserved hypothetical protein (4,197)
655	hypothetical protein; Gene family: conserved hypothetical protein (1,362)
147	hypothetical protein; Gene family: conserved hypothetical protein (345)
955	hypothetical protein; Gene family: conserved hypothetical protein (1,965)
155	hypothetical protein; Gene family: conserved hypothetical protein (261)
249	hypothetical protein; Gene family: conserved hypothetical protein (249)
110	hypothetical protein; Gene family: conserved hypothetical protein (1,362)
117	hypothetical protein; Gene family: conserved hypothetical protein (261)
104	hypothetical protein; Gene family: conserved hypothetical protein (225)
113	hypothetical protein; Gene family: conserved hypothetical protein (504)
657	hypothetical protein; Gene family: conserved hypothetical protein (936)
218	hypothetical protein; Gene family: conserved hypothetical protein (243)
197	hypothetical protein; Gene family: conserved hypothetical protein (231)
235	hypothetical protein; Gene family: conserved hypothetical protein (516)
597	lytic transglycosylase (603)
195	*mok*; modulator of post-segregation killing protein (213)
252	InsL (1,119)
640	*psiA*; plasmid SOS inhibition protein A (720)
251	plasmid-partitioning protein (972)
544	single-stranded DNA-binding protein (567)
240	plasmid SOS inhibition protein B (438)
1,141	*sopA*; plasmid-partitioning protein SopA (1,167)
441	adenine-specific DNA methylase (564)

4.4 Discussion and Conclusion

4.4.1 Conjugative Plasmids in IncI1 and IncF Groups

There are a large number of plasmids types known to occur among *E. coli* strains and these play an important role in bacterial adaptation (Frost et al., 2005). The results of this study demonstrate that the majority of plasmids in *E. coli* belongs to the RepFIB/RepFIIA (IncF) backbone types, while some of them belong to the RepI1 (IncI1) group. Conjugative plasmids belonging to the RepI1 (IncI1) group were found to be more homogeneous compared to those of the RepFIB/FIIA (IncF) plasmid group. By considering core components of plasmids: replication genes, stability genes, and the transfer region, IncI1 plasmids were found to be conserved, as they share their common core genome. By contrast, there is no core genome share by IncF plasmids. Instead, for most IncF plasmids, the balance of entire genome consists of genes unique to each plasmid.

In addition, by considering types of bacteriocin-encoding plasmids, almost all of plasmids belonging to the RepI1 group are ColIb or ColIa plasmids (except one ColIbMBimm plasmid of strain B367). By contrast, plasmids belonging to RepFIB/FIIA group comprised various types of bacteriocin-encoding plasmids. These lead to much more variable gene content found among plasmids belonging to the RepFIB/FIIA (IncF) group. Due to the plasticity of plasmids, the more plasmids included in the analysis, the more diversity in plasmids will be detected. This diversity of plasmids is what influences the core genome that will be shared by all plasmids.

4.4.2 Evolution of Colicin Ib and Ia Plasmids in the RepI1 (IncI1) Group in *E. coli*

A few completed *E. coli* plasmid sequences belonging to the RepI1 group are available and most of available RepI1 plasmids are ColIb plasmids from *Shigella* and *Salmonella*. In addition to these RepI1-ColIb plasmids in databases, the results in this study reveal the diversity of *E. coli* plasmids found in the RepI1 group as ColIa and coassociate-bacteriocin encoding plasmids (ColIbMBimm plasmid of strain B367) were detected to be a member of the RepI1 group. Many plasmids lack the colicin genes but still possess the RepI1 backbone were also detected. Among *E. coli* plasmids belonging to the RepI1 group, 2 plasmids: ColIb of pO113 and ColIbMBimm of strain B367 were found to be a hybrid RepI1/FIB plasmid. The balance of the entire genome of these 2 plasmids consists of genes largely unique to each plasmid. The pO113, a human ETEC plasmid encoding only colicin Ib, is one of plasmids divergent from the RepI1-ColIb plasmids. Noticeably, pO113 (RepI1/FIB plasmid) was clustered with 3 non-bacteriocin encoding plasmids (RepI1 plasmids) from human B2 strains on cluster B (Fig. 1A and 1B). These might suggest the evolutionary divergence of these plasmids from a common ancestor.

Among plasmids in the RepI1 group in this study, only RepI1/FIB-ColIbMBimm plasmid of strain B367 (phylo-group D) was found to carry a putative virulence cluster known to be plasmid associated: *hlyF*, *iroNEDCB*, *iss*, and *sitABCD*. These putative virulence genes are genes known to be associated with colicin Ia, V, B, and M plasmids except *hlyF* that is associated with colicin B and M plasmids (Christenson and Gordon, 2009, Jeziorowski and Gordon, 2007). Previous study has shown that the colicin Ia and microcin V coassociation involves the movement of the microcin V operon together with the *iroNEDCB* and *iss* genes on to *traT*- positive ColIa plasmids (Jeziorowski and Gordon, 2007). However, in addition to colicin Ia, microcin V has been found together with colicins Ib, B, and M encoded on conjugative plasmids in *E. coli* (Gordon et al., 2007). Therefore, the colicin Ib and MBimm coassociation of the plasmid from strain B367 might involve the acquisition of gene normally found on ColV plasmids as a remnant of microcin V operon (no *cvaA* and *cvaB*) was detected on this plasmid (Table 2). .In addition, that the colicin Ia and colicin Ib genes have remarkably physical and functional similarities (Cascales et al., 2007) might imply that the evolution of ColIa and ColIb plasmids with the coassociation of other bacteriocins might evolve in the similar manner. However, as only one representative of the ColIbMBimm plasmid possessing the RepI1 backbone found in this study, the coassociation of these bacteriocins still unclear.

4.4.3 Evolution of Bacteriocin Plasmids in the RepFIB and RepFIIA (IncF) Groups in *E. coli*

The most common colicins produced by *E. coli* are colicins B and M (ColBM) usually encoded adjacently on the same plasmids. However, the result in this study reveals the coassociation of colicins B, M, and Ia (ColBMIa) found in *E. coli* RepFIB/FIIA plasmids of strains TA206 (phylo-group B2) and B706 (phylo-group D). The diversity among these two ColBMIa plasmids was also found as they possess different transfer regions. The transfer region of ColBMIa plasmid of B2 strain TA206 shared similarity with the transfer region of *Salmonella* plasmid R64 (IncI1), while ColBMIa plasmid of D strain B706 shared similarity with the *E. coli* F sex factor transfer region (IncF). Due to only one representative ColBMIa found in each transfer region type, it is not possible to determine how ColBMIa plasmids evolved. However, the ColBMIa plasmid of D strain B706 appears to be closely related to ColIa plasmids as they were clustered together on the tree (Fig. 5). These might suggest that ColBMIa and ColIa plasmids shared a common ancestor.

By considering plasmids encoding only one bacteriocin type, ColV plasmids were very homogenous as all are members of the IncF group and possessed *E. coli* F sex factor transfer region. By contrast, ColIa plasmids were diverse as they were found to be members of both IncF and IncI1 groups. Plasmids encode only colicin Ia operon were found to have a variable pattern of VFs as they were isolated from different hosts and phylo-groups. This suggests that host specificity and phylo-groups of *E. coli* strains are influence on a variety of VF profiles of plasmids carried by those strains. ColIa plasmids belonging to animal B2 strain (TA435) and animal D strain (TA024) except animal B2 strain (TA103) found to carry several VFs responsible for ExPEC virulence including *etsABCD*, *iroBCDEN*, *iss*, and *traT* (*traT*: only found in TA024). By contrast, ColIa plasmids belonging to human B1 strain (H591) and two animal B1 strains (R529 and TA141) contain no investigated VFs except *traT*. This finding indicates that ColIa plasmids of *E. coli* belonging to phylogroups B2 and D in this study represent additional colicin-encoding plasmids responsible for ExPEC virulence in addition to ColV plasmids known to have long been associated with ExPEC virulence (Smith and Huggins, 1976).

In addition to plasmids encoding the colicin Ia operon, the ColIaimm plasmids carrying a colicin Ia immunity gene but lacking a colicin Ia toxin gene (1,881 bp inlength) were detected. ColIaimm plasmids of *E. coli* strains belonging to phylo-groups B2 and D are monophyletic except ColIaimm plasmid of human B2 strain (H413) as its VF profile differs from the others (Fig. 5 and 7). The detection of VFs on ColIaimm plasmids was infrequent but antimicrobial resistance genes investigated were rather detected. How ColIaimm plasmids evolved is still unknown. However, it is well known that most natural *E. coli* isolates are capable of resisting to most colicins (Riley and Gordon, 1992, Gordon et al., 1998). Therefore, these results propose that lacking the colicin Ia toxin gene might not affect on the competitive interactions among strains but strains carrying this type of plasmid have rather made use from MDR functions under specific circumstance. However, further investigation needs to be conducted to truly understand these plasmids.

Microcin V has been reported to be found together with colicins Ia, Ib, B, and M encoded on conjugative plasmids in *E. coli* (Gordon et al., 2007). As expected, phylogeny based on gene content data clustered several bacteriocin-encoding plasmids including, ColV, ColIaV, ColMBimm, and ColVMBimmIaimm plasmids together on the clade B (Fig. 5). In case of ColIaV plasmids in this study, three ColIaV plasmids of *E.coli* (pECOS88, H252, and H461) and one ColIaV plasmid of *Sallmonella* (pCVM29188-146) were found to encode intact colicin Ia and intact microcin V operons and also virulence factors normally associated with ColV plasmids. This result is consistent with the hypothesis that ColIaV plasmids have evolved as a consequence of the movement of an intact microcin V operon and associated virulence factors (*iss* and *iroBCDEN* always present) onto *traT*-positive ColIa plasmids encoding an intact colicin Ia operon (Jeziorowski and Gordon, 2007).

In case of ColMBimm plasmids, these plasmids possess an identical colicin gene structure, an F-type *traY* gene and have IncFII-related replicons (Christenson and Gordon, 2009). The ColMBimm plasmids in this study form two distinct subgroups, (i) pAPEC-O103-ColBM, H299, M718, pVM01, and pAPEC-O1-ColBM, and (ii) B921, H420, and pSMS35-130 on clade B (Fig. 7). The patterns of VFs of these plasmids also indicate that two distinct subgroups had the different pattern of the virulence cluster associated with ColV plasmids. The ColMBimm plasmids of strains

H299 and M718, and pAPEC-O1-ColBM were previously reported to have a truncated B activity gene called type I ColMBimm plasmid (Christenson and Gordon, 2009). Previous study has shown that the evolution of ColMBimm plasmids appears to have involved as a consequence of gene transfer between colicin BM and microcin V plasmids (Christenson and Gordon, 2009). They also suggest that transfer events involve the acquisition of genes associated with ColV plasmids onto ColBM plasmids, which has resulted in the loss of colicin B and microcin V activity and immunity. In agreement with the results of previous study (Christenson and Gordon, 2009), the result in this study reveals that the size of ColBM plasmid (pEC-B24) was smaller than ColMBimm plasmids. This indicates the transfer of the microcin V plasmid onto the ColBM plasmid, which has resulted in the gain of virulence traits associated with ColV plasmids and the loss of of colicin B and microcin V activity and immunity. The pattern of putatively plasmid-borne virulence traits also emphasized the differences between ColBM and ColMBimm plasmids, as pEC-B24 contains no virulence traits and was located on the different clade from ColMBimm plasmids (Fig. 7). In this study, plasmids from APEC strains including pAPEC-O103-ColBM and pVM01 were found to have a truncated B activity gene like those plasmids and were also found to contain a putative virulence cluster associated with ColV plasmids, as well as remmants of the microcin V operon.

ColIa, ColV, and ColBM plasmids were members of the IncFII group (Christenson and Gordon, 2009, Jeziorowski and Gordon, 2007). In this study, bacteriocin encoding plasmids belonging to RepFIB/FIIA group (RepFIIA and RepFIB/RepFIIA plasmids) shared a common RepFIIA ancestor. This indicates that the genomic diversity among these plasmids have evolved from a single plasmid backbone type, RepFIIA, as plasmids containing only RepFIB were not detected. However, in essence there is no core genome share by RepFIB/FIIA plasmid and RepFIIA plasmid. Although these plasmids in the IncF group can be grouped into 2 major clusters based on gene content data, overall, there are only small fractions of gene content share by IncF plasmids belonging to each cluster. Instead, for most IncF plasmids, the balance of entire genome consists of genes unique to each plasmid.

4.4.4 The Conjugative Ability of Plasmids in the IncF Group

The ability to conjugate of plasmids in the IncF group was determined by genes known as *tra* operon of *E. coli* F sex factor transfer region. Twenty-four transfer genes (*tra*) have been identified as transfer genes known to be involved in conjugation. However previous study has showed that plasmids lacking *traB* gene or carry truncated *traB* gene are unlikely conjugative as *traB* is an essential gene for conjugative transfer (Rump et al., 2012, Kim et al., 1993).

In this study, the result suggests that there is no *tra* gene cluster in common to all 55 plasmids in the IncF group. This means that in essence IncF plasmids carry a different combination of the *tra* genes. For example, the transfer regions of ColIa plasmid of strain TA435, pAPEC-O103-ColBM, and pAPEC-1, and are found to be truncated in a different manner in each plasmid. These plasmids lack a significant number of *tra* genes which might suggest that they are probably incapable to transfer. However ColIa plasmid of strain TA435 is still conjugative as it still carries *traB* gene (*traMYALEKBPVRCIX*). By contrast, pAPEC-O103-ColBM and pAPEC-1 seem not to be able to transfer as *traB* is missing in pAPEC-O103-ColBM (contains *traGSTDIX*) (Johnson et al., 2010) and truncated in pAPEC-1 (contains *traMJYALEKIX* and pseudo *traB*) (Mellata et al., 2009), respectively. How do these plasmids still persist in *E. coli* populations without the conjugative ability?

There are two ways for nontransmissible plasmids to persist in *E. coli* populations. One possibility is that plasmids may be becoming secondary chromosomes (Smillie et al., 2010). The other possibility is that these plasmids are hitchhiking alongside *E. coli* to confer a significant fitness advantage of their hosts. It is well known that pAPEC-O103-ColBM (ColMBimm) and pAPEC-1 (ColV) are virulence plasmids associated with ExPEC virulence (Johnson and Nolan, 2009). The virulence genes encoded by these plasmids probably contribute to the pathogenicity of extra-intestinal pathogenic *E. coli* (ExPEC) strains to cause extra-intestinal infections. This might suggest that even these plasmids lost the ability of conjugation; however, they still persist in *E. coli* populations by hitchhiking alongside to confer the selective advantage in particular the pathogenic potency of their hosts. Not only ExPEC plasmids but other virulence plasmids causing enterohemorrhagic *E. coli* (EHEC) of serotype O157:H7 have been reported to be potentially non-conjugative plasmids (Rump et al., 2012). These suggest that these particular plasmids may persist in population for long period of time by contributing the significantly useful genes for their host under natural selection. Interestingly, previous study has also suggested that about half of the plasmids are in fact nontransmissible (Smillie et al., 2010).

4.4.5 *E. coli* plasmids and their role in *E. coli* ecology

Large conjugative plasmids of *E. coli* originally encoded bacteriocins thought to mediate competitive interactions among strains. However, in human and animal strains, bacteriocin-encoding genes on conjugative plasmids have been loss and replaced by virulence-associated traits. Evidence suggests that most natural *E. coli* isolates are capable of resisting most bacteriocins (Gordon et al., 1998, Riley and Gordon, 1992). Therefore losing bacteriocin production might not affect on the competitive interactions among strains. On the other hand, the acquisition of virulence traits may be conferring a fitness advantage to their hosts. These might suggest that bacteriocin plasmids are changing their role from conferring the competitive advantage in microbial populations to conferring the selective advantage particularly for the pathogenic potency in relation to their hosts.

In conclusion, conjugative plasmids found in *E. coli* are very diverse. The majority of plasmids in *E. coli* belongs to the RepFIB or RepFIIA (IncF) backbone types, while some of them belong to the RepI1 (IncI1) group. Comparing between IncI1 and IncF plasmids, plasmids in the IncI1 group are more homogeneous. In this study, particularly for the large group of conjugative plasmids in the IncF group, the data strongly suggest that there are no such things as plasmid species. As in essence there is no core genome shared by IncF plasmids. In addition, overall for most plasmids the balance of the genome consists of genes that are unique each plasmid. Conjugative plasmids: key agents in the adaptation of *E. coli* populations have changed their role as mediators of intra- and interspecies interactions to become associated with *E. coli* virulence. Furthermore, plasmids derived from this study are broadly representative of the ancestral bacteriocin plasmids and the coassociation of bacteriocins found in *E. coli* species.

CHAPTER 5

Genetic and metabolic characteristics of phylo-group F

5.1 Introduction

E. coli is the best-known species of Enterobacteriaceae and represents the most numerous facultative anaerobe presenting in the lower intestinal tract of birds and mammals. The species includes both commensal strains with little ability to cause disease and pathogenic strains that are able to cause intestinal or extra-intestinal infections. Due to a highly genetically diversity of *E. coli*, there is the existence of extensive substructure within the species.

Based on methods such as multi-locus enzyme electrophoresis (MLEE) and multilocus sequence typing (MLST), *E. coli* can be subdivided into 5 main phylo-groups known as A, B1, B2, D, and E (Gordon et al., 2008). Although *E. coli* strains have phylogenetic cohesiveness, however among strains of the various phylo-groups, they differ in their phenotypic characteristics, genome size, and propensity to cause intestinal and extra-intestinal infections. Strains of the different phylo-groups are also associated with certain ecological niches and life-history characteristics. Among different phylo-groups, phylo-group B2 strains are most commonly isolated from mammalian hosts possessing hindgut modifications for microbial fermentation (Gordon and Cowling, 2003). Phylo-group B2 strains have been shown to persist as a resident strain than a transient strain than are strains of the other phylo-groups (Nowrouzian et al., 2006). In addition, in term of propensity to cause disease, phylo-group B2 and to a lesser extent phylo-group D strains are most likely to be responsible for extra-intestinal infections in humans (Jaureguy et al., 2008, Johnson and Russo, 2002). However, due to the diversity of the strains and the growing body of multi-locus sequence data and genome data for *E. coli*, the additional phylo-groups have been recently delineated. Based on the recent method: the extended quadruplex PCR phylo-group assignment, *E. coli* are now assigned to 8 phylo-groups including A, B1, B2, C, D, E, F, and *Escherichia* cryptic clade I (Clermont et al., 2013). The phylo-group F recently described as an additional group of *E. coli* strains has been proposed for a sister group to phylo-group B2 which known to be responsible for extra-intestinal infection (Jaureguy et al., 2008, Clermont et al., 2011). This makes phylo-group F strains interesting as the strains might have a propensity to cause diseases.

Previous studies have described phylo-group F as a group of strains closely related to the phylo-group B2. The phylogenetic relationships based on MLST analysis using 8 housekeeping genes (*dinB*, *icdA*, *pabB*, *polB*, *putP*, *trpA*, *trpB*, and *uidA*; 4,095 nt in total) showed that strains of the group (later assigned to phylo-group F) were sharply clustered separately from D and B2 strains and these strains seem to be more closely related to phylo-group B2 (Jaureguy et al., 2008). Phylogenetic relationships among *E. coli*, *E. fergusonii*, and *E. albertii* based on a set of 2,173 *E. coli* K-12 genes conserved in all strains under study also showed that two *E. coli* strains belonging to phylo-group F: SMS-3-5 and IAI39 were clustered with many extraintestinal pathogenic *E. coli* strains belonging to phylo-group B2 (Chaudhuri et al., 2010). However, *E. coli* IAI39 and *E. coli* SMS-3-5 currently belonging to phylo-group F were previously assigned to phylo-group D (Fricke et al., 2008, Jaureguy et al., 2008, Touchon et al., 2009). In addition, the MLST analysis based on the partial nucleotide sequence of 13 housekeeping genes (*aes*, *icd*, *pabB*, *polB*, *putP*, *trpA*, *trpB*, *adk*, *fumC*, *gyrB*, *mdh*, *purA* and *recA*; 9,819 nt in total) clustered phylo-group F strains on the same clade with phylo-group B2 strains (Clermont et al., 2013). Although, the previous phylogenetic relationships based on gene sequence analysis: MLST of 8 housekeeping genes (Jaureguy et al., 2008, Clermont et al., 2011), the set of 2,173 *E. coli* K-12 genes (Chaudhuri et al., 2010), and MLST of 13 housekeeping genes (Clermont et al., 2013) revealed that phylo-group F strains are more closely related to phylo-group B2 strains. However, there is some controversy as the evolutionary tree based on the metabolic distance demonstrated that metabolic networks of phlo-group F strains are closely related to phylo-group D which also known to be responsible for extra-intestinal infection (Vieira et al., 2011).

As it remains elusive on what phylo-group F is and whether F strains are more closely related to D or B2 strains. Therefore, to gain insights into the additional phylo-group F, the aim of this study was to characterize the phylogenetic relationships and metabolic diversity among the *E. coli* strains belonging to phylo-group F, D, and B2 based on whole genome-scale analysis. The newly environmental strain E1227 was sequenced as part of this study for inclusion in the analysis. The outcome of this study will be broadly representative of the strains belonging to phylo-group F to be found in *E. coli* as there is a few number of F strains currently available in the database. The study will also lead to significant advances in our understanding of phylogenetic diversity presented in this species.

5.2 Materials and Methods

5.2.1 Bacterial strains

The genome sequences of 23 *E. coli* strains were use for comparative genome analysis. The 23 *E. coli* strains investigated comprise 22 *E. coli* reference strains (complete genomes) from GenBank database at the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/) and the Broad Institute (draft genomes) (http://www.broadinstitute.org/), and the environmental *E. coli* E1227 (draft genomes in this study). The main characteristics of these strains were presented in Table 1. The genome sequences of *E. coli* strains were retrieved from NCBI and from the Broad Institute.

The environmental-source strain E1227 was genome sequenced and determined its phylo-group of *E. coli* as part of this study. The environmental strain E1227 was previously isolated from a water sample (Molonglo Sewage Treatment Plant, ACT) in Australia. The strain was found to represent the typical *E. coli* biochemical profile. However, due to the source of isolation and the phenotypic characteristic of the strain, the environmental strain E1227 was firstly suspected to be a membership of one of five novel *Escherichia* clades (C-I to C-V) that were recovered primarily from environmental sources and indistinguishable from typical *E. coli* based on traditional phenotypic tests (Walk et al., 2009). However, MLST analysis based on partial sequences of 22 housekeeping genes (*adk, aspC, torC, icdA, kdsA, fumC, fadD, metG, lysP, grpE, recA, mutS, rpoS, dnaG, mdh, aroE, mtlD, gyrB. cyaA, purA, arcA,* and *clpX*) revealed that the environmental strain E1227 is *E. coli.* The study by Luo and colleagues (2011) reported that the *E. coli* strains that were recovered primarily from environmental sources encoded all genes required for classification as typical *E. coli* (commensal or pathogenic). This means that the phenotype and taxonomy of the environmentally adapted *E. coli* lineages are indistinguishable from typical *E. coli* based on traditional phenotypic tests (Luo et al., 2011).

Table 1. Principal characteristics of 23 *Escherichia coli* genomes used in this study.

Strains	**Size (Mb)**	**Phylogenetic Group**	**Pathotype (1)**	**Host/ Source**	**Sample**	**Reference (2)**
IAI39	5.13	F	Pyeloneprhitis (ExPEC/ UPEC)	Human	Urine	NCBI
SMS-3-5	5.07	F	None	Environment	Sediments of Shipyard Creek	NCBI
MS21-1	5.28	F	Unknown	Human	Gut	NCBI
E1227	**5.20**	**F**	**Unknown**	**Environment**	**Water**	**This study**
B093	5.21	F	Unknown	Bird		Broad Institute
H299	5.32	unknown	Commensal	Human		Broad Institute
UMN026	5.20	D	Cystitis (ExPEC/ UPEC)	Human	Urine	NCBI
042	5.24	D	EAEC	Human	Unknown	NCBI
B706	6.13	D	Unknown	Bird		Broad Institute
B354	4.85	D	Unknown	Bird		Broad Institute
TA249	5.25	D	Unknown	Mammal		Broad Institute
TA255	4.91	D	Unknown	Mammal		Broad Institute
TA280	5.30	D	Unknown	Mammal		Broad Institute
M114	5.43	D	Unknown	Mammal		Broad Institute
CFT073	5.23	B2	Pyeloneprhitis (ExPEC/ UPEC)	Human	Blood	NCBI
ED1a	5.21	B2	Commensal	Human	Faeces	NCBI
S88	5.03	B2	Newborn meningitis (ExPEC/ NMEC)	Human	Cerebrospinal fluid	NCBI
536	4.94	B2	Pyeloneprhitis (ExPEC)	Human	Urine	NCBI
APEC-O1	5.08	B2	Colisepticemia (ExPEC/ APEC)	Chicken	Lung	NCBI
H001	5.05	B2	Unknown	Human		Broad Institute
H223	5.13	B2	Unknown	Human		Broad Institute
TA464	5.11	B2	Unknown	Mammal		Broad Institute
M605	5.48	B2	Unknown	Mammal		Broad Institute

The strain in bold correspond to the strain sequenced in this study.

(1) ExPEC (extraintestinal pathogenic *E. coli*), UPEC (urinary pathogenic *E. coli*), EAEC (enteroaggregative *E.coli*), NMEC (Neonatal meningitis-associated *E. coli*), APEC (avian pathogenic *E. coli*).

(2) Publically available genomes were downloaded from NCBI or the Broad Institute.

5.2.2 Sequencing and annotation of the environmental strain E1227

The genome of the environmental strain E1227 was sequenced using the Roche 454 Sequencer available at the John Curtin School of Medical Research, the Australian National University. A single-ended sequencing strategy was performed. Sequence reads derived from 454 Sequencer were assembled with the Newbler Assembler software that yielded the total number of 141 contigs with the total length is about 5.2 Mb. The DNA sequences of 141 contigs derived from Newbler Assembler software were submitted to JCVI Annotation Service. The Annotation Service was run through JCVI's prokaryotic annotation pipeline which included gene finding with Glimmer, Blast-extend-repraze (BER) searches, HMM searches, TMHMM searches, SignalP predictions, and automatic annotations from AutoAnnotate.

5.2.3 Phylo-group determination of the environmental strain E1227

Based on MLST analysis, in this study the complete sequence of the 22 housekeeping genes (*adk, aspC, torC, icdA, kdsA, fumC, fadD, metG, lysP, grpE, recA, mutS, rpoS, dnaG, mdh, aroE, mtlD, gyrB, cyaA, purA, arcA,* and *clpX*) of *E. coli* were applied to determine phylo-group of *E. coli* E1227. The 22 housekeeping genes of E1227 derived from genome annotation were used to compare with 30 known phylo-group *E. coli* strains (A, B1, B2, D, E, and F) available at NCBI and Broad Institute. The *E. coli* reference strains used for comparison comprised 8 A strains, 6 B1 strains, 7 B2 strains, 3 D strains, 4 E strains, and 2 F strains. *Escherichia* E1118 (clade V), *E. fergusonii* ATCC35469, and *E. albertii* B156 were also included as outgroups in the analysis. The 22 housekeeping genes of each strain were concatenated and loaded to the MEGA5 program (Tamura et al., 2011) to generate a phylogenetic tree.

The phylogenetic relationships among E1227 and known phylo-group *E. coli* reference strains (NCBI and Broad Institute) revealed that E1227 was closely related to *E. coli* SMS-3-5 and *E. coli* IAI39 which are members of phylo-group F of *E. coli* (Fig. 1). The *E. coli* E1227 belonging to phylo-group F was then included as a member of phylo-group F to use for the comparative genomic analysis.

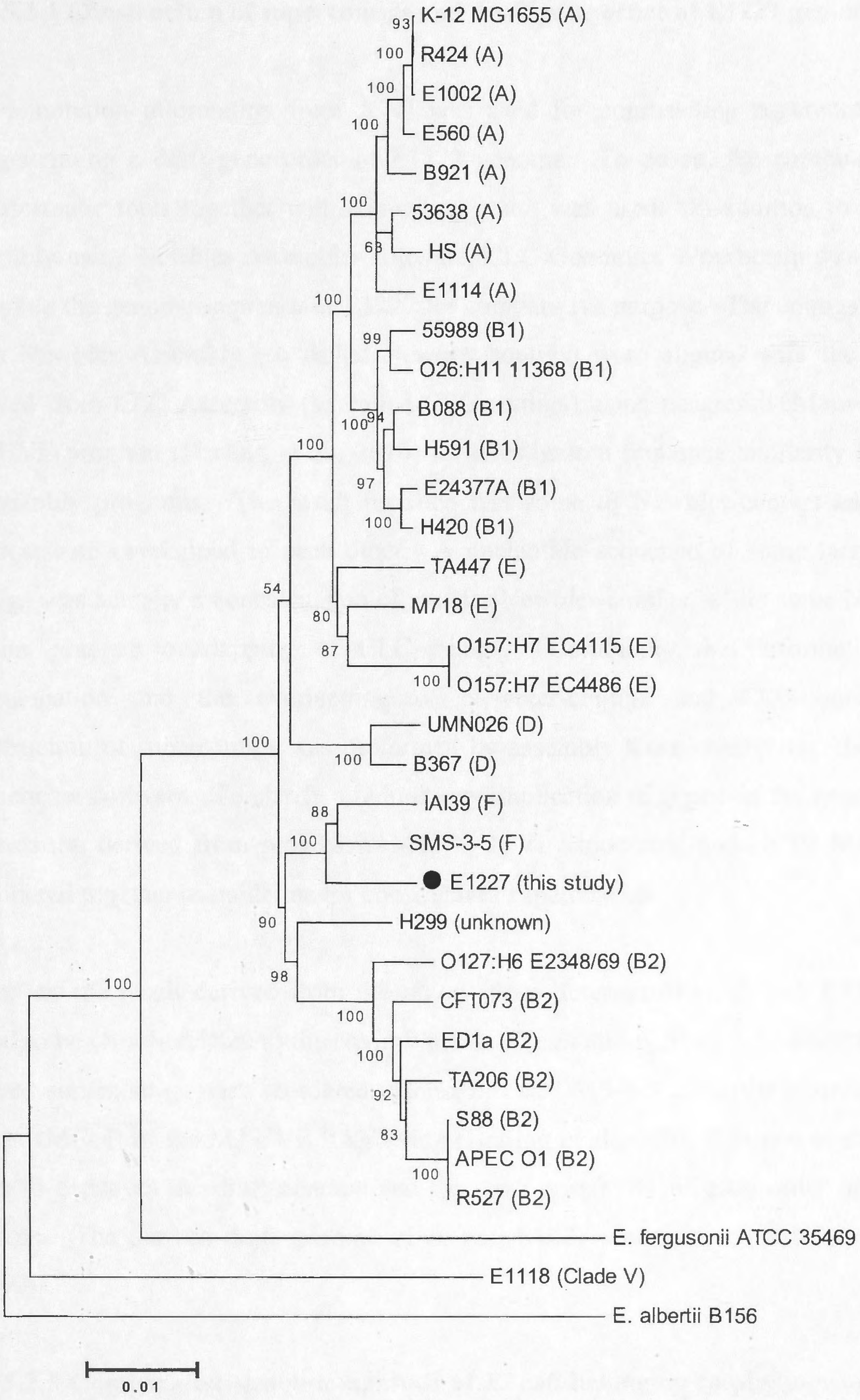


Figure 1. Maximum-likelihood tree of *E. coli* E1227 and known phylo-group *E. coli* strains based on the complete sequences of 22 housekeeping genes. Bootstrap values are based on 500 replicates. Bootstrap confidence values greater than 50% are listed to the left of the nodes. Phylogenetic analysis was constructed using MEGA5 (Tamura et al., 2011).

5.2.4 Construction of supercontigs and draft gene order of E1227 genome

The annotation information from JCVI was used for constructing supercontigs and for generating a draft gene order of E1227 genome. To do so, the combination of bioinformatic tools together with annotation result was used. In addition to genome assembly using Newbler Assembler software, CLC Genomics Workbench was used to assemble the genome sequence of E1227 for comparative purpose. The contigs derived from Newbler Assembly (so called Newbler-contigs) were aligned with the contigs derived from CLC Assembly (so called CLC-contigs) using progressiveMauve of the MAUVE program (Darling et al., 2010) to investigate a sequence similarity between 2 assembly programs. The result revealed that some of Newbler-contigs and CLC-contigs were overlapped to each other. A nucleotide sequence of some large CLC-contigs was actually a concatenation of smaller Newbler-contigs, while some Newbler-contigs was an overlapping of CLC-contigs. Combining this information: the concatenation and the overlapping of Newbler-contigs and CLC-contigs, the construction of supercontigs was performed by assembly these contigs together using Sequencher software. To clarify a redundancy/duplication of genes on the genome, the information derived from progressiveMauve, JCVI annotation, and NCBI blast were considered together as guideline for constructing supercontigs.

Based on the result derived from the phylo-group determination, *E. coli* E1227 was found to be closely related to the environmental *E. coli* strain, SMS-3-5. Therefore, the derived supercontigs were reordered against *E. coli* SMS-3-5 using the Mauve Contig Mover (MCM) of the MAUVE Reordering (Darling et al., 2004, Rissman et al., 2009) so as to construct the draft genome and the most possibility of gene order on E1227 genome. The derived draft genome of *E. coli* E1227 was further used for the next analyses.

5.2.5 Comparative genomic analysis of *E. coli* belonging to phylo-group F, D, and B2

The phylogenetic analysis of 23 *E. coli* strains (NCBI: 10, Broad: 12, and this study: 1) were performed using whole-genome sequences. The *E. coli* genome sequences included in the study were 9 B2 strains, 8 D strains, 5 F strains, and 1 unknown phylo-group strain (Table 1).

To construct the phylogeny of *E. coli* strains, the core genome was used. For the draft genomes of 12 *E. coli* strains retrieved from the Broad Institute, the draft genomes of each strain were firstly reordered against a closely related *E. coli* strain belonging to the same phylo-group from GenBank using the MCM of the MAUVE program to reorder contigs in their draft genomes relative to their related *E. coli* reference strains (Rissman et al., 2009). The derived MCM output files in the eXtended Multi-FastA file format (.xmfa) of each strain were then used for multiple genome alignments with the other 10 *E. coli* GenBank reference strains and *E. coli* E1227 belonging to phylo-group F (this study) using the progressiveMauve of the MAUVE program (Darling et al., 2010). The output files derived from progressiveMauve: the .xmfa file contains the complete genome alignment, the .bbcols file contains a region of the alignment where one or more genomes have a sequence element that one or more others lack, and the .backbone file contains regions conserved among subsets of the genomes under study were then used as input files for the next steps. The core genome was extracted from the .xmfa and .bbcols files using stripSubsetLCBs script (available at http://gel.ahabs.wisc.edu/mauve/snapshots/) to generate the .xmfa file of the core genome sequences (core alignment blocks) greater than 500 nt. The derived core genome sequence in the .xmfa file format was converted to the FASTA file format using xmfa2fasta script (personal communication) and was then used as an input file for constructing a maximum likelihood tree using PhyML (Guindon et al., 2010).

The phylogeny inferred from the variable genome data (the dispensable genome), that is genes not found in all *E. coli* strains was also constructed for comparative purpose. The gene content data were extracted from the .backbone file (derived from the progressiveMauve) using bbFilter script (available at http://gel.ahabs.wisc.edu/mauve/snapshots/) to generate the .bin file containing binary features (presence/absence) of blocks in a particular genome. The gene content data derived from the .bin file was converted to Nexus format (.nex) using GenAlEX (Peakall and Smouse, 2006). The derived gene content data in the .nex file was used to construct the phylogenetic tree using SplitTree4 (Parameters: GeneContentDistance>UPGMA>

ConsensusTree >Phylogram) (Huson and Bryant, 2006) and was exported to Newick format (.nwk) to use as an input file for the next step.

The variable genome with feature gain and loss and the distribution of the variable genome among the different *E. coli* strains were visualized using GenoPlast (Didelot et al., 2008). As GenoPlast requires fully bifurcating tree, however, PhyML did not yield the maximum likelihood tree inferred from the core genome in a required topology. Therefore, the core genome sequences in FASTA format was used as an input file to construct the UPGMA tree (fully bifurcating tree) using MEGA5 (Tamura et al., 2011). Trees inferred from the core and the variable genome together with the file tabulating the presence/ absence of each DNA block in each of *E. coli* strains were used as input files for GenoPlast.

The particular blocks of the variable genome present in one and more phylo-groups were also identified from the .backbone file generated from the progressiveMauve. The nucleotide sequences of these DNA blocks were investigated using Blast similarity search via the Board Institute and NCBI blast to determine whether they were unique to the phylo-group or shared by more phylo-groups in this study (F, D, and B2).

5.2.6 Metabolic profiling

The set of KEGG (Kyoto Encyclopedia of Genes and Genomes) map metabolic pathways predicted for the 23 *E. coli* genomes were compared. The genome sequences of all strains were submitted to RAST (Rapid Annotation using Subsystem Technology: http://rast.nmpdr.org/) (Aziz et al., 2008) to construct KEGG map metabolic pathway for each strain. For *E. coli* E1227, the draft genome (re-ordered supercontigs) from the previous step 2.4 was submitted. The derived data in Tab Delimited file format, the number of enzymes (distinct ECs) for pathway x in a given organism/total number of enzymes (distinct ECs) in the same pathway x defined in the KEGG database, was then used to perform a hierarchical clustering (HCL) according to their metabolic capabilities using the pathway completion values via the MeV software (MultiExperiment Viewer: http://www.tm4.org/mev/) (Saeed AI, 2003, Saeed AI, 2006).

5.3 Results

5.3.1 Phylogenetic relationships of *E. coli* belonging to phylo-groups F, D, and B2

A comparative genomics approach was used to investigate the phylogenetic relationships of 23 *E. coli* genomes belonging to phylo-groups F, D, B2, and 1 unknown phylo-group strain H299. The genome of these *E. coli* strains range from 4.85 to 6.13 Mb and the average length of those 23 genomes is approximately 5 Mb. On average 68% (3.52 Mb) represents the core genome. The maximum likelihood and UPGMA trees of the 23 *E. coli* strains inferred from the core genome were constructed (Fig. 2 and 3A, repectively). The results reveal that these strains spilt into 2 major clusters: cluster A (phylo-group F, D, and unknown phylo-group H299 strains) and cluster B (phylo-group B2 strains). The phylo-group F and D strains share their core genome and appear to be sister groups, while phylo-group B2 strains were clustered together clearly distinct from those phylo-groups F and D.

The phylogeny for the 23 *E. coli* genomes inferred from the gene content data (the dispensable genome) that is genes not found in all *E. coli* strains was also constructed (Fig. 4A) to compare with the tree inferred from the core genome (Fig. 3A). The phylogeny based on the gene content data reveals that these *E. coli* strains fall into 2 major clusters: cluster A (phylo-group F and D strains) and cluster B (phylo-group B2 and unknown phylo-group H299 strains) (Fig. 4A). The results indicate that phylo-group F and D strains were always branched on the same cluster in the phylogenies based on both core genome and gene content data. However the F strain E1227 was found to be located within phylo-group D on the phylogeny based on the gene content data. This suggests that although *E. coli* E1227 belonging to phylo-group F shares core genome with phylo-group F strains, the strain has more similar gene content to strains belonging to phylo-group D. While the unknown phylo-group H299 that previously clustered together with phylo-groups F and D based on tree inferred from the core genome data now was located as part of cluster B together with phylo-group B2 strains based on tree inferred from the gene content data (Fig. 4A). Therefore, that the strain H299 has an ambiguous status could be due to the fact that the strain shares core

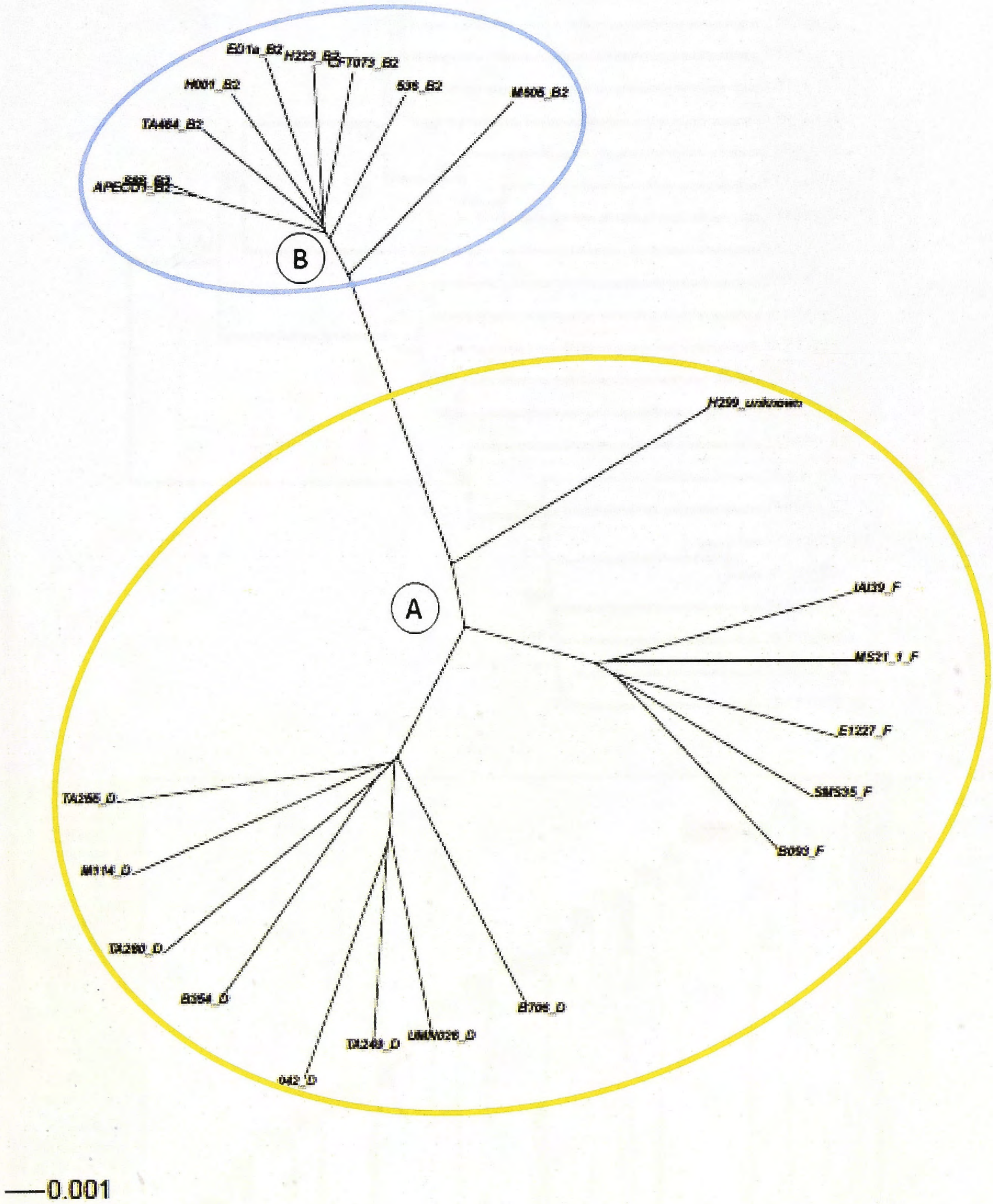


Figure 2. Maximum likelihood tree of 23 *E. coli* strains (phylo-group F, D, and unknown phylo-group H299 strains) inferred from the core genome data based on 3.52 Mb of nucleotide sequence. The tree was constructed using PhyML (Model: GTR, Likelihood ratio test: SH-like branch support).

genome with phylo-group F and D strains, while having more similar gene content to phylo-group B2 strains.

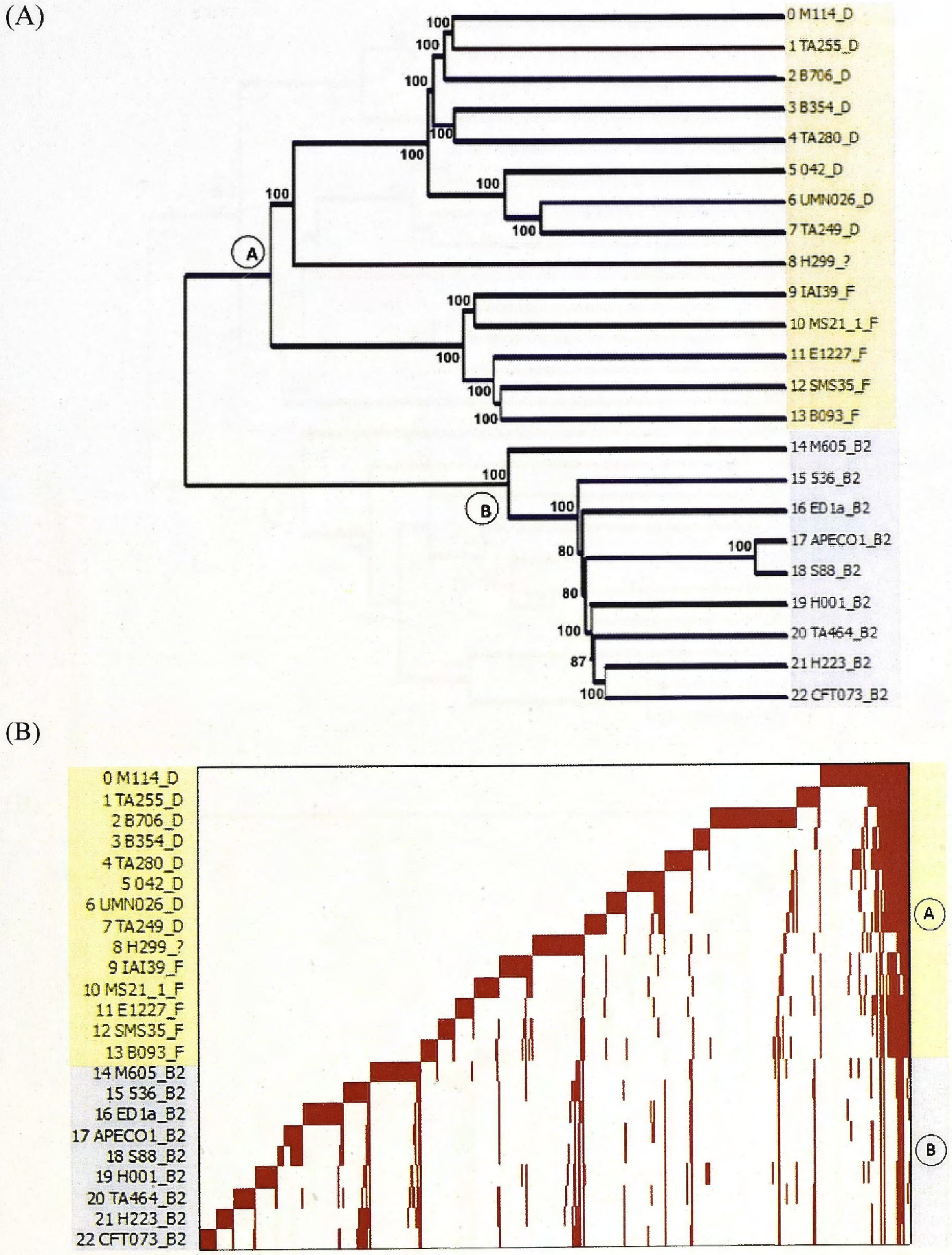


Figure 3. The phylogeny of the 23 *E. coli* genomes inferred from the core genome data (3.52 Mb) implemented using GenoPlast. (A) The UPGMA tree with feature gain in red below and loss in blue above the branches. Bootstrap values are based on 500 replicates and bootstrap confidence values greater than 50% derived from MEGA5 are listed to the left of the nodes. (B) The variable genome of each node of the phylogeny shown in (A). Each row of red bars corresponds to a strain in the phylogeny. The horizontal width of the red bars corresponds to the size of the variable genome for each strain, and overlapping red bars corresponds to variable genome content shared by two or more strains.

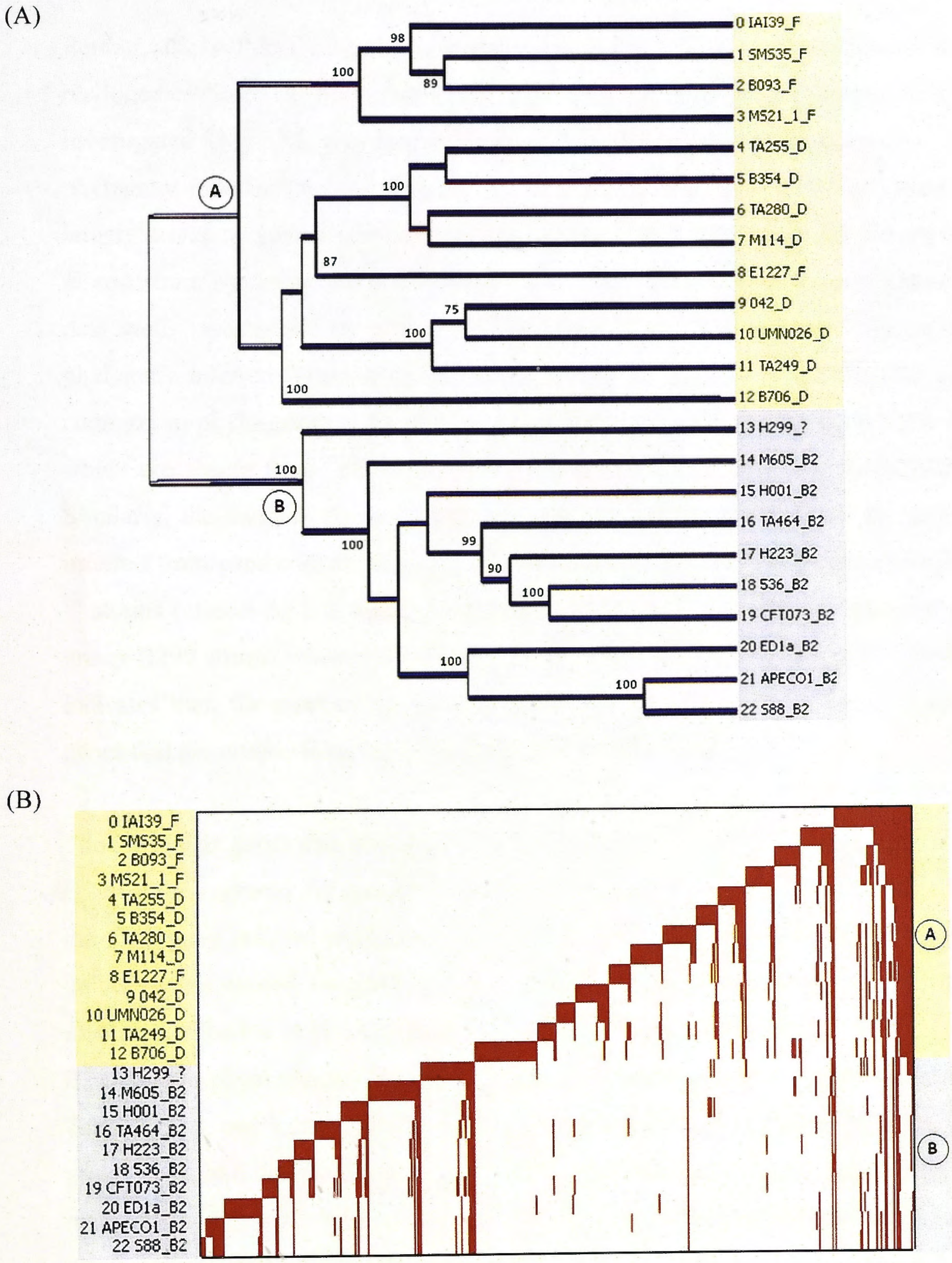


Figure 4. The phylogeny of the 23 *E. coli* genomes inferred from the gene content data implemented using GenoPlast. (A) The UPGMA tree with feature gain in red below and loss in blue above the branches. Bootstrap values are based on 500 replicates and bootstrap confidence values greater than 50% derived from MEGA5 are listed to the left of the nodes. (B) The variable genome of each node of the phylogeny shown in (A). Each row of red bars corresponds to a strain in phylogeny. The horizontal width of the red bars corresponds to the size of the variable genome for each strain, and overlapping red bars corresponds to variable genome content shared by two or more strains.

Further, the patterns of genome content evolution of 23 *E. coli* genomes for the phylogenies inferred both from the core genome and gene content data were investigated (Fig. 3A and 4A, respectively). The results of genome flux on the phylogeny indicate that the diversity of these 23 *E. coli* genomes was found to be largely driven by gene acquisition and loss events. The variable genome pattern of each *E. coli* strain shown on the phylogeny inferred from the core genome and gene content data were investigated as shown by Fig. 3B and 4B, respectively. Based on the phylogeny inferred from core genome data, there are regions of the variable genome common to phylo-group F, D, and unknown phylo-group H299 strains (cluster A) but which are absent from phylo-group B2 strains (cluster B) and vice versa (Fig. 3B). Similarly, the variable genome pattern of each *E. coli* strain based on the phylogeny inferred from gene content data also reveals the sharing of genes by phylo-group F and D strains (cluster A) but which are absent from phylo-group B2 and unknown phylo-group H299 strains (cluster B) and vice versa (Fig. 4B). The result in this study also indicates that, for most of the *E. coli* strains, the balance of the genome consists of genes that are unique to each *E. coli* strain (Fig. 3B and 4B).

The particular genes that are shared by phylo-group F, D, and unknown phylo-group H299 strains (cluster A) and phylo-group B2 strains (cluster B) based on the result of the phylogeny inferred from core genome data (Fig. 3B) were investigated. Also the unique genes shared by phylo-group F strains and by phylo-group D strains were identified. The list of genes sharing by strains in cluster A, sharing by strains in cluster B, unique to phylo-group F strains, and unique to phylo-group D strains was shown in Table 2, 3, 4, and 5, respectively. The result shows that most of genes share by phylo-group F, D, and unknown phylo-group H299 strains (cluster A) are genes involved in metabolism (Table 2). While most of genes share by phylo-group B2 strains (cluster B) are genes with unknown function (Table 3).

Focusing on strains belonging to phylo-group F, most of the genes shared by phylo-group F strains are gene with known function (Table 4). Some of genes conserved among F strains (absent in D, B2, and H299) were not restricted to phylo-group F as they also are present in strains belonging to other phylo-groups (Table 4). Four gene products: Mandelate racemase/muconate lactonizing enzyme family protein, Transporter; major facilitator family, Transcriptional regulator; GntR family, and SMS domain protein were also detected in *E. coli* O7:K1 str. CE10, but not in other *E. coli* strains in the NCBI database (data not shown). *E. coli* O7:K1 str. CE10 is Neonatal meningitis-associated *E. coli* (NMEC) strain belonging to phylo-group D (Lu et al., 2011). Based on genome-based phylogenetic analysis by Lu and colleagues (2011), the O7:K1 str. CE10 was reported to be strain closely related to IAI39 (O7:K1), which is the uropathogenic *E. coli* (UPEC) strain belonging to phylogroup F. This result emphasizes a closely relationship between *E. coli* strains belonging to phylo-groups F and D, where they appear to be a sister group.

Table 2. Genes on conserved segments (alignment blocks greater than 500 nt) shares by strains in cluster A (phylo-group F, D, and unknown phylo-group H299 strains).

Gene/ annotation	nt	Gene product	Function
cynX	1155	putative cyanate transporter	Transport and binding proteins/ Detoxification
cynR	900	DNA-binding transcriptional dual regulator	Central intermediary metabolism/ Transcription/ DNA interactions/ Detoxification
yagU	615	conserved hypothetical protein; putative inner membrane protein	Unknown
ECIAI39_0467	1017	conserved hypothetical protein	Unknown
yeaX	966	putative dioxygenase subunit	Enzymes of unknown specificity
yeaW	1125	putative 2Fe-2S cluster-containing dioxygenase subunit	Enzymes of unknown specificity
yeaV	1446	putative transporter	Transport and binding proteins
puuC	1488	gamma-Glu-gamma-aminobutyraldehyde dehydrogenase, NAD(P)H-dependent	Amino acids and amines
puuB	1281	gamma-Glu-putrescine oxidase, FAD/NAD(P)-binding	Amino acids and amines
puuE	1266	GABA aminotransferase, PLP-dependent	Pyridoxine/ Energy metabolism
yddV	1383	putative diguanylate cyclase YddV	Control
lsrB	1023	AI2 transporter ; periplasmic-binding component of ABC superfamily	Carbohydrates, organic alcohols, and acids
lsrF	876	putative aldolase	Scavenge (Catabolism)
lsrG	291	conserved hypothetical protein	Unknown
ECIAI39_1904	2,649	putative Outer membrane autotransporter barrel, putative pectin lyase fold	Transport and binding proteins
hcaR	891	DNA-binding transcriptional activator of 3-phenylpropionic acid catabolism	Energy metabolism/ Transcription/ DNA interactions
hcaE	1362	3-phenylpropionate dioxygenase, large (alpha) subunit	Energy metabolism

Table 2. (cont.)

Gene/ annotation	nt	Gene product	Function
hcaF	519	3-phenylpropionate dioxygenase, small (beta) subunit	Energy metabolism
hcaC	321	3-phenylpropionate dioxygenase, putative ferredoxin subunit	Energy metabolism
hcaB	813	2,3-dihydroxy-2,3-dihydrophenylpropionate dehydrogenase	Energy metabolism
hcaD	1203	phenylpropionate dioxygenase, ferredoxin reductase subunit	Energy metabolism
yliK	2145	methylmalonyl-CoA mutase	Central intermediary metabolism/ Energy metabolism
argK	996	membranc ATPase/protein kinase	Transport and binding proteins
rtcA	1017	RNA 3'-terminal phosphate cyclase	RNA processing/ tRNA and rRNA base modification
rtcB	1227	conserved hypothetical protein	Unknown
rtcR	1599	sigma 54-dependent transcriptional regulator of rtcBA expression	Transcription/ DNA interactions
ECIAI39_3996	321	conserved hypothetical protein	Unknown
ECIAI39_3997	1014	putative membrane protein	Unknown
ECIAI39_3998	1224	conserved hypothetical protein	Unknown
yiaM	747	transporter	Transport and binding proteins
yiaN	1278	transporter	Transport and binding proteins
ECIAI39_4096	861	putative transcriptional regulator protein	Transcription/ DNA interactions
ECIAI39_4475	540	conserved hypothetical protein	Unknown
yjcO	690	conserved hypothetical protein	Unknown
cynS	471	cyanate aminohydrolase	Nitrogen metabolism/ Detoxification
cynT	660	carbonic anhydrase	Detoxification
sfmA	543	putative fimbrial-like adhesin protein	Surface structures/ Protect/ Explore
sfmC	693	pilin chaperone, periplasmic	Protein folding and stabilization
sfmD	2610	putativc outer membrane export usher protein	Circulate/ Shape
sfmH	978	putative fimbrial-like adhesin protein	Surface structures/ Chemotaxis and motility
ybiU	1260	conserved hypothetical protein	Unknown
yeaU	1086	putative tartrate dehydrogenase	Fermentation/ Circulate
yeaT	924	putative DNA-binding transcriptional regulator	Transcription/ DNA interactions
ECIAI39_1334	1899	putative ankyrin repeat regulatory protein	Regulatory functions
lhr	4617	putative ATP-dependent helicase	DNA replication, recombination, and Repair/ Circulate

Table 2. (cont.)

Gene/ annotation	nt	Gene product	Function
puuP	1386	putrescine importer	Transport and binding proteins
puuA	1419	gamma-Glu-putrescine synthase	Glutamate family/ Nitrogen metabolism/ Amino acids and amines
puuD	765	gamma-Glu-GABA hydrolase	Amino acids and amines
puuR	558	DNA-binding transcriptional repressor	Regulatory functions
ycaM	1431	putative transporter	Transport and binding proteins/ Control
yfeT	858	putative DNA-binding transcriptional regulator	Control
ygfH	1479	propionyl-CoA:succinate-CoA transferase	Central intermediary metabolism
arsB	1290	arsenite/antimonite transporter	Anions/ Detoxification
lyxK	1497	L-xylulose kinase	Sugars
sgbU	861	putative L-xylulose 5-phosphate 3-epimerase	Central intermediary metabolism
sgbE	696	L-ribulose-5-phosphate 4-epimerase	Sugars
ECIAI39_4097	1,401	putative permease of the major facilitator superfamily	Transport and binding proteins

Note: *E. coli* IAI39 (GenBank accession no. CU928164.2) was used as reference genes.

Table 3. Genes on conserved segments (alignment blocks greater than 500 nt) sharesby strains in cluster B (phylo-group B2 strains)

Gene / annotation	nt	Product	Function
c1819	309	Putative conserved protein	putative transport
ydbA_1	1008	possible pseudogene (bigA), internally repeated	Unknown
ydjF	759	transcriptional regulator YdjF	putative regulator
ydjG	981	hypothetical oxidoreductase ydjG	Unknown
ydjH	969	hypothetical sugar kinase ydjH	Unknown
ydjI	837	hypothetical protein ydjI	Unknown
ydjJ	1044	Hypothetical zinc-type alcohol dehydrogenase-like protein ydjJ	Unknown
ydjK	1380	Hypothetical metabolite transport protein ydjK	putative transport
c2180	153	hypothetical protein	Unknown
ydjL	1077	Hypothetical zinc-type alcohol dehydrogenase-like protein ydjL	Unknown
yfaV	1329	hypothetical transport protein YfaV	putative transport
yfaW	1218	hypothetical protein yfaW	Unknown
c3302	1044	hypothetical protein	Unknown
c3303	147	hypothetical protein	Unknown
c3304	552	hypothetical protein	Unknown
c4013	1047	hypothetical protein	Unknown
c4014	963	hypothetical protein	Unknown
c4015	990	Ribose transport system permease protein rbsC	transport; Transport of small molecules: Carbohydrates, organic acids, alcohols
c4016	1527	Ribose transport ATP-binding protein rbsA	transport; Transport of small molecules: Carbohydrates, organic acids, alcohols
c4017	897	Putative ribose ABC transporter	Unknown
gatY	855	Tagatose-bisphosphate aldolase gatY	enzyme; Degradation of small molecules: Carbon compounds
c4019	129	hypothetical protein	Unknown
c4020	831	hypothetical protein	Unknown
c4021	1011	hypothetical protein	Unknown
sucA	2820	2-oxoglutarate dehydrogenase E1 component	enzyme; Energy metabolism, carbon: TCA cycle
c5033	201	hypothetical protein	Unknown
c5034	1056	dihydrolipoamide succinyltransferase component of 2-oxoglutarate dehydrogenase complex	enzyme; Energy metabolism, carbon: TCA cycle
c5035	1419	Putative 2-oxoglutarate dehydrogenase	Unknown
c5036	1170	Succinyl-CoA synthetase beta chain	enzyme; Energy metabolism, carbon: TCA cycle
c5037	873	Succinyl-CoA synthetase alpha chain	enzyme; Energy metabolism, carbon: TCA cycle
c5038	1506	Putative membrane-bound protein	Unknown
c5039	1110	Putative lactate dehydrogenase	Unknown
c5040	1359	Putative c4-dicarboxylate transport transcriptional Regulatory protein	Unknown
c5041	1821	Putative transport sensor protein	Unknown
c5042	135	Hypothetical protein	Unknown
c5043	141	Hypothetical protein	Unknown
c5044	153	Hypothetical protein	Unknown

Table 3. (cont.)

Gene / annotation	nt	Product	Function
yjcO	744	Hypothetical protein yjcO precursor	Unknown
yddO	705	Hypothetical ABC transporter ATP-binding protein yddO	putative transport
c5078	897	Putative oligopeptide ABC transporter	Unknown
yddQ	834	hypothetical ABC transporter permease protein yddQ	putative transport
yddR	1143	hypothetical ABC transporter permease protein yddR	putative transport
c5081	1569	putative conserved protein	putative factor

Note: *E. coli* CFT073 (GenBank accession no. AE014075.1) was used as reference genes.

Table 4. Genes on conserved segments (alignment blocks greater than 500 nt) shareed by phylo-group F strains (absent from phylo-group D, B2, and unknown phylo-group H299 strain)).

Gene/annotation (nt)	Product	Function	Comment
dinJ (261)	antitoxin of YafQ-DinJ toxin-antitoxin system	DNA replication, recombination, and repair/ Control	**Broad Blast Hits**: phylogroup A and E strains
yafQ (249)	toxin of the YafQ-DinJ toxin-antitoxin system	Control	**Broad Blast Hits**: phylogroup A and E strains
ECIAI39_3907 (993)	conserved hypothetical protein; Putative ATP binding protein of ABC transporter	Unknown	**Broad Blast Hits**: phylogroup B1 and E strains
ECIAI39_4411 (813)	putative Shikimate dehydrogenase	Amino acid biosynthesis	**Broad Blast Hits**: *E. albertii* B156
ECIAI39_0931 (297)	conserved hypothetical protein, putative plasmid stabilisation system protein	Unknown	-
ECIAI39_0932 (297)	conserved hypothetical protein	Unknown	-
ECIAI39_3066 (687)	conserved hypothetical protein; putative membrane protein	Unknown	-
ECIAI39_3067 (717)	conserved hypothetical protein; putative membrane protein	Unknown	-

Note: *E. coli* IAI39 (GenBank accession no. CU928164.2) was used as reference genes.

Table 5. Genes on conserved segments (alignment blocks greater than 500 nt) shared by phylo-group D strains.

Gene/ annotation	nt	Product	Function
yagR	2199	putative oxidoreductase with molybdenum-binding domain	Enzymes of unknown specificity
yagQ	957	conserved hypothetical protein	Unknown
opdE	1194	transcription regulatory protein opdE	DNA interactions
ECUMN_0311	885	putative LysR-like transcriptional regulator	DNA interactions
ECUMN_0312	1158	conserved hypothetical protein	Unknown
yagS	957	putative oxidoreductase with FAD-binding domain	Enzymes of unknown specificity
siiEA	20778	adhesin for cattle intestine colonization	Explore
ECUMN_3885	1152	putative fimbrial adhesin	Surface structures
ECUMN_3886	804	putative periplasmic pilin chaperone	Explore
ECUMN_3887	525	putative exported fimbrial-like adhesin protein	Surface structures
ECUMN_3888	573	putative fimbrial minor structural subunit	Explore
aufC	2592	putative outer membrane export usher protein AufC	Cell envelope
ECUMN_3890	753	putative fimbrial chaperone protein precursor	Explore
aufA	705	Auf fimbriae major fimbrial subunit AufA	Surface structures
ECUMN_0457	1572	conserved hypothetical protein	Unknown
yihN	1266	putative transporter	Transport and binding proteins
yjcF	1293	conserved hypothetical protein	Unknown
ycgV	2868	putative adhesin ; putative autotransporter	Surface structures
arpA	2187	regulator of acetyl CoA synthetase	Biosynthesis/ Regulatory functions
ymjC	639	putative NAD(P) binding enzyme	Enzymes of unknown specificity

Note: *E. coli* UMN026 (GenBank accession no. CU928163.2) was used as reference genes.

5.3.2 Metabolic profile diversity

The metabolic pathways predicted for 23 *E. coli* genomes were compared (Table 6). The hierarchical clustering of metabolic profiles of 23 *E. coli* strains clustered the F and D strains (also unknown phylo-group H299 strain) together separately from the B2 strains (Fig. 5). Only one D strain, M114, was found to have a distinct metabolic profile from other D strains and was clustered with the B2 strains. This result reveals that *E. coli* strains belonging to phylo-group F and D are closely related and likely to share more metabolic pathways. In addition, this result is congruent with the result of the investigated genes common to phylo-group F, D, and unknown phylo-group H299 strains that most of genes found to be genes involved in metabolism (Table 2).

Table 6. The number of distinct enzymes combinations (distinct ECs) for pathway x in a given organism/total number of distinct ECs in the same pathway x defined in the KEGG database.

		(%) = The number of distinct ECs for pathway x in a given organism / Total number of distinct ECs in the same pathway x				
KEGG metabolic pathway	**Distinct ECs**	**IAI39 (F)**	**E1227 (F)**	**B093 (F)**	**MS-21-1 (F)**	**SMS-3-5 (F)**
Alanine, aspartate and glutamate metabolism	43	55.8	55.8	55.8	53.5	55.8
Arginine and proline metabolism	97	42.3	43.3	42.3	42.3	42.3
Ascorbate and aldarate metabolism	44	36.4	38.6	36.4	36.4	36.4
Biotin metabolism	12	50.0	50.0	50.0	50.0	50.0
Carbon fixation in photosynthetic organisms	25	60.0	56.0	56.0	56.0	56.0
Citrate cycle (TCA cycle)	22	63.6	63.6	63.6	63.6	63.6
Cysteine and methionine metabolism	64	35.9	35.9	35.9	35.9	35.9
Fatty acid biosynthesis	21	52.4	52.4	52.4	52.4	52.4
Fatty acid elongation in mitochondria	8	37.5	37.5	37.5	37.5	37.5
Fatty acid metabolism	29	41.4	48.3	41.4	41.4	37.9
Folate biosynthesis	25	48.0	44.0	44.0	44.0	44.0
Fructose and mannose metabolism	65	41.5	41.5	41.5	41.5	41.5
Glutathione metabolism	40	50.0	50.0	50.0	50.0	50.0
Glycerolipid metabolism	36	25.0	30.6	25.0	25.0	25.0
Glycerophospholipid metabolism	50	38.0	38.0	40.0	40.0	40.0
Glycine, serine and threonine metabolism	57	45.6	47.4	45.6	43.9	45.6
Glycolysis / Gluconeogenesis	41	53.7	56.1	53.7	53.7	53.7
Glycosaminoglycan degradation	16	31.2	31.2	31.2	31.2	31.2
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	8	50.0	50.0	37.5	37.5	37.5
Glyoxylate and dicarboxylate metabolism	58	32.8	41.4	37.9	36.2	37.9

Table 6. (cont.)

		(%) = The number of distinct ECs for pathway x in a given organism Total number of distinct ECs in the same pathway x				
KEGG metabolic pathway	**Distinct ECs**	**IAI39 (F)**	**E1227 (F)**	**B093 (F)**	**MS-21-1 (F)**	**SMS-3-5 (F)**
Histidine metabolism	37	40.5	35.1	43.2	32.4	35.1
Inositol phosphate metabolism	40	12.5	12.5	12.5	12.5	12.5
Lipopolysaccharide biosynthesis	22	86.4	86.4	86.4	86.4	86.4
Lysine biosynthesis	31	54.8	54.8	54.8	54.8	54.8
Lysinc degradation	54	24.1	22.2	18.5	18.5	24.1
Methane metabolism	33	27.3	27.3	27.3	27.3	27.3
N-Glycan biosynthesis	29	3.4	3.4	3.4	3.4	3.4
Nicotinate and nicotinamide metabolism	47	40.4	40.4	40.4	40.4	40.4
One carbon pool by folate	24	54.2	54.2	54.2	54.2	54.2
Pantothenate and CoA biosynthesis	28	53.6	53.6	57.1	57.1	57.1
Pentose and glucuronate interconversions	56	46.4	46.4	46.4	44.6	46.4
Pentose phosphate pathway	37	64.9	64.9	64.9	64.9	64.9
Phenylalanine metabolism	51	29.4	23.5	25.5	25.5	25.5
Phenylalanine, tyrosine and tryptophan biosynthesis	31	71.0	71.0	71.0	71.0	71.0
Phenylpropanoid biosynthesis	30	23.3	20.0	20.0	20.0	20.0
Porphyrin and chlorophyll metabolism	66	33.3	33.3	33.3	33.3	33.3
Primary bile acid biosynthesis	19	5.3	5.3	5.3	5.3	5.3
Propanoate metabolism	47	42.6	44.7	40.4	40.4	40.4
Purine metabolism	104	55.8	55.8	55.8	55.8	55.8
Pyrimidine metabolism	64	50.0	48.4	50.0	50.0	50.0
Pyruvate metabolism	64	50.0	53.1	51.6	51.6	51.6
Reductive carboxylate cycle (CO2 fixation)	13	69.2	69.2	69.2	69.2	69.2
Riboflavin metabolism	16	68.8	62.5	68.8	68.8	68.8
Sphingolipid metabolism	29	31.0	27.6	27.6	27.6	27.6
Steroid biosynthesis	27	7.4	7.4	7.4	7.4	7.4
Sulfur metabolism	30	33.3	33.3	33.3	33.3	33.3
Synthesis and degradation of ketone bodies	6	16.7	16.7	16.7	33.3	16.7
Terpenoid backbone biosynthesis	27	48.1	44.4	44.4	44.4	44.4
Thiamine metabolism	16	50.0	50.0	50.0	50.0	50.0
Tyrosine metabolism	63	17.5	22.2	17.5	22.2	17.5
Ubiquinone and other terpenoid-quinone biosynthesis	25	64.0	60.0	56.0	56.0	56.0
Valine, leucine and isoleucine biosynthesis	18	72.2	72.2	72.2	72.2	72.2
Valine, leucine and isolcucinc degradation	34	26.5	32.4	26.5	26.5	23.5
Vitamin B6 metabolism	26	30.8	30.8	30.8	30.8	30.8

Table 6. (cont.)

		(%) = The number of distinct ECs for pathway x in a given organism / Total number of distinct ECs in the same pathway x							
KEGG metabolic pathway	**Distinct ECs**	**B354 (D)**	**B706 (D)**	**O42 (D)**	**M114 (D)**	**UMN026 (D)**	**TA249 (D)**	**TA255 (D)**	**TA280 (D)**
Alanine, aspartate and glutamate metabolism	43	55.8	53.5	55.8	53.5	53.5	55.8	53.5	53.5
Arginine and proline metabolism	97	42.3	43.3	42.3	42.3	43.3	42.3	42.3	42.3
Ascorbate and aldarate metabolism	44	38.6	40.9	36.4	38.6	40.9	38.6	38.6	38.6
Biotin metabolism	12	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Carbon fixation in photosynthetic organisms	25	52.0	56.0	60.0	56.0	56.0	56.0	56.0	56.0
Citrate cycle (TCA cycle)	22	59.1	63.6	63.6	63.6	63.6	63.6	63.6	63.6
Cysteine and methionine metabolism	64	35.9	35.9	35.9	35.9	39.1	35.9	35.9	35.9
Fatty acid biosynthesis	21	52.4	52.4	52.4	52.4	52.4	52.4	52.4	52.4
Fatty acid elongation in mitochondria	8	37.5	37.5	37.5	37.5	37.5	37.5	37.5	37.5
Fatty acid metabolism	29	37.9	41.4	41.4	37.9	41.4	37.9	41.4	37.9
Folate biosynthesis	25	44.0	48.0	48.0	44.0	44.0	44.0	44.0	44.0
Fructose and mannose metabolism	65	41.5	41.5	40.0	41.5	41.5	40.0	41.5	41.5
Glutathione metabolism	40	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Glycerolipid metabolism	36	27.8	27.8	25.0	27.8	30.6	27.8	25.0	27.8
Glycerophospholipid metabolism	50	40.0	38.0	38.0	38.0	38.0	38.0	40.0	40.0
Glycine, serine and threonine metabolism	57	45.6	45.6	43.9	45.6	45.6	45.6	45.6	45.6
Glycolysis / Gluconeogenesis	41	51.2	56.1	53.7	53.7	56.1	53.7	53.7	53.7
Glycosaminoglycan degradation	16	31.2	31.2	31.2	31.2	31.2	31.2	31.2	31.2
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	8	37.5	37.5	50.0	50.0	37.5	37.5	37.5	37.5
Glyoxylate and dicarboxylate metabolism	58	37.9	37.9	37.9	37.9	37.9	37.9	37.9	37.9
Histidine metabolism	37	35.1	37.8	32.4	32.4	37.8	35.1	35.1	35.1
Inositol phosphate metabolism	40	22.5	12.5	12.5	12.5	15.0	12.5	12.5	12.5
Lipopolysaccharide biosynthesis	22	86.4	86.4	86.4	86.4	86.4	86.4	86.4	86.4
Lysine biosynthesis	31	54.8	54.8	54.8	54.8	54.8	54.8	54.8	54.8
Lysine degradation	54	18.5	25.9	18.5	18.5	25.9	18.5	18.5	18.5
Methane metabolism	33	27.3	27.3	27.3	27.3	27.3	27.3	27.3	27.3
N-Glycan biosynthesis	29	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4
Nicotinate and nicotinamide metabolism	47	40.4	40.4	40.4	40.4	40.4	40.4	40.4	40.4
One carbon pool by folate	24	54.2	54.2	54.2	54.2	54.2	54.2	54.2	54.2
Pantothenate and CoA biosynthesis	28	57.1	57.1	53.6	57.1	57.1	57.1	57.1	57.1
Pentose and glucuronate interconversions	56	46.4	46.4	46.4	46.4	46.4	48.2	46.4	46.4
Pentose phosphate pathway	37	64.9	64.9	64.9	64.9	64.9	64.9	64.9	64.9
Phenylalanine metabolism	51	23.5	27.5	27.5	23.5	23.5	23.5	23.5	23.5
Phenylalanine, tyrosine and tryptophan biosynthesis	31	71.0	71.0	71.0	71.0	71.0	71.0	71.0	71.0
Phenylpropanoid biosynthesis	30	20.0	20.0	23.3	20.0	20.0	20.0	20.0	20.0
Porphyrin and chlorophyll metabolism	66	34.8	31.8	33.3	33.3	33.3	33.3	31.8	33.3

Table 6. (cont.)

		(%) = The number of distinct ECs for pathway x in a given organism / Total number of distinct ECs in the same pathway x							
KEGG metabolic pathway	Distinct ECs	B354 (D)	B706 (D)	O42 (D)	M114 (D)	UMN026 (D)	TA249 (D)	TA255 (D)	TA280 (D)
Primary bile acid biosynthesis	19	5.3	5.3	5.3	5.3	5.3	5.3	5.3	5.3
Propanoate metabolism	47	40.4	42.6	42.6	40.4	42.6	40.4	40.4	40.4
Purine metabolism	104	55.8	55.8	56.7	54.8	55.8	55.8	55.8	55.8
Pyrimidine metabolism	64	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Pyruvate metabolism	64	50.0	53.1	50.0	51.6	53.1	51.6	51.6	51.6
Reductive carboxylate cycle (CO2 fixation)	13	69.2	69.2	69.2	69.2	69.2	69.2	69.2	69.2
Riboflavin metabolism	16	68.8	68.8	68.8	68.8	68.8	68.8	68.8	68.8
Sphingolipid metabolism	29	27.6	27.6	31.0	27.6	27.6	27.6	27.6	27.6
Steroid biosynthesis	27	7.4	7.4	7.4	7.4	7.4	7.4	7.4	7.4
Sulfur metabolism	30	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
Synthesis and degradation of ketone bodies	6	16.7	33.3	16.7	33.3	16.7	16.7	50.0	16.7
Terpenoid backbone biosynthesis	27	44.4	44.4	48.1	44.4	44.4	44.4	44.4	44.4
Thiamine metabolism	16	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Tyrosine metabolism	63	15.9	25.4	15.9	22.2	22.2	22.2	15.9	28.6
Ubiquinone and other terpenoid-quinone biosynthesis	25	52.0	52.0	60.0	52.0	52.0	52.0	52.0	52.0
Valine, leucine and isoleucine biosynthesis	18	72.2	72.2	72.2	72.2	72.2	72.2	72.2	72.2
Valine, leucine and isoleucine degradation	34	23.5	26.5	26.5	23.5	26.5	23.5	29.4	23.5
Vitamin B6 metabolism	26	30.8	30.8	30.8	30.8	30.8	30.8	30.8	30.8

Table 6. (cont.)

		(%) = The number of distinct ECs for pathway x in a given organism / Total number of distinct ECs in the same pathway x									
KEGG metabolic pathway	Distinct ECs	ED1a (B2)	536 (B2)	APEC-O1 (B2)	CFT073 (B2)	H001 (B2)	H223 (B2)	M605 (B2)	S88 (B2)	TA464 (B2)	H299 (?)
Alanine, aspartate and glutamate metabolism	43	53.5	55.8	55.8	55.8	53.5	53.5	53.5	53.5	53.5	53.5
Arginine and proline metabolism	97	39.2	40.2	40.2	40.2	39.2	39.2	39.2	38.1	39.2	43.3
Ascorbate and aldarate metabolism	44	36.4	34.1	36.4	36.4	36.4	34.1	34.1	34.1	34.1	38.6
Biotin metabolism	12	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Carbon fixation in photosynthetic organisms	25	56.0	60.0	60.0	60.0	56.0	56.0	56.0	56.0	56.0	56.0
Citrate cycle (TCA cycle)	22	63.6	63.6	63.6	63.6	63.6	63.6	63.6	63.6	63.6	63.6
Cysteine and methionine metabolism.	64	35.9	34.4	35.9	35.9	35.9	35.9	35.9	39.1	35.9	35.9
Fatty acid biosynthesis	21	47.6	47.6	47.6	47.6	47.6	47.6	47.6	47.6	47.6	52.4
Fatty acid elongation in mitochondria	8	37.5	37.5	37.5	37.5	37.5	37.5	37.5	37.5	37.5	37.5
Fatty acid metabolism	29	34.5	37.9	44.8	41.4	37.9	37.9	37.9	34.5	37.9	44.8
Folate biosynthesis	25	48.0	48.0	48.0	48.0	48.0	48.0	48.0	48.0	48.0	44.0

Table 6. (cont.)

KEGG metabolic pathway	Distinct ECs	(%) = The number of distinct ECs for pathway x in a given organism / Total number of distinct ECs in the same pathway x									
		ED1a (B2)	536 (B2)	APEC-O1 (B2)	CFT073 (B2)	H001 (B2)	H223 (B2)	M605 (B2)	S88 (B2)	TA464 (B2)	H299 (?)
Fructose and mannose metabolism	65	41.5	41.5	40.0	40.0	41.5	41.5	41.5	41.5	41.5	41.5
Glutathione metabolism	40	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Glycerolipid metabolism	36	25.0	25.0	27.8	27.8	25.0	25.0	25.0	25.0	27.8	27.8
Glycerophospholipid metabolism	50	38.0	38.0	38.0	40.0	38.0	38.0	40.0	38.0	38.0	38.0
Glycine, serine and threonine metabolism	57	40.4	43.9	43.9	43.9	43.9	43.9	43.9	43.9	43.9	45.6
Glycolysis / Gluconeogenesis	41	53.7	53.7	56.1	56.1	53.7	53.7	53.7	53.7	53.7	56.1
Glycosaminoglycan degradation	16	31.2	31.2	31.2	31.2	31.2	31.2	31.2	31.2	31.2	31.2
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	8	37.5	50.0	50.0	50.0	37.5	50.0	37.5	37.5	37.5	37.5
Glyoxylate and dicarboxylate metabolism	58	37.9	36.2	37.9	37.9	37.9	37.9	37.9	37.9	37.9	37.9
Histidine metabolism	37	32.4	32.4	35.1	35.1	32.4	32.4	32.4	32.4	32.4	35.1
Inositol phosphate metabolism	40	22.5	10.0	10.0	10.0	12.5	10.0	12.5	10.0	10.0	12.5
Lipopolysaccharide biosynthesis	22	86.4	86.4	86.4	86.4	86.4	86.4	86.4	86.4	86.4	86.4
Lysine biosynthesis	31	51.6	54.8	51.6	51.6	51.6	51.6	51.6	51.6	54.8	54.8
Lysine degradation	54	24.1	18.5	25.9	25.9	18.5	18.5	18.5	18.5	24.1	25.9
Methane metabolism	33	27.3	27.3	27.3	27.3	27.3	27.3	27.3	27.3	27.3	27.3
N-Glycan biosynthesis	29	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4
Nicotinate and nicotinamide metabolism	47	40.4	38.3	38.3	38.3	40.4	38.3	40.4	38.3	38.3	40.4
One carbon pool by folate	24	54.2	54.2	54.2	54.2	54.2	54.2	54.2	54.2	54.2	54.2
Pantothenate and CoA biosynthesis	28	57.1	53.6	53.6	53.6	57.1	57.1	57.1	57.1	57.1	57.1
Pentose and glucuronate interconversions	56	46.4	46.4	46.4	50.0	50	46.4	44.6	46.4	50	50
Pentose phosphate pathway	37	64.9	64.9	64.9	64.9	64.9	64.9	64.9	64.9	64.9	64.9
Phenylalanine metabolism	51	19.6	17.6	17.6	17.6	21.6	17.6	21.6	17.6	17.6	25.5
Phenylalanine, tyrosine and tryptophan biosynthesis	31	71.0	71.0	71.0	71.0	71.0	71.0	71.0	71.0	71.0	71.0
Phenylpropanoid biosynthesis	30	20.0	23.3	23.3	23.3	20.0	20.0	20.0	20.0	20.0	20.0
Porphyrin and chlorophyll metabolism	66	33.3	33.3	33.3	33.3	34.8	33.3	33.3	33.3	33.3	33.3
Primary bile acid biosynthesis	19	5.3	5.3	5.3	5.3	5.3	5.3	5.3	5.3	5.3	5.3
Propanoate metabolism	47	38.3	38.3	40.4	40.4	36.2	36.2	36.2	36.2	36.2	42.6
Purine metabolism	104	55.8	56.7	56.7	56.7	55.8	55.8	55.8	55.8	55.8	55.8
Pyrimidine metabolism	64	48.4	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Pyruvate metabolism	64	50.0	51.6	51.6	51.6	50.0	50.0	50.0	50.0	50.0	53.1
Reductive carboxylate cycle (CO2 fixation)	13	69.2	69.2	69.2	69.2	69.2	69.2	69.2	69.2	69.2	69.2
Riboflavin metabolism	16	68.8	68.8	68.8	68.8	68.8	68.8	68.8	68.8	68.8	68.8
Sphingolipid metabolism	29	27.6	31.0	31.0	31.0	27.6	27.6	27.6	27.6	27.6	27.6
Steroid biosynthesis	27	7.4	7.4	7.4	7.4	7.4	7.4	7.4	7.4	7.4	7.4

Table 6. (cont.)

		(%) = The number of distinct ECs for pathway x in a given organism									
		Total number of distinct ECs in the same pathway x									
KEGG metabolic pathway	**Distinct ECs**	**ED1a (B2)**	**536 (B2)**	**APEC-O1 (B2)**	**CFT073 (B2)**	**H001 (B2)**	**H223 (B2)**	**M605 (B2)**	**S88 (B2)**	**TA464 (B2)**	**H299 (?)**
Sulfur metabolism	30	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
Synthesis and degradation of ketone bodies	6	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
Terpenoid backbone biosynthesis	27	44.4	48.1	48.1	48.1	44.4	44.4	44.4	44.4	44.4	44.4
Thiamine metabolism	16	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Tyrosine metabolism	63	17.5	17.5	17.5	17.5	17.5	17.5	17.5	17.5	17.5	19.0
Ubiquinone and other terpenoid-quinone biosynthesis	25	56.0	64.0	64.0	64.0	56.0	56.0	56.0	56.0	56.0	56.0
Valine, leucine and isoleucine biosynthesis	18	72.2	72.2	72.2	72.2	72.2	72.2	72.2	72.2	72.2	72.2
Valine, leucine and isoleucine degradation	34	26.5	26.5	32.4	29.4	26.5	26.5	26.5	23.5	26.5	29.4
Vitamin B6 metabolism	26	30.8	30.8	30.8	30.8	30.8	30.8	30.8	30.8	30.8	30.8

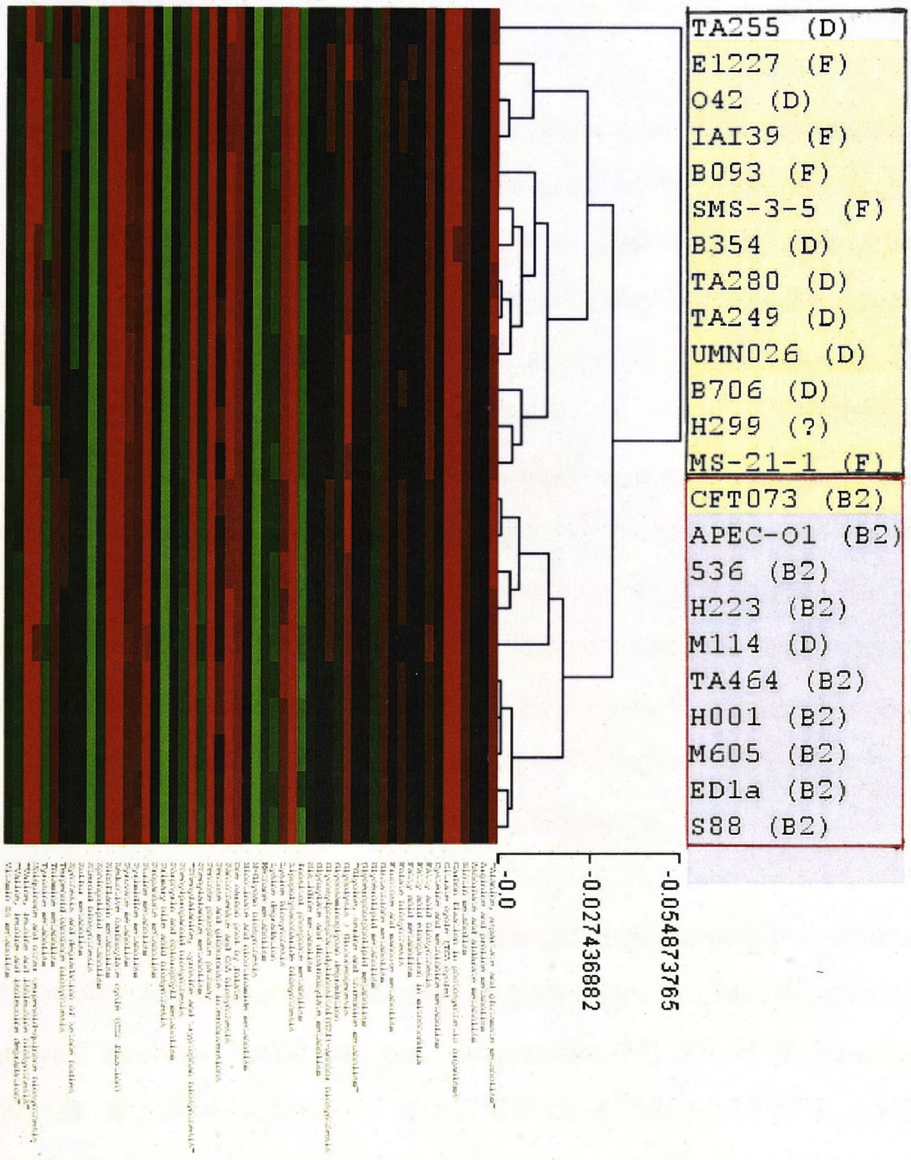


Figure 5. The hierarchical clustering of metabolic profiles of 23 *E. coli* genomes. *E. coli* strains were clustered based on the pathway completion values using the MeV software. Highlighted in yellow are the F, D, and unknown phylo-group strain H299. Highlighted in blue are the B2 strains and the D strain M114. Each row corresponds to the 23 *E. coli* strains and each column represents one pathway.

5.4 Discussion and Conclusion

Based on MLST analysis using the complete sequences of 22 housekeeping genes, the newly environmental *E. coli* E1227 was assigned to be a membership of the phylo-group F. This phylo-group F has been recently described as an additional group of *E. coli* strains and suggested to be a sister group of phylo-group B2 strains known to be responsible for extra-intestinal infection by the phylogenetic analysis based on gene sequences (Jaureguy et al., 2008, Clermont et al., 2011, Chaudhuri et al., 2010, Clermont et al., 2013). However, the evolutionary tree based on the metabolic distance showed that metabolic networks of phylo-group F strains are closely related to strains belonging to phylo-group D which also known to be responsible for extra-intestinal infection (Vieira et al., 2011). In this study, the result of the phylogenetic relationships and metabolic profile diversity based on whole-genome sequence analysis clearly indicates that phylo-group F and D strains appear to be sister groups, while phylo-group B2 strains are more divergent from those F and D strains.

A factor that might affect the ability to distinguish phylo-group F strains from phylo-groups D and/or B2 is probably because of genes used for investigation. The different phylogenetic analysis based on gene sequences: MLST of 8 housekeeping genes (Jaureguy et al., 2008, Clermont et al., 2011), a set of 2,173 *E. coli* K-12 genes (Chaudhuri et al., 2010), and MLST of 13 housekeeping genes (Clermont et al., 2013) revealed that phylo-group F strains were more closely related to phylo-group B2 strains. However the other phylogenetic analysis using a different set of investigated genes assigned these suspicious strains (later assigned to phylo-group F) to phylo-group D (Fricke et al., 2008, Jaureguy et al., 2008, Touchon et al., 2009).

The result in this study indicates that the diversity of *E. coli* strains is largely driven by gene acquisition and loss event. However the nature of gain and loss of genes is still unclear. Although the ability of the extended quadruplex PCR phylo-group assignment method, using a set of 4 genes: *arpA*, *chuA*, *yjaA*, and TspE4.C2 enables an *E. coli* isolate to be assigned to one of the eight phylo-groups (A, B1, B2, C, D, E, F, and *Escherichia* cryptic clade I) (Clermont et al., 2013). However there appear to be the exception for a small fraction of strains to be correctly assigned to the appropriate phylo-group due to the result of the gain and loss of genes and recombination event

(Clermont et al., 2013). In this study, the *arpA* (regulator of acetyl CoA synthetase) gene, for example, was absent among phylo-group F and B2 strains but was present in strains belonging to phylo-group D (Table 5). A previous study has reported that *arpA* is a gene common in all phylo-group A and B1 strains (Clermont et al., 2004). Therefore the presence of *arpA* in phylo-group A, B1, and D strains emphasis the relationships among these strains as phylo-group D is described as a sister group to phylo-group A and B1, which they are thought to be one clade (Lecointre, 1998). However, based on the extended quadruplex method, *arpA* could not be detected in a few D strains (reptile and environmental strains) due to total deletion of *arpA* (Clermont et al., 2013). Evidence has suggested that the absence of *arpA* is strongly linked to neonatal meningitis in human host caused by extraintestinal-virulence *E. coli* strains (Clermont et al., 2004). This is a reasonable explanation for the absence of *arpA* from phylo-group B2 strains known to be responsible for extraintestinal disease. While the absence of *arpA* from phylo-group F strains might imply the propensity to cause neonatal meningitis of the strains; however, it is still unclear. In addition, the lack of *arpA* in phylo-group D strains: reptile and environmental isolates (Clermont et al., 2013) may suggest that host specificity might play an important role in genome content of these isolates.

As the result of the gain and loss of genes and recombination event may yield misleading results to assign phylo-group of strains correctly when using a set of genes or MLST data. The phylogenetic relationships based on whole-genome sequence analysis is being the most sophisticated way of inferring the phylogenetic relationships of *E. coli* known to be a highly genetically diverse species. The analysis using whole-genome dataset provides a large amount of informative genome data that can overcome the impact of recombination signals that may occur when using the less number of individual genes or MLST which are more susceptible to the effect of recombination (Chaudhuri and Henderson, 2012, Touchon et al., 2009) .

In this study, 5 F strains including 2 human isolates, 2 environmental isolates, and 1 bird isolate were investigated. Among current members of phylo-group F, human strain IAI39 represents a clinical condition as extraintestinal pathogenic *E. coli* (ExPEC) (Touchon et al., 2009). The strains SMS-3-5 and E1227 are both environmental isolates; however, it is likely that they differ in their propensity to cause disease.

The strain SMS-3-5 is unlikely to present a risk to public health (Fricke et al., 2008), while the strain E1227 was found to carry putative virulence genes (i.e. *malX, papA, ibe, ompT, iss, astA, sitABCD;* data not shown). However, as a few numbers of environmental F strains are available, it is difficult to draw a conclusion about propensity to cause disease of the environmental F strains. Therefore, in order to truly understand the pathogenesis of strains belonging to phylo-group F from various sources (humans, animals, and environment), a comparison among more F strains must be conducted.

In conclusion, the phylogenetic relationship and metabolic profile diversity based on the whole-genome analysis indicate that phylo-group F, the recently described group of *E. coli* strains, is very closely related to phylo-group D strains know to be responsible for extra-intestinal infection. The function of genes unique to strains belonging to phylo-group F in this study is still unknown. Some of genes conserved among F strains (absent in D, B2, and H299) were not restricted to phylo-group F as they are also present in strains belonging to phylo-groups A, B1, and E (data not shown). However, for the most of the *E. coli* strains, the balance of the genome consists of genes that are unique to each strain. The diversity of these *E. coli* strains is largely driven by gene acquisition and loss events. The host/source might have an important role in the variable genome shared by F strains, which might also involve in the propensity of strains to cause disease. However, as a few F strains included in this study, the relationships among the phylo-group membership, host specificity, and the pathogenesis of strains are still unclear. Therefore, the study would be conducted with more strains: human, animal, and environmental strains, so that these factors can be identified.

CHAPTER 6

Conclusion and Future Directions

In this research, a variety of bioinformatics tools were used to investigate the genetic diversity and reconstruct the evolutionary history of *E. coli* derived from variety of sources: humans, animals, and environment based on whole genome-scale analysis. The thesis includes three main themes: 1. Distribution of extra-intestinal virulence traits among *E. coli* isolated from native Australian vertebrates with those isolated from humans living in Australia, 2. Investigation of the evolution of conjugative plasmids in *E. coli* and their changing role in *E. coli* ecology, and 3. Genetic and metabolic characteristics of phylo-group F, with a specific objective for each chapter. The outcomes of this research lead to significant advances in our understanding presented in *E. coli* species.

6.1 Summary of major findings

Chapter 3: Distribution of Extra-intestinal Virulence Traits among *E. coli* Isolated from Native Australian Vertebrates with Those Isolated from Humans Living in Australia

The frequency and distribution of some traits associated with PAIs were found to be significantly correlated with and concentrated in phylo-group B2 strains. The difference of virulence gene profiles among *E. coli* B2 strains varies with the source of isolation, humans *versus* animals. Among B2 strains, traits typically associated with PAIs are absent or very rare in animal isolates. The frequency observed might be greatly determined by the relative abundance of particular STs that are very common in human strains. High recombination rates were observed in B2 strains suggesting that this is a considerable evolutionary adaptation for attaining virulence and that there were distinct differences in virulence gene profiles between human B2 strains and the small subset of animal B2 strains.

Chapter 4: **Investigation of the Evolution of Conjugative Plasmids in *E. coli* and Their Changing Role in *E. coli* Ecology**

Conjugative plasmids found in *E. coli* are very diverse. The majority of plasmids in *E. coli* belong to the RepFIB or RepFIIA (IncF) backbone types, while some of them belong to the RepI1 (IncI1) group. IncI1 plasmids are found to be more homogeneous and genetically conserved than IncF plasmids. In this study, particularly for the large group of conjugative plasmids in the IncF group, the data strongly suggest that there are no such things as plasmid species. As in essence there is no core genome shared by IncF plasmids. Overall for most plasmids the balance of the genome consists of genes that are unique each plasmid. Conjugative plasmids: key agents in the adaptation of *E. coli* populations have changed their role as mediators of intra- and interspecies interactions to become associated with *E. coli* virulence.

Chapter 5: **Genetic and Metabolic Characteristics of Phylo-group F Strains**

The comparative genomic approach was used to characterize phylogenetic relationship and metabolic profile diversity of the phylo-group F, the recently described group of *E. coli* strains. In comparison to phylo-groups B2 and D, strains belonging to phylo-group F are very closely related to phylo-group D strains know to be responsible for extra-intestinal infection. Most of genes unique to phylo-group F strains had known function. Whilst some of genes conserved among F strains (absent in D, B2, and H299) were present in strains belonging to phylo-groups A, B1, and E (data not shown). However, the balance of the genome consists of genes that are unique to each strain. The diversity of these *E. coli* strains is largely driven by gene acquisition and loss events. The host/source might have an important role in the variable genome shared by F strains, which might also involve in the propensity to cause disease of strains.

6.2 Future work

Due to the limited number of representative *E. coli* plasmids and also *E. coli* strains belonging to phylo-group F from a variety of sources, many aspects of the species are yet to be truly understood. The evolution of ColIb plasmids with the coassociation of other bacteriocins is still unclear, as only one representative of the ColIbMBimm plasmid possessing the RepI1 backbone was found in this study. Therefore, more representative plasmids are required. To better understand the evolution and the coassociation of ColIbMBimm plasmids, a comparison among these plasmids at the whole-genome level (i.e., phylogenetic analysis) and a statistical analysis (i.e., genotypic frequency analysis) would be conducted.

Moreover, as a few strains belonging to phylo-group F from various sources (humans, animals, and environment) are available, it is difficult to draw a conclusion about propensity to cause disease of strains belonging to phylo-group F. The absence of *arpA* (regulator of acetyl CoA synthetase) from F strains as well as B2 strains, which known to be responsible for extraintestinal infections, might imply the propensity to cause extraintestinal disease of F strains similar to B2 strains; however, it is still unclear. Therefore more representative *E. coli* strains belonging to phylo-group F from various sources (humans, animals, and environment) are required for the statistical analysis to be performed. In addition, the total deletion of *arpA* in phylo-group D strains: reptile and environmental ioslates is questioning the role of host specificity in genome content of these isolates. Thanks to the rapid genome sequencing and the evolving comparative tools, the identification of virulence-associated genes among these strains using a comparative pathogenomic approach (Lehmann et al., 2013) would be conducted. This could lead to a truly understanding of pathogenesis of *E. coli* strains belonging not only to phylo-group F but also to phylo-groups B2 and D.

6.3 Broad implications to *E. coli* evolution, diversity and ecology

Variation in the genomic changes, that underlie evolutionary adaptation, is subject to many influences and complications (Barrick et al., 2009). Many plasmids types (Inc groups) are known to occur among *E. coli* strains and they play an important role in the adaptation of bacterial populations (Frost et al., 2005). The results from this study shed light on the evolution of bateriocin plasmids in the IncF and IncI1 groups. These bacteriocin plasmids are changing their role from conferring the competitive advantage in microbial populations to conferring the selective advantage particularly for the pathogenic potency in relation to their hosts. This genomic-based comparative approach wil provivd a guideline for elucidating reletionships between gene content of these plasmids and adaptation to the ecological niche of their host (de Muinck et al., 2013).

The extent of genomic diversity existing in *E. coli* has been the subject to debate (Jackson et al., 2011). Based on the recent method: the extended quadruplex PCR phylo-group assignment, *E. coli* are now assigned to 8 phylo-groups including A, B1, B2, C, D, E, F, and *Escherichia* cryptic clade I (Clermont et al., 2013). The results in this study, however, provide novel insight into the phylogenetic relationship and metabolic profile diversity of the additional phylo-group F. Although the propensity to cause disease of F strains in still unclear, the absence of *arpA* from F strains as well as B2 strains might imply the evolution of *E. coli* pathogenesis between these two phylo-groups. Moreover the study provides information about the frequency and distribution of some virulence traits that are very common in human isolates but absent or very rare in animal isolates. In combination, these data (genomes, genes, and metabolic profiles) will enable a comprehensive analysis of genome-wide measurements (Yoon et al., 2012) to determine the relationships among the phylo-group membership, host specificity, and the pathogenesis in *E. coli.*

In addition, conjugative plasmids in this study (the second theme) will be additional representative of the ancestral bacteriocin plasmids and the coassociation of bacteriocins to be found in *E. coli*. Similarly, the newly sequenced environmental strain E1227 (the third theme) also will be a significant addition of the strain belonging to phylo-group F to be found in *E. coli* as there is a few of F strains currently available in the database. Therefore the outcomes of this study will provide a guideline for addressing and elucidating a range of questions concerning *E. coli* species.

BIBLIOGRAPHY

AMABILE-CUEVAS, C. F. & CHICUREL, M. E. 1992. Bacterial plasmids and gene flux. *Cell,* 70, 189-199.

AZIZ, R., BARTELS, D., BEST, A., DEJONGH, M., DISZ, T., EDWARDS, R., FORMSMA, K., GERDES, S., GLASS, E., KUBAL, M., MEYER, F., OLSEN, G., OLSON, R., OSTERMAN, A., OVERBEEK, R., MCNEIL, L., PAARMANN, D., PACZIAN, T., PARRELLO, B., PUSCH, G., REICH, C., STEVENS, R., VASSIEVA, O., VONSTEIN, V., WILKE, A. & ZAGNITKO, O. 2008. The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics,* 9, 75.

BAHRANI-MOUGEOT, F. K., BUCKLES, E. L., LOCKATELL, C. V., HEBEL, J. R., JOHNSON, D. E., TANG, C. M. & DONNENBERG, M. S. 2002. Type 1 fimbriae and extracellular polysaccharides are preeminent uropathogenic *Escherichia coli* virulence determinants in the murine urinary tract. *Molecular Microbiology,* 45, 1079-1093.

BARBIERI, N. L., TEJKOWSKI, T. M., DE OLIVEIRA, A. L., DE BRITO, B. G. & HORN, F. 2012. Characterization of Extraintestinal *Escherichia coli* Isolated from a Peacock (*Pavo cristatus*) With Colisepticemia. *Avian Diseases Digest,* 7, e64-e65.

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, 1243-1247.

BERGSTROM, C. T., LIPSITCH, M. & LEVIN, B. R. 2000. Natural Selection, Infectious Transfer and the Existence Conditions for Bacterial Plasmids. *Genetics,* 155, 1505-1519.

BERGTHORSSON, U. & OCHMAN, H. 1998. Distribution of chromosome length variation in natural isolates of *Escherichia coli. Mol Biol Evol,* 15, 6-16.

BIDET, P., METAIS, A., MAHJOUB-MESSAI, F., DURAND, L., DEHEM, M., AUJARD, Y., BINGEN, E., NASSIF, X. & BONACORSI, S. 2007. Detection and Identification by PCR of a Highly Virulent Phylogenetic Subgroup among Extraintestinal Pathogenic *Escherichia coli* B2 Strains. *Appl. Environ. Microbiol.,* 73, 2373-2377.

BONACORSI, S., BIDET, P., MAHJOUB, F., MARIANI-KURKDJIAN, P., AIT-IFRANE, S., COURROUX, C. & BINGEN, E. 2009. Semi-automated rep-PCR for rapid differentiation of major clonal groups of *Escherichia coli* meningitis strains. *International Journal of Medical Microbiology,* In Press, Corrected Proof.

BOYD, E. F. & HARTL, D. L. 1998. Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a phylogenetically clustered distribution. *J Bacteriol,* 180, 1159–1165.

BRUEN, T. C., PHILIPPE, H. & BRYANT, D. 2006. A Simple and Robust Statistical Test for Detecting the Presence of Recombination. *Genetics,* 172, 2665-2681.

CASCALES, E., BUCHANAN, S. K., DUCHE, D., KLEANTHOUS, C., LLOUBES, R., POSTLE, K., RILEY, M., SLATIN, S. & CAVARD, D. 2007. Colicin Biology. *Microbiol. Mol. Biol. Rev.,* 71, 158-229.

CAUGANT, D. A., LEVIN, B. R., ORSKOV, I., ORSKOV, F., SVANBORG EDEN, C. & SELANDER, R. K. 1985. Genetic diversity in relation to serotype in *Escherichia coli. Infection and Immunity,* 49, 407-413.

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., FRANKEL, G., TANG, C. M., DUDLEY, E. G., ROBERTS, I. S., RASKO, D. A., PALLEN, M. J., PARKHILL, J., NATARO, J. P., THOMSON, N. R. & HENDERSON, I. R. 2010. Complete Genome Sequence and Comparative Metabolic Profiling of the Prototypical Enteroaggregative *Escherichia coli* Strain 042. *PLoS ONE,* 5, e8801.

CHRISTENSON, J. K. & GORDON, D. M. 2009. Evolution of colicin BM plasmids: the loss of the colicin B activity gene. *Microbiology,* 155, 1645-1655.

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

CLERMONT, O., BONACORSI, S. & BINGEN, E. 2001. The Yersinia high-pathogenicity island is highly predominant in virulence-associated phylogenetic groups of *Escherichia coli*. *FEMS Microbiol Lett,* 196, 153-157.

CLERMONT, O., BONACORSI, S. & BINGEN, E. 2004. Characterization of an Anonymous Molecular Marker Strongly Linked to Escherichia coli Strains Causing Neonatal Meningitis. *Journal of Clinical Microbiology,* 42, 1770-1772.

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

CLERMONT, O., OLIER, M., HOEDE, C., DIANCOURT, L., BRISSE, S., KEROUDEAN, M., GLODT, J., PICARD, B., OSWALD, E. & DENAMUR, E. 2011. Animal and human pathogenic *Escherichia coli* strains share common genetic backgrounds. *Infection, Genetics and Evolution,* 11, 654-662.

CZARAN, T. L., HOEKSTRA, R. F. & PAGIE, L. 2002. Chemical warfare between microbes promotes biodiversity. *Proceedings of the National Academy of Sciences,* 99, 786-790.

DARLING, A. C. E., MAU, B., BLATTNER, F. R. & PERNA, N. T. 2004. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research,* 14, 1394-1403.

DARLING, A. E., MAU, B. & PERNA, N. T. 2010. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE,* 5, e11147.

DE MUINCK, E., LAGESEN, K., AFSET, J., DIDELOT, X., RØNNINGEN, K., RUDI, K., STENSETH, N. & TROSVIK, P. 2013. Comparisons of infant Escherichia coli isolates link genomic profiles with adaptation to the ecological niche. *BMC Genomics,* 14, 1-21.

DIDELOT, X., DARLING, A. & FALUSH, D. 2008. Inferring genomic flux in bacteria. *Genome Research.*

DIDELOT, X. & FALUSH, D. 2007. Inference of Bacterial Microevolution Using Multilocus Sequence Data. *Genetics,* 175, 1251-1266.

EBERHARD, W. G. 1990. Evolution in bacterial plasmids and levels of selection. *The Quarterly review of biology,* 65, 3-22.

FRICKE, W. F., WRIGHT, M. S., LINDELL, A. H., HARKINS, D. M., BAKER-AUSTIN, C., RAVEL, J. & STEPANAUSKAS, R. 2008. Insights into the

Environmental Resistance Gene Pool from the Genome Sequence of the Multidrug-Resistant Environmental Isolate *Escherichia coli* SMS-3-5. *J. Bacteriol.,* 190, 6779-6794.

FROST, L. S., LEPLAE, R., SUMMERS, A. O. & TOUSSAINT, A. 2005. Mobile genetic elements: the agents of open source evolution. *Nat Rev Micro,* 3, 722-732.

GORDON, D. M. 1992. Rate of plasmid transfer among *Escherichia coli* strains isolated from natural populations. *Journal of General Microbiology,* 138, 17-21.

GORDON, D. M. 2004. "The Influence of Ecological Factors on the Distribution and the Genetic Structure of *Eschercihia coli.*" *Escherichia coli* and *Salmonella typhimurium*: cellular and Molecular Biology. Neidhardt, F. et al. (eds). *American Society for Microbiology, Washington, DC.*

GORDON, D. M. 2010. Strain Typing and the Ecological Structure of *Escherichia coli. AOAC International Journal,* 93, 974-984.

GORDON, D. M. & COWLING, A. 2003. The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. *Microbiology,* 149, 3575-3586.

GORDON, D. M. & O'BRIEN, C. L. 2006. Bacteriocin diversity and the frequency of multiple bacteriocin production in *Escherichia coli. Microbiology,* 152, 3239-3244.

GORDON, D. M., OLIVER, E. & LITTLEFIELD-WYER, J. 2007. The Diversity of Bacteriocins in Gram-Negative Bacteria. *In:* RILEY, M. A. & CHAVAN, M. A. (eds.) *Bacteriocins: Ecology and Evolution* New York: Springer.

GORDON, D. M., OLIVIER, C., HEATHER, T. & ERICK, D. 2008. Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. *Environmental Microbiology,* 10, 2484-2496.

GORDON, D. M., RILEY, M. A. & PINOU, T. 1998. Temporal changes in the frequency of colicinogeny in *Escherichia coli* from house mice. *Microbiology,* 144, 2233-2240.

GORDON, D. M., STERN, S. E. & COLLIGNON, P. J. 2005. Influence of the age and sex of human hosts on the distribution of *Escherichia coli* ECOR groups and virulence traits. *Microbiology,* 151, 15-23.

GUINDON, S., DUFAYARD, J.-F., LEFORT, V., ANISIMOVA, M., HORDIJK, W. & GASCUEL, O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology,* 59, 307-321.

HACKER, J. & KAPER, J. B. 2000. Pathogenicity Islands and the Evolution of Microbes. *Annual Review of Microbiology,* 54, 641-679.

HARKINS, T. & JARVIE, T. 2007. Megogenomics analysis using the genome sequencer FLX system. *Nat Methods,* 4: application notes iii–v.

HUSON, D. H. & BRYANT, D. 2006. Application of Phylogenetic Networks in Evolutionary Studies. *Mol Biol Evol,* 23, 254-267.

ISHII, S., MEYER, K. P. & SADOWSKY, M. J. 2007. Relationship between Phylogenetic Groups, Genotypic Clusters, and Virulence Gene Profiles of *Escherichia coli* Strains from Diverse Human and Animal Sources. *Applied and Environmental Microbiology,* 73, 5703-5710.

JACKSON, S., PATEL, I., BARNABA, T., LECLERC, J. & CEBULA, T. 2011. Investigating the global genomic diversity of Escherichia coli using a multi-genome DNA microarray platform with novel gene prediction strategies. *BMC Genomics,* 12, 1-17.

JAUREGUY, F., LANDRAUD, L., PASSET, V., DIANCOURT, L., FRAPY, E., GUIGON, G., CARBONNELLE, E., LORTHOLARY, O., CLERMONT, O., DENAMUR, E., PICARD, B., NASSIF, X. & BRISSE, S. 2008. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. *BMC Genomics,* 9, 560.

JEZIOROWSKI, A. & GORDON, D. M. 2007. Evolution of Microcin V and Colicin Ia Plasmids in *Escherichia coli. J. Bacteriol.,* 189, 7045-7052.

JOHNSON, J. R. 1991. Virulence factors in *Escherichia coli* urinary tract infection. *Clin. Microbiol. Rev.,* 4, 80-128.

JOHNSON, JAMESÂ R., CLERMONT, O., MENARD, M., KUSKOWSKI, MICHAELÂ A., PICARD, B. & DENAMUR, E. 2006. Experimental Mouse Lethality of *Escherichia coli* Isolates, in Relation to Accessory Traits, Phylogenetic Group, and Ecological Source. *The Journal of Infectious Diseases,* 194, 1141-1150.

JOHNSON, J. R., DELAVARI, P., KUSKOWSKI, M. & STELL, A. L. 2001. Phylogenetic distribution of extraintestinal virulence-associated traits in *Escherichia coli Journal of Infectious Diseases,* 183, 1834-1835.

JOHNSON, J. R., OWENS, K., MANGES, A. R. & RILEY, L. W. 2004. Rapid and Specific Detection of *Escherichia coli* Clonal Group A by Gene-Specific PCR. *J. Clin. Microbiol.,* 42, 2618-2622.

JOHNSON, J. R. & RUSSO, T. A. 2002. Extraintestinal pathogenic *Escherichia coli:* "The other bad E coli". *Journal of Laboratory and Clinical Medicine,* 139, 155-162.

JOHNSON, T. J., JORDAN, D., KARIYAWASAM, S., STELL, A. L., BELL, N. P., WANNEMUEHLER, Y. M., ALARCÓN, C. F., LI, G., TIVENDALE, K. A., LOGUE, C. M. & NOLAN, L. K. 2010. Sequence Analysis and Characterization of a Transferable Hybrid Plasmid Encoding Multidrug Resistance and Enabling Zoonotic Potential for Extraintestinal Escherichia coli. *Infection and Immunity,* 78, 1931-1942.

JOHNSON, T. J. & NOLAN, L. K. 2009. Pathogenomics of the Virulence Plasmids of *Escherichia coli. Microbiol. Mol. Biol. Rev.,* 73, 750-774.

KAPER, J. B., NATARO, J. P. & MOBLEY, H. L. T. 2004. Pathogenic *Escherichia coli. Nat Rev Micro,* 2, 123-140.

KAUFFMANN, F. 1947. The Serology of the Coli Group. *The Journal of Immunology,* 57, 71-100.

KIM, S. R., FUNAYAMA, N. & KOMANO, T. 1993. Nucleotide sequence and characterization of the traABCD region of IncI1 plasmid R64. *Journal of Bacteriology,* 175, 5035-5042.

KIRKUP, B. C. & RILEY, M. A. 2004. Antibiotic-mediated antagonism leads to a bacterial game of rock-paper-scissors in vivo. *Nature,* 428, 412-414.

KRČMÉRY, V., FREDERICQ, P., WIEDEMANN, B. & HURWITZ, C. 1971. MOBILIZATION OF EXTRACHROMOSOMAL DETERMINANTS FOR STREPTOMYCIN RESISTANCE BY TRANSFERABLE COLICINOGENIC FACTORS. *Annals of the New York Academy of Sciences,* 182, 118-122.

LE GALL, T., CLERMONT, O., GOURIOU, S., PICARD, B., NASSIF, X., DENAMUR, E. & TENAILLON, O. 2007. Extraintestinal Virulence Is a Coincidental By-Product of Commensalism in B2 Phylogenetic Group *Escherichia coli* Strains. *Mol Biol Evol,* 24, 2373-2384.

LECOINTRE, G., RACHDI, L., DARLU, P. & DENAMUR, E. 1998 *Escherichia coli* Molecular Phylogeny Using the Incongruence Length Difference Test *Molecular Biology and Evolution* 15, 1685-1695.

LEHMANN, J. S., FOUTS, D. E., HAFT, D. H., CANNELLA, A. P., RICALDI, J. N., BRINKAC, L., HARKINS, D., DURKIN, S., SANKA, R., SUTTON, G., MORENO, A., VINETZ, J. M. & MATTHIAS, M. A. 2013. Pathogenomic Inference of Virulence-Associated Genes in *Leptospira interrogans*. *PLoS Negl Trop Dis,* 7, e2468.

LESK, A. M. 2005. *Introduction to Bioinformatics,* New York, Oxford University Press Inc.

LEYTON, D. L., SLOAN, J., HILL, R. E., DOUGHTY, S. & HARTLAND, E. L. 2003. Transfer Region of pO113 from Enterohemorrhagic *Escherichia coli*: Similarity with R64 and Identification of a Novel Plasmid-Encoded Autotransporter, EpeA. *Infect. Immun.,* 71, 6307-6319.

LU, S., ZHANG, X., ZHU, Y., KIM, K. S., YANG, J. & JIN, Q. 2011. Complete Genome Sequence of the Neonatal-Meningitis-Associated *Escherichia coli* Strain CE10. *Journal of Bacteriology,* 193, 7005.

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,* 108, 7200-7205.

MAIDEN, M. C. J., BYGRAVES, J. A., FEIL, E., MORELLI, G., RUSSELL, J. E., URWIN, R., ZHANG, Q., ZHOU, J., ZURTH, K., CAUGANT, D. A., FEAVERS, I. M., ACHTMAN, M. & SPRATT, B. G. 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenicĩ€ microorganisms. *Proceedings of the National Academy of Sciences,* 95, 3140-3145.

MARTINEZ, J. L. & BAQUERO, F. 2002. Interactions among Strategies Associated with Bacterial Infection: Pathogenicity, Epidemicity, and Antibiotic Resistance. *Clinical Microbiology Reviews,* 15, 647-679.

MCVEAN, G., AWADALLA, P. & FEARNHEAD, P. 2002. A Coalescent-Based Method for Detecting and Estimating Recombination From Gene Sequences. *Genetics,* 160, 1231-1241.

MELLATA, M., TOUCHMAN, J. W. & CURTISS, R. 2009. Full Sequence and Comparative Analysis of the Plasmid pAPEC-1 of Avian Pathogenic *Escherichia coli* χ7122 (O78 : K80 : H9). *PLoS ONE* [Online], 4(1): e4232. .

MOHAPATRA, B. R., BROERSMA, K., NORDIN, R. & MAZUMDER, A. 2007. Evaluation of repetitive extragenic palindromic-PCR for discrimination of fecal *Escherichia coli* from humans, and different domestic- and wild-animals. *Microbiol Immunol,* 51 (8), 733-40

MORA, A., LOPEZ, C., DABHI, G., BLANCO, M., BLANCO, J., ALONSO, M., HERRERA, A., MAMANI, R., BONACORSI, S., MOULIN-SCHOULEUR, M. & BLANCO, J. 2009. Extraintestinal pathogenic *Escherichia coli* O1:K1:H7/NM from human and avian origin: detection of clonal groups B2 ST95 and D ST59 with different host distribution. *BMC Microbiology,* 9, 132.

MORENO, E., PLANELLS, I., PRATS, G., PLANES, A. M., MORENO, G. & ANDREU, A. 2005. Comparative study of *Escherichia coli* virulence determinants in strains causing urinary tract bacteremia versus strains causing pyelonephritis and other sources of bacteremia. *Diagnostic Microbiology and Infectious Disease,* 53, 93-99.

NATARO, J. P. 2005. Enteroaggregative *Escherichia coli* pathogenesis. *Current Opinion in Gastroenterology,* 21, 4-8.

NATARO, J. P. & KAPER, J. B. 1998. Diarrheagenic *Escherichia coli. Clin. Microbiol. Rev.,* 11, 142-201.

NOWROUZIAN, F. L., ADLERBERTH, I. & WOLD, A. E. 2006. Enhanced persistence in the colonic microbiota of *Escherichia coli* strains belonging to phylogenetic group B2: role of virulence factors and adherence to colonic cells. *Microbes and Infection,* 8, 834-840.

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

PEAKALL, R. & SMOUSE, P. E. 2006. genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes,* 6, 288-295.

POWER, M. L., LITTLEFIELD-WYER, J., GORDON, D. M., VEAL, D. A. & SLADE, M. B. 2005. Phenotypic and genotypic characterization of encapsulated *Escherichia coli* isolated from blooms in two Australian lakes. *Environmental Microbiology,* 7, 631-640.

RILEY, M. A. & GORDON, D. M. 1992. A survey of Col plasmids in natural isolates of *Escherichia coli* and an investigation into the stability of Col-plasmid lineages. *Journal of General Microbiology,* 138, 1345-1352.

RILEY, M. A. & GORDON, D. M. 1999. The ecological role of bacteriocins in bacterial competition. *Trends in Microbiology,* 7, 129-133.

RILEY, M. A. & WERTZ, J. E. 2002. Bacteriocin diversity: ecological and evolutionary perspectives. *Biochimie,* 84, 357-364.

RISSMAN, A. I., MAU, B., BIEHL, B. S., DARLING, A. E., GLASNER, J. D. & PERNA, N. T. 2009. Reordering contigs of draft genomes using the Mauve Aligner. *Bioinformatics*, btp356.

RUMP, L. V., MENG, J., STRAIN, E. A., CAO, G., ALLARD, M. W. & GONZALEZ-ESCALONA, N. 2012. Complete DNA sequence analysis of EHEC pO157_2 in GUD+ Escherichia coli O157:H7 reveals a novel evolutionary path. *Journal of Bacteriology.*

SAEED AI, B. N., BRAISTED JC; LIANG W, SHAROV V, HOWE EA, ET AL. 2006. TM4 microarray software suite. *Methods in Enzymology.* , 411, 134-93.

SAEED AI, S. V., WHITE J, LI J, LIANG W, BHAGABATI N, ET AL. 2003. TM4: a free, open-source system for microarray data management and analysis. . 34.

SAMPEI, G.-I., FURUYA, N., TACHIBANA, K., SAITOU, Y., SUZUKI, T., MIZOBUCHI, K. & KOMANO, T. 2010. Complete genome sequence of the incompatibility group I1 plasmid R64. *Plasmid,* 64, 92-103.

SAVAGEAU, M. A. 1983. *Escherichia coli* Habitats, Cell Types, and Molecular Mechanisms of Gene Control *The American Naturalist.* , 22(6), 732-744.

SCHUBERT, S., PICARD, B., GOURIOU, S., HEESEMANN, J. & DENAMUR, E. 2002. Yersinia highpathogenicity island contributes to virulence in *Escherichia coli* causing extraintestinal infections. *Infect Immun,* 70, 5335–5337.

SCHUBERT, S. R., DARLU, P., CLERMONT, O., WIESER, A., MAGISTRO, G., HOFFMANN, C., WEINERT, K., TENAILLON, O., MATIC, I. & DENAMUR, E. 2009. Role of Intraspecies Recombination in the Spread of Pathogenicity Islands within the *Escherichia coli* Species. *PLoS Pathog,* 5, e1000257.

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, 873-884.

SMILLIE, C., GARCILLÁN-BARCIA, M. P., FRANCIA, M. V., ROCHA, E. P. C. & DE LA CRUZ, F. 2010. Mobility of Plasmids. *Microbiology and Molecular Biology Reviews,* 74, 434-452.

SMITH, W. H. & HUGGINS, M. B. 1976. Further Observations on the Association of the Colicine V Plasmid of *Escherichia coli* with Pathogenicity and with Survival in the Alimentary Tract. *Journal of General Microbiology,* 92, 335-350.

SUMMERS, D. K. 1996. *The Biology of Plasmids,* Oxford, Blackwell Science Ltd. .

TAMURA, K., DUDLEY, J., NEI, M. & KUMAR, S. 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution,* 24, 1596-1599.

TAMURA, K., PETERSON, D., PETERSON, N., STECHER, G., NEI, M. & KUMAR, S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution,* 28, 2731-2739.

TENAILLON, O., SKURNIK, D., PICARD, B. & DENAMUR, E. 2010. The population genetics of commensal *Escherichia coli. Nat Rev Micro,* 8, 207-217.

TOUCHON, M., HOEDE, C., TENAILLON, O., BARBE, V., BAERISWYL, S., BIDET, P., BINGEN, E., BONACORSI, S., BOUCHIER, C., BOUVET, O., CALTEAU, A., CHIAPELLO, H., CLERMONT, O., CRUVEILLER, S., DANCHIN, A., DIARD, M., DOSSAT, C., KAROUI, M. E., FRAPY, E., GARRY, L., GHIGO, J. M., GILLES, A. M., JOHNSON, J., LE BOUGUENEC, C., LESCAT, M., MANGENOT, S., MARTINEZ-JEHANNE, V., MATIC, I., NASSIF, X., OZTAS, S., PETIT, M. A., PICHON, C., ROUY, Z., RUF, C. S., SCHNEIDER, D., TOURRET, J., VACHERIE, B., VALLENET, D., MEDIGUE, C., ROCHA, E. P. C. & DENAMUR, E. 2009. Organised Genome Dynamics in the *Escherichia coli* Species Results in Highly Diverse Adaptive Paths. *PLoS Genet,* 5, e1000344.

VERSALOVIC, J., KOEUTH, T. & LUPSKI, R. 1991. Distribution of repetitive DNA sequences in eubacteria and application to finerpriting of bacterial enomes. *Nucleic Acids Research,* 19, 6823-6831.

VIEIRA, G., SABARLY, V., BOURGUIGNON, P.-Y., DUROT, M., LE FÈVRE, F., MORNICO, D., VALLENET, D., BOUVET, O., DENAMUR, E., SCHACHTER, V. & MÉDIGUE, C. 2011. Core and Panmetabolism in Escherichia coli. *Journal of Bacteriology,* 193, 1461-1472.

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. Appl. Environ. Microbiol.,* 75, 6534-6544.

WELCH, R. 2006. The Genus *Escherichia. The Prokaryotes.*

YOON, S., HAN, M.-J., JEONG, H., LEE, C., XIA, X.-X., LEE, D.-H., SHIM, J., LEE, S., OH, T. & KIM, J. 2012. Comparative multi-omics systems analysis of Escherichia coli strains B and K-12. *Genome Biology,* 13, R37.

CURRICULUM VITAE

Name: Phataraporn Khumphai

Birth Place: Suphan Buri, Thailand

Academic Backgroud:

2000 - 2004 Master of Science (Genetic Engineering)
Interdisciplinary Graduate Program, Kasetsart University, Thailand
Thesis: *Cloning and Nucleotide Sequencing of the Protease Encoding Gene from Halotolerant Bacterium, Pseudoalteromonas flavipulchra MFKU126*

1996 - 2000 Bachelor of Science (Biotechnology)
Faculty of Science and Technology, Thammasat University, Thailand
Senior project: *Identification of Caster Bean varieties by Random Amplified Polymorphic DNA (RAPD)*

Professional Experience:

Aug 2005 - present Lecturer, Department of Biotechology, Faculty of Science and Technology, Thammasat University, Thailand

Dec 2004 - Jul 2005 Researcher (Molecular Laboratory and Microarray Laboratory), Research Centre, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Thailand