COMPARATIVE GENOMICS OF SELECTED SPECIES OF GRAM-NEGATIVE BACTERIA

by

CHUAN-PENG REN

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Division of Immunity and Infection
The Medical School
The University of Birmingham
June 2009

ABSTRACT

Investigation of genomic diversity can provide insight into the evolutionary history of bacterial species. However, complete genome sequencing is not yet practical for large strain collections at the beginning of this project. In this project PCR-based methods to investigate the genomic diversity of non-sequenced strains were successfully developed. In *Escherichia coli*, the distribution of two Type III secretion system, ETT2 (*E. coli* Type Three Secretion 2) and Flag-2 (*E. coli* Flagellar system 2), were surveyed among a collection of 79 strains. Remnants of both clusters were found to be present in most strains, suggesting that both have a long evolutionary history within *E. coli*.

The PCR-based methods were also developed for application as part of genome sequencing projects. They were used to explore the co-linear and variable regions between *Campylobacter jejuni* M1 and the genome sequenced strain *C. jejuni* strain 11168. The *C. jejuni* M1 genome was assembled into thirty-four genomic contigs relative to strain 11168, and the size and position of insertions/deletions were characterised. Similar methods were used to facilitate the finishing of the genome of *Francisella tularensis* strain FSC198, using sequence information from strain Schu S4. The completed genome sequence of FSC198 showed it to be almost identical to that of Schu S4. The two genomes differ at only 11 loci, eight SNPs (single nucleotide polymorphisms) and 3 VNTRs (variable number tandem repeats). This surprising finding suggested that the European isolate FSC198 may be derived from the US laboratory strain Schu S4.

Two virulence factors, IglA and IglB, from a pathogenicity island of strain FSC198 were further investigated and found to interact at the protein level. These proteins are possibly involved in Type VI secretion, and may represent potential vaccine candidates.

DEDICATION

I dedicate this thesis to my dear husband, Jia, for loving me through all the joys and disappointments of writing this thesis.

"Two are better than one, because they have a good return for their work."

Ecclesiastes 4:9

ACKNOWLEDGEMENTS

I thank the BBSRC for funding the TP-PCR *Escherichia coli* project, and the HPA (Porton Down) for funding the *Francisella tularensis* FSC198 genome sequencing project. I thank the Professors Duncan Maskell and Diane Newell, for their collaboration on the *Campylobacter jejuni* project.

I am particularly thankful to my supervisor, Professor Mark Pallen, for initiating my interest in research, and his support and guidance throughout this PhD project.

I am greatly indebted to my friend Dr. Roy Chaudhuri, for giving me generously his time, helpful advice and expert knowledge, and for proof-reading and correcting this entire thesis.

For helpful advice on practical work and invaluable friendship in the laboratory, I would like to thank my colleagues: Lihong Zhang, Scott Beaston, Sophie Matthews, Martin Antonio, Helen Betts, Arshad Khan, Rob Shaw, Lewis Bingle, Ian Henderson, Anthony Scott-Tucker, and Rasha Younis.

A special thanks to those who have supported me with their prayers and encouragement especially during difficult times. I am appreciative of my church members both in Birmingham Chinese Evangelical Church and Guildford Chinese Christian Fellowship.

I thank Dr. Georgina Lloyd for proof-reading this thesis.

My greatest thanks is reserved to my parents, Fude Ren and Weizhen Shan, for their immeasurable and invaluable love and support. Thank you, my beloved dad and mum. Without you, all this might not have been possible.

DECLARATION

I declare I myself carried out all the experiments described in this thesis, except where indicated in the text. The work took place in the Division of Immunity and Infection, the Medical School, the University of Birmingham.

I acknowledge the contribution from Dr. Chaudhuri on *Francisella tularensis* FSC198 genome sequencing project listed in the Chapter Five. Dr. Chaudhuri designed the primers for both whole-genome PCR scanning and finishing stage (listed in Appendix 1 and 2). Professor Pallen and Dr. Chaudhuri analysed the sequencing data on those sites of SNPs and VNTRs and confined the final differences. I declare I performed the experiments on whole genome PCR scanning, PCR and sequencing for both gap closure and genome finishing in that project.

I acknowledge the following figures/tables present in this thesis and in the relevant publications were from other hands:

- Figure 3-1 “Schematic representation of genomic analysis of ETT2 cluster” was reproduced with permission from an original by Dr. Chaudhuri.
- Table 3-1 “Lists of ETT2 genes” was reproduced with permission from an original by Dr. Chaudhuri.
- Figure 3-2 “Schematic representation of structure of the Eip island in EAEC strain 042” was reproduced with permission from an original by Dr. Chaudhuri.
- Table 3-2 “Gene within the *E. coli* Flag-2 gene cluster” was reproduced with permission form an orginal by Dr. Beatson.

- Figure 3-3 “Schematic representation of the Flag-2 gene cluster” was reproduced with permission from an original by Dr. Beatson.
- Figure 3-5 “The strategy of Tiling-Path PCR” was reproduced with permission from an original by Professor Pallen.
- Table 5-2 “SNPs differences between the FSC198 and Schu S4 genome sequence” documents work performed by Dr. Chaudhuri.
- Table 5-3 “VNTRs differences between the FSC198 and Schu S4 genome sequence” documents work performed by Dr. Chaudhuri.
- Figure 5-4 “Circular representation of the complete genome sequence of FSC198” was reproduced from http://xBase.bham.ac.uk.

Some of the work in Chapter Three and Chapter Five has been published in the following papers:

Ren, C. P., Chaudhuri, R. R., Fivian, A., Bailey, C. M., Antonio, M., Barnes, W. M. & Pallen, M. J. (2004). The ETT2 gene cluster, encoding a second type III secretion system from *Escherichia coli*, is present in the majority of strains but has undergone widespread mutational attrition. *Journal of Bacteriology* **186**, 3547-3560

Ren, C. P., Beatson, S. A., Parkhill, J. & Pallen, M. J. (2005). The Flag-2 locus, an ancestral gene cluster, is potentially associated with a novel flagellar system from *Escherichia coli*. *Journal of Bacteriology* **187**, 1430-1440.

Chaudhuri, R.R., Ren, C.P., Desmond, L., Vincent, G.A., Silman, N.J., Brehm, J.K., Elmore, M.J., Hudson, M.J., Forsman, M., Isherwood, K.E., Gurycová, D., Minton, N.P., Titball, R.W., Pallen, M.J., Vipond, R. (2007). Genome sequencing shows that European isolates of *Francisella tularensis* subspecies *tularensis* are almost identical to US laboratory strain Schu S4. *PLoS ONE* **2**, e352

These papers may be found bound at the back of this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS USED IN THIS THESIS

ABC	ATP-binding cassette
AD	Transcriptional activation domain
A/E	Attaching and effacing
Amp	Ampicillin
APS	Adenosine 5´ phosphosulfate
BD	DNA-binding domain
CDS	Coding sequence
CGH	Comparative genome hybridisation
Chl	Chloramphenicol
COG	Cluster of orthologous group
DAEC	Diffusely adherent *E. coli*
DEC	Diarrhoeagenic *E. coli* Collection
Dot	Defect in organelle trafficking
DR	Direct repeat
DS-PCR	Deletion-scanning PCR
DUF	Domain with unknown function
EAEC	Enteroaggregative *E. coli*
ECOR	*Escherichia coli* Reference Collection
EHEC	Enterohaemorrhagic *E. coli*
EIEC	Enteroinvasive *E. coli*
Eip	*E. coli* invasion proteins
EM	Extracellular milieu

EP	Extracellular polysaccharide
EPEC	Enteropathogenic *E. coli*
ETEC	Enterotoxigenic *E. coli*
ETT2	*E. coli* type three secretion system 2
ExPEC	Extraintestinal pathogenic *E. coli*
Flag-2	*E. coli* second flagellar system
FPI	*Francisella* pathogenicity island
FSC	*Francisella* strain collection
GEIs	Genomic islands
HGT	Horizontal gene transfer
HRP	Horseradish peroxidase
IAHP	IcmF-associated homologous proteins
Icm	Intracellular multiplication
Igl	Intracellular growth locus
In-del	Insertions and deletions
IPTG	Isopropyl-β-D thiogalactopyranoside
IR	Inverted repeat
IS	Insertion sequence elements
LB	Luria-Bertani
LEE	Locus of enterocyte effacement
LOS	Lipooligosaccharide
LR-PCR	Long Range PCR
LVS	Live vaccine strain
MBP	Maltose Binding Protein

MFP	Membrane fusion protein
MGEs	Mobile genomic elements
MLEE	Multi-Locus Enzyme Electrophoresis
MLVA	Multilocus variable number of tandem repeat analysis
NC	Needle complex
NEB	New England Biolabs
NMEC	Neonatal meningitis *E. coli*
OMP	Outer membrane protein
PAIs	Pathogenicity islands
PBS	Phosphate Buffered Saline
PT	Pertussis toxin
SDS-PAGE	Sodium Dodecyl lauryl Sulphate-PolyAcrylamide Gel Electrophoresis
SNPs	Single nucleotide polymorphisms
SP-PCR	Single-Primer PCR
T1SS	Type I secretion system
T2SS	Type II secretion system
T3SS	Type III secretion system
T4SS	Type IV secretion system
T5SS	Type V secretion system
T6SS	Type VI secretion system
Tn	Transposon
Tpase	Tansponsase
TP-PCR	Tiling-path PCR scanning
UPEC	Uropathogenic *E. coli*

VF	Virulence factor
VNTRs	Variable-number tandem repeats
WPGS	Whole-genome PCR scanning

CHAPTER ONE

GENERAL INTRODUCTION

1.1 Bacterial genome evolution

Bacteria must retain their genetic information from one generation to the next. On the other hand, bacteria need to evolve strategies allowing the generation of new genetic diversity in order to survive and adapt to the continually changing environmental conditions and niches in which they live. Therefore, genetic plasticity mirrors different bacterial lifestyles and physiological versatilities (Dobrindt and Hacker, 2001). Three main forces are shaping bacterial genomes: gene gain, gene loss and gene change. The molecular and genetic mechanisms leading to these changes are discussed below.

1.1.1 Point mutations or Single-nucleotide polymorphisms (SNPs)

Point mutations or SNPs represent the smallest scale of bacterial genome change. Evolution caused by point mutations is considered as a slow evolutionary process (Ziebuhr *et al.*, 1999). There are two classes of SNPs, referred to as synonymous and nonsynonymous SNPs. Nonsynonymous SNPs (nsSNPs), as the name suggests, result in a substitution that alter the encoding amino acid and hence provide substrate for evolutionary selection. In contrast, synonymous SNPs (sSNPs) or silent SNPs do not change the protein structure and are evolutionarily neutral or nearly so (Kimura, 1983).

Mutations in structural genes of putative virulence factors can modify or attenuate the encoded proteins and influence their function during pathogenesis. Many of these mutations often enable single bacterial clone to become more pathogenic without the acquisition of additional genes. This mechanism is due to random mutagenesis which offers the bacterium a strong advantage under a selective pressure, and thus represents so-

called pathogenicity-enhancing (or 'pathoadaptive') mutations (Ziebuhr *et al.*, 1999). For example, random mutagenesis in the type 1 fimbrial *fimH* gene of *Salmonella typhimurium* is caused by this mechanism. Mutational variation in the *fimH* gene can produce different increases in cell binding, biofilm formation and host-colonization properties (Boddicker *et al.*, 2002; Weissman *et al.*, 2003).

Although slow-evolving, SNPs help to differentiate those bacteria with little genetic variety. The approach of SNPs detection is an important typing tool for those genetically homologous pathogens, such as *Mycobacterium leprae*, *Mycobacterium tuberculosis*, *Yersinia pestis and Bacillus anthracis*, etc (Pallen and Wren, 2007).

1.1.2 Genetic change and repeat units

The repeat units can be found either dispersed throughout the entire genome or located in one genomic area (van Belkum, 1999a). These repeats are either identical (homogeneous) or vary in their sequences due to point mutations (heterogeneous). The number of the repeat units at a certain locus can vary drastically from strain to strain within a given species, and thus they are so-called "variable number of tandem repeats" (VNTRs). The mechanism that brings about most of the repeat number variability is due to slipped strand misparing (SSM) (Torres-Cruz and van der Woude, 2003). Since stretches of relatively short arrays of repeat units exist, aberrant base-pairing often occurs during replication and leads to the mutations in the type of insertions and deletions (in-del) of their repeat units (van Belkum, 1999b).

The roles of the repeat units have been attributed to many biological functions. Many genes responsible for virulence from pathogenic bacteria contain the repetitive elements. For example, the *Staphylococcus aureus* gene encoding the immunoglobulin-binding protein A harbours a region with repeat units that are 21 nucleotide long (Frenay *et al.*, 1996). In addition, loss or gain of repeats can occur within the promoter region, where the alteration can affect the relative position of flanking promoter components and influences transcription (van Ham *et al.*, 1994). Alternatively, variations in the number of repeat units can occur in a protein-coding region. This change can lead to drastic alterations in gene products and in the expression levels, due to frameshift mutations (van Belkum *et al.*, 1998; Sreenu *et al.*, 2003). Such reversible mutations can also cause phase variations, which can be exploited by bacteria to switch gene and/or protein expression on or off. Phase variation, which imparts greater defensive capability to the pathogens to escape a hostile host environment, is a widespread source of intraspecific genotypic and phenotypic variation (van Belkum *et al.*, 1998; Pallen and Wren, 2007). For example, in *Campylobacter jejuni*, 10-100 of short homopolymeric nucleotide repeats are commonly found in genes encoding the biosynthesis or modification of surface structures (Parkhill *et al.*, 2000). These repeats can lead to slippage during DNA replication, resulting in phage variation of surface properties (Parkhill *et al.*, 2000). Another case can be seen in a thorough analysis of the complete genome sequence of *Neisseria meningitides* MC58 (Tettelin *et al.*, 2000). A repertoire of 82 putative phase variable genes was identified in *N. meningitides* MC58 and more than half of these are homopolymeric tracts (Saunders *et al.*, 2000). Phage variable structures in *Neisseria* are critical for them to adapt to local environmental conditions, for example, capsule confers serum resistance and affects cell interactions; pili and pilus modification affect adhesion; some surface proteins, e.g. PorA, have roles in the formation

of surface pores (Saunders *et al.*, 2000). In fact, for many microorganisms, genetic changes that occur at repeat unit regions often confer a certain degree of selective advantage and aid the pathogens in their adaptation.

Repeat units, if short, are also known as short sequence repeats (SSRs) or micro-satellites. They vary in size, but are generally short sequences of less than 30 bp. These short repeat units can also be used to assess strain relatedness and strain discrimination. Since repeat units evolve at a relatively high speed, they can be used to monitor short term bacterial evolution (van Belkum, 1999b). This is particularly useful in strain discrimination for those pathogenic bacteria that are genetically very homologous. The experimental assessment of the variability in repeat number of different loci is called "multilocus variable number of tandem repeat analysis" (MLVA) (Lindstedt, 2005). MLVA is based on the combination of the polymorphous nature of the VNTRs and the use of PCR methodology (Le Fleche *et al.*, 2001). To date, the availability of more than 600 bacterial whole genome sequences allows for the design of specific PCR primers aimed at the amplification of the repeat and some bordering consensus sequences (van Belkum, 2007). MLVA has been proven as an appropriate way to address the genetic diversity of highly monomorphic species such as *Bacillus anthracis* and *Yersinia pestis* (Le Fleche *et al.*, 2001). And now, MLVA has been developed for typing purposes for several bacterial species, such as *Francisella tularensis*, *Haemophilus influenzae*, *Mycobacterium tuberculosis*, and enteric pathogens like *E. coli* and *Salmonella enterica* etc (Johansson *et al.*, 2000; Johansson *et al.*, 2004).

In addition to short repeat units, some genomes contain long sequence repeats, such as ribosomal gene clusters, for example, the *rrn* operons of *Salmonella enterica* serovar Typhi,

or in horizontally acquired DNA, for example, the duplicated *Francisella* pathogenicity island (Liu and Sanderson, 1995; Nano *et al.*, 2004). It is obvious that sequence multiplicity could increase expression simply through greater gene dosage. This is particularly important for bacterial pathogenesis if the multiplied genes are involved in bacterial virulence. Alternatively, repeated sequences can serve as targets for homologous recombination, as a consequence of genomic rearrangements such as deletions, duplications, and inversions.

1.1.3 Horizontal gene transfer (HGT) and mobile genetic elements (MGEs)

HGT generates rapid and extremely dynamic genomes, rather than evolution through the modification of existing genetic information (Ochman *et al.*, 2000). There are three fundamentally distinct mechanisms by which HGT can occur: transformation, conjugation and transduction. Transformation refers to the process when a cell takes up isolated DNA from the environment and has the potential to transfer DNA between distantly related organisms. Many bacterial species are naturally competent to uptake DNA, such as *Bacillus subtilis*, *Haemophilus influenzae*, and *Neisseria* spp. A second mechanism is conjugation, which is defined as the direct transmission of DNA from one cell to another. In contrast to transformation, which is DNase-sensitive and does not require cell-to-cell contact, conjugation is DNase-resistant and does require cell-to-cell contact. A third mechanism is transduction, which is thought to be phage-mediated transfer of genetic materials. The role of phages will be discussed later.

Bacterial genomes generally consist of stable regions called the “core genome” and variable regions, obtained from the “flexible gene pool” (Hacker *et al.*, 2003). The “core genome” of a certain species has a fairly homogeneous G+C content and codon usage, and often encodes housekeeping functions and carries gene clusters with relatively low mutational capacity (Hacker and Kaper, 2000). In contrast, the” flexible gene pool” represents the total amount of foreign DNAs available for recipient cells. If genes from the “flexible gene pool” were obtained from a source organism with a different mutational bias, their G+C content and codon usage will be different from the rest of the genome. Most often, these foreign genes are carried by MGEs, which are able to move within or between genomes via HGT. The acquisition of sequences from distantly related organisms may confer very new phenotypic characteristics on the recipient cell over a short evolutionary timescale. In the past few years, there has been growing evidence that HGT has played a vital role in the evolution of bacterial genomes (Ochman *et al*., 2000).

MGEs include plasmids, transposons, bacteriophages, integrons, insertion sequence elements and genomic islands. Their size ranges from a few hundred base pairs up to 100 kb (Dobrindt *et al.*, 2004).

1.1.3.1 Insertion Sequence (IS) elements

IS elements are small, genetically compact DNA sequences, normally less than 2.5 kb in length (Mahillon and Chandler, 1998; Mahillon *et al.*, 1999). The overall structure of most IS elements is very similar, and includes a central tansponsase (Tpase) gene flanked by inverted repeat (IR) sequences that exactly define the borders of the elements (Mahillon and Chandler, 1998). Generally, IS elements encode no functions other than their own

translocation, and the central Tpases function to transpose their own elements both within and between genomes (Mahillon and Chandler, 1998; Mahillon *et al.*, 1999). The IRs are short, between 10 and 40 bp, and often contain the promoter for the Tpase gene at one side (Mahillon *et al*., 1999). In addition, IS elements are also flanked by further short (between 2 and 14 bp), duplicated direct repeated sequences. However, these direct repeats (DR) does not belong to IS elements, but arise from duplication in the recipient DNA at the insertion site (Ou *et al.*, 2006). Because of the presence of repeated sequences, IS elements can also be regarded as repetitive sequences that are randomly scattered throughout the bacterial genome. The presence of two copies of an IS element can lead to homologous recombination between them, which results in inversion, deletion or replicon fusion of the region between two IS elements (Siguier *et al.*, 2006b). For this reason, IS elements play an important role in promoting genome rearrangement.

Over 1,600 different ISs have been identified to date (www-is.biotoul.fr) (Siguier *et al*., 2006b). These ISs have been classified into about 20 families, based on similarities and conservation in the sequence of their Tpases, genetic organisation and IRs (Mahillon *et al*., 1999; Siguier *et al.*, 2006a). In addition to chromosomal DNA, IS elements are also commonly found in bacterial plasmids, and are an integral part of many naturally occurring bacterial plasmids (Mahillon *et al*., 1999). The ISs in plasmids function in plasmid transfer and integration into the host chromosome. Furthermore, many antibiotic resistance genes are often encoded on plasmids and spread within bacterial populations with the aid of ISs (Mahillon *et al*., 1999). Thus, although the basic function of ISs is simply translocation, they are involved in other activities to shape genomic variation.

In the post-genomic era, complete genome sequences and genomic comparisons provide advantages in defining IS classification, understanding mechanisms of distribution and identifying their role in evolution. For example, when the first *Francisella* genome Schu S4 was published in 2005, the genome was found to be notable for its large complement of IS elements (Larsson *et al.*, 2005). It is now known that about 1 to 5% of annotated *Francisella* genes are transposase genes (Titball and Petrosino, 2007). Five different types of IS elements (IS*Ftu1*-IS*Ftu5*) were found in the Schu S4 genome, and are conserved in other sequenced *Francisella* genomes such as OSU18 (Petrosino *et al.*, 2006). More surprisingly, IS*Ftu1* belongs to the IS*630* Tc-1 mariner family of transposons, of which about 50 copies were found in the Schu S4 genome and 58 copies in the OSU18 genome (Larsson *et al*., 2005; Petrosino *et al*., 2006). However, the presence of this kind of Tc-1 transposon is not usual in bacteria, but they are generally found in eukaryotes and have been reported in a range of invertebrates such as nematodes and insects (Larsson *et al*., 2005). The high level of *Francisella* IS*630* elements was suggested to be acquired originally from the infected insect vectors, which commonly mediate *Francisella*'s transmission (Larsson *et al*., 2005).

The expansion of the transposable elements, particularly IS elements, in bacterial pathogens is a recent emerging common feature (Siguier *et al*., 2006a). This is because IS elements facilitate homologous recombination within a genome, a process that can provoke large-scale genome rearrangements. These genomic changes often disrupt the ancestral gene order, and bring about a high level of gene inactivation and gene loss. Horizontally acquired genes have been balanced by gene loss, so that the bacterial genome size will not continuously increase. For example, comparison among three members of the Bordetella

family, *Bordetella bronchiseptica*, *Bordetella parapertussis* and *Bordetella pertussis*, revealed a smaller genome accompanied by more ISs (Parkhill *et al.*, 2003). A similar scenario can be seen in the Yersinae, of which a *Yersinia pestis* CO92 strain shows amplification of at least four different ISs: 66 copies of IS*1541*, 44 copies of IS*100* (a member of the IS*21* family), 21 copies of IS*285* (a member of the IS*256* family), and 9 copies of IS*1661* (a member of the IS*3* family) (Parkhill *et al.*, 2001). Overall numbers of IS copies in the CO92 strain are around ten-fold higher than those in another strain, *Yersinia pseudotuberculosis* (Parkhill *et al*., 2001). In addition, the genome displays anomalies in GC base-composition bias, indicating frequent intragenomic recombination through ISs (Parkhill *et al.*, 2001).

1.1.3.2 Bacteriophages

Bacterophages (or phages for short) contain either DNA or RNA enclosed in a protein coat. They are simply viruses that infect bacteria. The shapes and sizes of phages are quite variable, for example, those that infect *E. coli* are revealed to be of several types: λ and T4 with head and tail structures, ΦX174 and MS2 with simple head structures, and M13 with filamentous structures, etc (Hendrix, 2003). Prophages refer to phages that are inserted into the bacterial DNA and replicate as part of the bacterial chromosome (Freeman, 1951). Prophages are often identified as a by-product of bacterial genome sequencing. A surprisingly large number of prophages were discovered when an *E. coli* O157: H7 strain, Sakai strain, was genome sequenced in 2001 (Hayashi *et al.*, 2001). In this case, O157 Sakai contains 18 prophages or prophage remnants (Sakai prophages, Sp1-18), most of which are λ-like phages (Hayashi *et al.*, 2001). These phages account for about 16% of the chromosomal DNA of the Sakai genome and half of the strain-specific sequences (Ohnishi

et al., 2001). In addition, the Sakai strain contains six large chromosome segments that appear to be prophage-like genetic elements (Sakai prophage-like elements, SpLE1-6). Similarly, the K-12 genome also contains a total of 11 prophages, prophage remnants and phage-related elements (Hayashi *et al*., 2001). With such a high proportion of phage genes in *E. coli*, phage-mediated gene transfer was thought to play a predominant role in shaping the genomic diversity of *E. coli*.

Phage DNA fulfils a number of criteria for being an ideal vehicle for horizontal gene transfer. The residual footprints of prophages are different G+C contents and codon usage from the host's genome, the adjacent tRNA genes and the IS elements in the phage-essential regions. However, the prophages in bacterial genome sequences are accurately recognised through the similarity of their genes to known phage genes. Although phage genomes encompass an enormous amount of sequence diversity, their genes appear to be highly conserved (Casjens, 2003). For example, all the λ-like phages on the O157 Sakai chromosome are very similar and some share identical or nearly identical DNA segments of >20 kb (Hayashi *et al*., 2001). Among the prophages or phage-like elements of K-12, three share some sequence homology with the O157 Sakai prophages (Hayashi *et al*., 2001). More interestingly, some of the homologous prophages are integrated at analogous loci, such as Rac of K-12, Sp10 of Sakai, and CP-933R of another O157 strain, *E. coli* EDL933 (Ohnishi *et al*., 2001). Are these phage homologies the result of integration by different phages at the same bacterial attachment site, or are they the descendants of the same progenitor prophage? The first alternative seems unlikely because it rarely happens that two independent isolated infectious λ-phages are so identical over complicated regions such as Sp10 and CP-933R. It seems more reasonable that Sp10 and CP-933R are

descendants of the same original prophage. The same relationship is also seen in other prophage pairs, like Sp14/CP-933U, Sp4/CP-933M and Sp5/CP933V (Perna *et al.*, 2001). This suggests the presence of a common ancestor of EDL933 and Sakai, and that prophages were integrated before their divergence.

Bacteriophages, in order to co-exist with and inhabit the host, often accumulate deleterious mutations. Most prophages from sequenced bacterial genomes have shown attenuating point mutations, inactivating DNA insertions (often transposases) or progressive DNA deletions (Casjens, 2003). This leads to the appearance of defective prophages, prophage remnants and isolated prophage genes in bacterial genomes (Casjens, 2003). The movement of these defective prophages can be mobilised by transmission of the co-infecting or helper phages(Dobrindt and Hacker, 2001). For example, the defective prophage P4 can be mobilised by the P2 helper phage (Christie and Calendar, 1990). In the O157 Sakai genome, P4-like phage (Sp2) and P2-like phage remnants (Sp13) are thought to be involved in the transfer of some SpLE elements of O157 (Ohnishi *et al*., 2001).

The significance of bacteriophages in bacterial pathogenesis is obvious, since many important virulence genes are bacteriophage-encoded, such as Shiga toxin (Stx) and Type III secretion effectors of Enterohaemorrhagic *E. coli*, Cholera toxin (CTXΦ) of *Vibrio cholerae*, and Staphylococcal exterotoxin (Φ42) of *Staphylococcus aureus* (Dobrindt and Hacker, 2001; Tobe *et al.*, 2006). The prophages have a widespread role in driving the diversification of these bacterial pathogens.

1.1.3.3 Genomic Islands

Genomic islands are referred to as large chromosomal regions that contain a cluster of functionally related genes. They are often flanked by direct repeat sequences and are located near an integrase or transposase gene and also close to a tRNA gene (Dobrindt *et al.*, 2004). Genomic islands encoding virulence factors of pathogenic bacteria have been designated "pathogenicity islands" (PAIs) (Hacker *et al.*, 1997). Pathogenicity islands are characterised by five features: (i) they are large clusters (10-200 kb in size) present in the genomes of pathogenic strains but absent from those non-pathogenic strains of the related species; (ii) their G+C content differs from the rest of the host genome; (iii) they are often associated with tRNA genes; (iv) they are presumed to be generated by horizontal gene transfer; (v) they are recognised by conferring upon the host bacterium a complex and distinctive virulence phenotype in a single step (Hacker *et al*., 1997). An example could be taken from a well-characterised pathogenicity island known as the Locus for Enterocyte Effacement (LEE). LEE PAIs are found in the genomes of enteropathogenic *Escherichia coli* (EPEC) and enterohaemorrhagic *E. coli* (EHEC) (Nataro and Kaper, 1998), as well as the closely related mouse pathogen *Citrobacter rodentium*. The LEE gene cluster encodes a bacterial Type III secretion system, and can be present on a single horizontally transferable plasmid (Deng *et al.*, 2001), or inserted into the EPEC or EHEC genome at *selC*, *pheV*, or *pheU* tRNA sites (Jores *et al.*, 2004). The entire LEE is 35.6 kb in size, with a GC content of 38.36%, far below the *E. coli* genomic average (50.8%). LEE-encoded proteins are responsible for the development of a characteristic histopathological feature known as “attaching and effacing” (A/E) lesions (Frankel *et al.*, 1998).

1.2 Bacterial protein secretion systems

Gram-negative bacteria employ several protein secretion systems to transport proteins across the cell envelope and interact with host cells (Hueck, 1998). These systems often play important roles in bacterial pathogenicity during host-pathogen interactions. Five major secretion systems, numbered from type I to V (or designated as T1SS to T5SS for Type 1 to 5 Secretion System) have been well characterised (Henderson *et al.*, 2004). Recently, a type VI secretion system has also been described in the study of virulent gene secretion from *Vibrio cholerae* (Pukatzki *et al.*, 2006). By utilising these secretion systems, proteins destined for the extracellular environment of Gram-negative bacteria have to cross two membranes, the cytoplasmic (inner) membrane and the outer membrane, which are separated by the periplasmic compartment. These secretion systems are recognised by a set of core components, which build up a secretion device through the cell envelope.

1.2.1 Type I secretion system (T1SS)

The T1SS was first identified by the secretion of *E. coli* alpha-hemolysin (HlyA). HlyA is mainly produced by uropathogenic *E. coli* (UPEC). It is an important virulence factor owing to its cytolytic and cytotoxic activity against a wide range of mammalian cell types, including erythrocytes, granulocytes, monocytes and endothelial cells (Gentschev *et al.*, 2002). Once HlyA is exported to the plasma membrane of host cells, it causes pore formation and release of cytoplasmic contents (Gentschev *et al.*, 2002). Additionally, T1SS has been shown to be involved in the secretion of metalloprotease of *Erwinia chrysanthemi*, the leukotoxin of *Pasteurella haemolytica*, and the adenylate cyclase of *Bordetella pertussis* (Henderson *et al.*, 2004).

The translocator of T1SS is made up of three proteins that span the cell envelope: a pore-forming outer membrane protein (OMP), a membrane fusion protein (MFP) and an inner membrane ATP-binding cassette (ABC) protein. Translocation occurs via a Sec-independent pathway without a periplasmic intermediate. The ABC transporter provides energy through ATP hydrolysis, and possibly represents the initial channel across the inner membrane (Holland *et al.*, 2005). ABC transporter links with MFP, which spans the initial part of the periplasm and forms a continuous channel to the surface with an outer membrane protein. In *E. coli*, the translocation process is triggered by the C-terminus of effectors bound to the ABC transporter (HlyB), subsequently causing the assembly and interaction between MFP (HlyD) and an outer membrane protein (TolC). But before HlyB binding of the effectors, the interaction between HlyB and HlyD was identified. It is worth mentioning that the C-terminus of the effector carries a poorly conserved secretion signal, which specifically recognises the ABC protein (Delepelaire, 2004). In most cases, the secretion signals are not cleaved off during or after secretion. Through this trans-envelope complex, the effector molecules are secreted into the extracellular milieu.

1.2.2 Type II secretion system (T2SS)

The T2SS is widely distributed among proteobacteria. Proteins secreted by T2SS include proteases, cellulases, pectinases, phospholipases, lipases, and toxins, many of which contribute to pathogenesis in plant or animal cells (Sandkvist, 2001a). The type II pathway was first discovered in *Klebsiella oxytoca*, where it was found to be required for secretion of the starch-hydrolysing lipoprotein, pullulanase (PulA). Other well-studied members of

this system include the cholera toxin of *Vibrio cholerae*, exotoxin A of *P. aeruginosa*, and several cell wall-degrading enzymes (Henderson *et al*., 2004).

T2SS secretes proteins in a two-step Sec-dependent manner. Proteins to be secreted are synthesised with N-terminal signal peptides, which allows for Sec-dependent or Tat pathway translocation across the inner membrane (Sandkvist, 2001b). After export across the inner membrane, signal peptides are removed and proteins undergo further modifications. For example, PulA folds into a stable, inner membrane-anchored intermediate in the periplasm with at least one disulfide bridge being formed (Francetic and Pugsley, 2005). Some evidence showed that a disulfide bond isomerase, DsbA, is required to facilitate the formation of the disulfide bond (Thanassi, 2002). The Sec machinery is composed of an ATPase, namely SecA, which provides energy for the transport of the protein, the signal peptidase removal, and the release of the remainder of the protein into the periplasm (Henderson *et al*., 2004).

Translocation across the outer membrane is mediated by the Type II secretion apparatus or secretons, which are made up of 12-15 proteins (Sandkvist, 2001b). These proteins and their encoding genes have kept a consensus nomenclature using the letters A to O, and S. In *Pseudomonas*, however, the letters P-Z and A have been used. These genes are organised in a single operon with overlapping regions between each other, and are transcribed at the same time. The pore-forming protein has been identified as protein D, which belongs to a large family of homologous proteins called secretins. These secretins share similarity with Type IV pilus biosynthesis, filamentous phage extrusion and Type III secretion (Sandkvist, 2001b). The C-terminal domain of protein D is conserved and

thought to be embedded in the outer membrane, while the N-terminal domain is variable and thought to be exposed to the periplasm and to interact with other components, like proteins N, B and C (Sandkvist, 2001b). In *Klebsiella* and *Erwinia*, a small outer membrane lipoprotein, protein S, is required for stabilisation of protein D. Protein C is thought to interact with the integral cytoplasmic membrane components, L and M. Proteins L and M form a stable complex, which interacts with protein E. Protein E normally remains in the cytoplasm, and functions as a kinase that regulates the secretion process by providing energy to promote translocation and assembly of the pilin-like subunits, proteins G to K (Henderson *et al.*, 2004). Protein G is likely to be assembled into long pilus-like bundles, which could assist in protein secretion. Interestingly, in a Type II apparatus, most, if not all, of the components interact to form a multiprotein complex spanning both the inner and outer membranes. The homologies between T2SS and Type IV pili suggest they are evolutionarily related.

1.2.3 Type III secretion system (T3SS)

T3SSs are widespread in many gram-negative bacteria pathogenic for animals and plants, including *Yersinia* spp., *Salmonella enterica*, *Shigella flexneri*, *E. coli*, *Ralstonia solancearum*, *Pseudomonas syringae*, and *Chlamydia trachomatis* (Hueck, 1998). Like T1SS, T3SS translocates its effector proteins in a Sec-independent manner. The striking feature of T3SSs is their ability to target effector proteins directly into eukaryotic cells. Therefore, the host and pathogen interact without the interference of the extracellular milieu.

The structure of the Type III secretion system is built up by about 20-25 proteins spanning both the inner and outer membranes of the bacterial envelope. A study from *Salmonella typhimurium* showed a needle-like structure visualised under an electron microscope, therefore called a "needle complex (NC)" injectisome (Kubori *et al.*, 1998). In *S. typhimurium,* the needle complex consists of two parts: a base (or cylinder) spanning both membranes and a needle-like projection extending outward from the bacterial surface (Kubori *et al.*, 2000). The base contains three T3SS proteins: InvG (a member of the secretin family), and two lipoproteins, PrgH and PrgK, while the needle contains one protein, PrgI (Kubori *et al*., 2000). Proteins homologous to components of the needle complex are widely distributed among T3SSs in plant and animal pathogenic bacteria. In *Shigella flexneri*, four proteins, MxiD, MxiG, MxiJ and MxiH, which are homologous to InvG and PrgH, PrgK and PrgJ, respectively, were responsible for NC composition (Blocker *et al.*, 2001). Within them, MxiG and MxiJ are predicted to be inner membrane proteins to form the NC base, while MxiD is predicted to be an outer membrane protein, forming the outer ring. MxiH is the major needle component, together with another protein, MxiI, which serves to cap the external needle tip (Blocker *et al*., 2001). The function of the needle complex is to serve as a hollow channel through which the T3SS secreted proteins cross the two bacterial membranes. Interestingly, similarity exists between the architecture of the needle complex and the basal components of the flagellar export machinery, suggesting both are evolved from a common ancestral secretion system (Blocker *et al.*, 2003). Therefore, T3SSs were broadly divided into two major groups: flagellar T3SSs, which are associated with flagellar biosynthesis, and non-flagellar T3SSs (NF-T3SS, later referred to T3SS) (Pallen *et al.*, 2005).

Additionally, T3SSs require some proteins called “translocators” that are located on the bacterial surface, and which can contact with the host cell membrane and translocate another set of proteins into the host cell cytoplasm and across their plasma membrane (Ghosh, 2004). For example, the T3SS of *Yersinia spp.* comprises of about 25 Yop proteins as a secretion apparatus, also called Ysc. Among these Yop proteins, YopB, YopD, and LcrV are translocators responsible for the exportation of other Yop proteins (intracellular effectors) (Cornelis and Wolf-Watz, 1997). These translocated effectors are usually deleterious to the host cell. The effector molecules also vary vastly among different bacterial pathogens, even within a pathotype of some species. For example, more than 60 putative effector genes were identified in the Sakai strain of EHEC O157:H7 (Tobe *et al.*, 2006).

Many proteins secreted by T3SS also rely on specific T3SS chaperones, sometimes described as “bodyguards” for efficient translocation from the bacterial cytosol into host cells. The chaperones have been predicted to function to stablise proteins, to prevent inappropriate protein-protein interaction and to assist in secretion (Elliott *et al.*, 1999). There are two distinct functional classes of chaperone: the class-I chaperones, which bind to effectors, and the class-II chaperones, which bind to translocators (Page and Parsot, 2002). These chaperones are small, about 15-20 KDa in size, and located in the cytoplasm. For example, in the LEE T3SS of EPEC, the interaction between a translcoated protein, Tir, and an outer membrane intimin (the product of *eae*) is essential for the formation of A/E lesions and virulence. The chaperone, CesT, has been shown to bind and stabilise Tir in the bacterial cytoplasm prior to delivery of the complex into host cells (Elliott *et al.*, 1999). In addition to secretion, CesT was also identified to as being involved in recruiting multiple

type III effectors to the T3SS, including EscN (the membrane-associated ATPase of the T3SS) (Thomas *et al.*, 2005).

Interestingly, in some bacteria, such as *Salmonella enterica*, there exist more than one distinct type III secretion systems with different phenotypes (Galan and Collmer, 1999). *Salmonella* Spi-1 is required for the initial interaction with, and invasion of, intestinal epithelial cells, whereas Spi-2 is expressed only after *S. enterica* has gained access to host cells, and is required for systemic infections (Galan, 1999; Dobrindt *et al.*, 2004). Sequence comparison indicates that the LEE of EPEC and EHEC is similar to the Spi-2 system (Perna *et al.*, 1998). Many of the Spi-1 genes show significant similarity to the *mxi/spa* invasion genes of *Shigella* spp. (Hansen-Wester and Hensel, 2001). These scenarios suggest that horizontal gene transfer has taken place.

1.2.4 Type IV secretion system (T4SS)

Like T3SS, T4SSs mediate secretion by direct cell-to-cell transfer of virulence factors from many Gram-negative animal, human and plant pathogens, including *Agrobacterium tumefaciens*, *Bartonella tribocorum*, *Bordetella pertussis*, *Brucella suis*, *Helicobacter pylori*, *Legionella pneumophila* and *Rickettsia prowazekii* (Baron *et al.*, 2002; Juhas *et al.*, 2008). T4SSs share many similarities with the ancestral conjugation machinery, for instance, the one encoded by the F plasmid. Therefore, T4SSs can contribute to both bacterial virulence and genome plasticity and evolution mediated by horizontal gene transfer (Juhas *et al.*, 2008).

T4SSs deliver both effector proteins and nucleoproteins (protein-DNA complexes), which is unique in all types of secretory systems (Christie and Vogel, 2000). The prototypical T4SS is that of the *Agrobacterium tumefaciens* nucleoprotein T-DNA transfer system (Baron *et al.*, 2002). This has been studied in most detail and serves as a model system for T4SS research. The T4SSs of *A. tumefaciens* are composed of 12 protein components, named the VirB/VirD4 system. These proteins are encoded on the Ti (tumor-inducing) plasmid and are responsible for the transfer of oncogenic genes into plant cells (Christie and Vogel, 2000). The structures of the T4SS include a secretion channel and a pilus, which has functions in motility, adhesion, secretion and DNA transfer. The transferred conjugation-intermediate DNA is not naked DNA but single-stranded DNA associated with one or more proteins. The genetic makeup of T4SS includes 11 VirB proteins (VirB1-VirB11) encoded on a single operon, and VirD4 encoded on a separate operon. Among these, VirB6, VirB7, VirB8, VirB9 and VirB10 constitute the components of the transmembrane channel, whereas VirB2 and VirB5 are pilus components and VirB3 and VirB7 are pilus-associated proteins (Christie *et al.*, 2005). Three nucleoside triphosphatases (NTPase), VirB4, VirB11 and VirD4, provide the energy for transfer. In addition to T-DNA encoding oncogenic proteins, other effector proteins transferred via the T4SS have been identified, including VirD5, VirE2, VirE3 and VirF (Christie *et al*., 2005). VirE2 has been recognised as a single-stranded DNA binding protein. These proteins are delivered to the host plant cell, which increases the infectious ability of the bacteria.

VirB/VirD4-like T4SSs were also identified in several other bacteria, including *Helicobacter pylori*, *Bartonella henselae* and *Bordetella pertussis* (Christie *et al.*, 2005). The *cag* T4SS of *H. pylori* secretes CagA, the only virulence factor identified to date.

CagA can interact with several host cell proteins, resulting in altered physiology of the host cells and an increased chance of successful infection (Stein *et al.*, 2000). In contrast, the intracellular *Bartonella* spp. could deliver several translocated effector proteins, called Beps. The Bep proteins are involved in a wide variety of bacterial pathogenesis, including activation of pro-inflammfation, apoptosis and cytoskeleton rearrangements (Schulein *et al.*, 2005). However, *Bordetella pertussis* Ptl T4SS delivers pertussis toxin (PT, a A/B toxin family) not directly into the host cell but rather into the extracellular milieu (EM) (Rambow-Larsen and Weiss, 2004). The B domain of PT interacts with host cell glycoprotein receptors at the EM, and mediates translocation of the A domain across the host cell membrane (Rambow-Larsen and Weiss, 2004). All of these VirB/VirD4-like T4SSs are classified as Type IVA systems.

In contrast, the Type IVB secretion systems resemble those with the archetypal *dot* (defect in organelle trafficking) and *icm* (intracellular multiplication) gene system from two intracellular bacterial pathogens: *Legionella pneumophila* and *Coxiella burnetii*. The Dot/Icm systems share an extensive similarity with the Tra system of the Incl Collb-P9 plasmid of *Shigella flexneri* and are involved in bacterial conjugation (Christie and Vogel, 2000). The degree of relatedness between Type IVA and Type IVB is very limited. The T4SS of *L. pneumophila* was first identified by screening for mutants defective in multiplication in host macrophages (Segal *et al.*, 1998; Vogel *et al.*, 1998). These genes were named *icm* (intracellular multiplication) and *dot* (defective organelle trafficking) genes, and include about 25 genes within two operons. Mutation of these genes imparted a variety of bacterial virulence, such as phagocytosis, pore formation in the host cell membrane, inhibition of phagosome-lysosome fusion, apoptosis of the host cell and escape

from phagosomes (Segal *et al.*, 2005). Several effector proteins delivered via the T4SS of *L. pneumophila* have been identified. For example, RalF is required for the localisation of ARF (ADP-ribosylation factor) on phagosomes containing *L. pneumophila*; LidA is required for the formation of the replicative vacuole; LepA and LepB are required for escape of the bacteria from the phagosome; and several Sids proteins (substrates of Icm/Dot transporter) are involved in apoptosis (Segal *et al.*, 2005; Juhas *et al.*, 2008). In addition, these secretion components and effectors are all regulated by two regulators, PmrA and CpxR, suggesting that these Icm/Dot genes are functionally related.

However, in contrast to all other *L. pneumophila dot/icm* genes, only *icmF* and *dotU* have homologues in other gram-negative bacteria that do not possess a T4SS. This suggests that the *icmF*-like and *dotU*-like genes are not necessarily associated with T4SSs. Studies on DotU and IcmF showed that the two proteins are localised in the *L. pneumophila* inner membrane. The loss of IcmF caused a reduced level of DotU protein and affected the stability of DotU, suggesting that the two proteins interact with each other. Moreover, the lack of IcmF and/or DotU affected the stability of three other Dot proteins (DotH, G, and F) (Sexton *et al.*, 2004). IcmF mutants appeared to be lysed more rapidly, suggesting a role for IcmF in cell surface reorganisation, which results in increased adherence to host cells.

1.2.5 Type V secretion system (T5SS)

Among all the types of secretory systems, perhaps T5SS is the simplest one. This system was first described in the secretion of immunoglobulin A1 (IgA1) protease from *Neisseria gonorrhoeae* (Pohlner *et al.*, 1987). In that experiment, DNA sequencing of a 4.6 kb cloned fragment revealed a large open reading frame encoding a 169 KDa precursor of

IgA1 protease (Pohlner *et al*., 1987). A characteristic of the secretion of IgA1 protease is that it acquires an active confirmation as its extracellular transport proceeds and it is released as a proform from the membrane-bound helper by autoproteolysis (Henderson *et al.*, 2004). In other words, this precursor protein directs the transport of IgA1 protease from the periplasm to the extracellular milieu, without the involvement of any other accessory protein. Therefore, the proteins secreted via T5SSs are also called "autotransporters" (Henderson *et al.*, 2004). As shown by work on IgA1 protease, three essential domains are required for a typical T5SS autotransporter. An N-terminal signal sequence allows targeting of the protein to the inner membrane for its further export into the periplasm. Next is the passenger domain, which confers the diverse effector functions of the various autotransporters. The C-terminal end of the translocation unit consists of a short linker region with an α-helical secondary structure and a β-core that adopts a β-barrel tertiary structure when embedded in the outer membrane (Henderson *et al*., 2004).

The inner membrane translocation has been identified as a Sec-dependent mechanism, in which SecB was assumed to be a molecular chaperone. Several studies have identified autotransporters with typical signal sequences, including *H. pylori* AlpA, *B. pertussis* SphB1, and *N. meningitidis* AspA/NalP proteins (Henderson *et al.*, 2004). During export through the inner membrane, the signal sequences are cleaved while the autotransporter proteins might exist in a partially folded status in the periplasm. However, the status of the protein in periplasm is still controversial. After formation of the β-barrel in the outer membrane, the passenger domain inserts into the pore and is translocated to the bacterial cell surface. The released passenger domains of the autotransporter proteins are very diverse in terms of virulence. They could display enzymatic activity, such as protease,

peptidase, lipase, and esterase, mediate actin-promoted bacterial motility, act as adhesins, immunomodulatory proteins, toxins or cytotoxins, or permit the maturation of other virulence proteins (Henderson *et al.*, 2004).

In addition to the autotransporter T5SS, which is classified as Type Va, two other subgroups, Type Vb and Type Vc, have been characterised. In Type Vb, the passenger domain and the pore-forming β domain are translated as two separate proteins, rather than a single polypeptide in Type Va (Henderson *et al*., 2004). Therefore, Type Vb also refers to a two-partner secretion (TPS) pathway. Type Vc possess a particular outer membrane topology, belonging to the Oca (for Oligomeric Coiled-coil Adhesins) family. These family members are probably considered as a subfamily of surface-attached oligomeric autotransporters, exemplified by *Y. pestis* YadA (Henderson *et al.*, 2004).

1.3 Bacterial genome sequencing

1.3.1 Shotgun sequencing

The commonest method of DNA sequencing known is Sanger sequencing, or chain termination sequencing using dideoxynucleotides (Sanger *et al.*, 1977b). The first bacterial genome sequence was derived from a laboratory strain of *Haemophilus influenzae*, and was followed by an isolate of *Mycoplasma genitalium* in the same year (Fleischmann *et al.*, 1995; Fraser *et al.*, 1995). Prior to the birth of the first bacterial genome sequence of *H. influenzae*, there were doubts about whether the approach of sequencing random pieces of DNA and then assembling them into an entire genome could work in practice. Therefore, when the project to sequence the 1.8 Mbp *H. influenzae* genome started in 1994, it was not

certain that the whole-genome sequencing strategy would be successful, even though the chosen organism had a small genome. However, today the random sequencing phase of a genome project is relatively routine and remains widely used, though many "next-generation" sequencing technologies are emerging (Schuster, 2008).

The whole-genome random sequencing procedure, also referred to as shotgun sequencing, was used in the *H. influenzae* genome sequencing project (Fleischmann *et al.*, 1995). In contrast to the traditional ordered sequencing approach, the shotgun procedure obtains sequence from randomly broken short fragments. Therefore, no prior knowledge, such as genetic or physical maps of the genome, is required. Firstly, the genomic DNA was broken into random fragments using a sonicator. Those fragments in the size range 1.6-2.0 kb were identified by gel electrophoresis and purified and ligated into a plasmid vector. Then the end-sequences from each fragment were determined by Sanger DNA sequencing. At this stage, the sequence reads obtained corresponded to six times the length of the *H. influenzae* genome (Fleischmann *et al.*, 1995). Next, using a computer, the overlaps from all the fragments were searched and assembled, and the identified matching sequences were collected and assembled into contiguous sequences, referred to as "contigs". Typically, before the finishing stage, more than 99% of the genome can be spanned by using a shotgun sequencing approach (Fraser *et al.*, 2002).

Once the shotgun library is established and the contigs are assembled, the next step is called genome finishing. This includes closing the gaps between contigs, resolving all ambiguities and improving overall quality to minimise the possibility of errors in the sequence. Currently, genome sequencing is producing shotgun sequence data at a very high

rate using advanced sequencing technologies (Hall, 2007). However, genome finishing is much more labour-intensive and time-consuming than the shotgun phase. The finishing stage is therefore a bottleneck in large-scale sequencing efforts. The sequence data from many genomes are now often released unfinished (Pallen and Wren, 2007).

Annotation is another bottleneck in sequencing projects. Annotation of a genome involves the description or prediction of the boundaries of genes and other genomic elements, and of the function of the gene products (Riley *et al.*, 2006). In the post-genomic era, annotating a gene involves the integration of information obtained through genomic sequencing efforts, bioinformatic analyses and direct experimental validation. However, since both sequence analysis and experimental evidence are constantly expanding, no annotation can ever be considered complete (Riley *et al.*, 2006). For example, when the genome of *Escherichia coli* K-12 was initially annotated in 1997, of 4,288 protein-coding genes annotated, 38% had no attributed function (Blattner *et al.*, 1997). A functional update in 2001 amended the number of gene products to 4,401, of which 19.6% had no functional assignment (Serres *et al.*, 2001). The latest version updated that to a total of 4,506 *E. coli* annotated gene products, of which 471 were of unknown function, representing 10% of *E. coli* chromosomal genes (Riley *et al*., 2006). The number of functionally unknown genes is continually falling due to characterisation of new gene functions, a process facilitated by the release of the annotated genomes of *Salmonella, Shigella* and other closely related organisms.

1.3.2 Cost of genome sequencing

Advances in technology have led to a reduction in the time and money required to sequence a genome. However, at the beginning of this PhD project, sequencing a whole genome, together with the challenges of genome finishing and annotation, required resources that were beyond the scope of a typical laboratory (Gresham *et al.*, 2008). It was not yet practical and routine to sequence every individual genome using the traditional methods. However, complete genome sequencing remains invaluable in many areas of microbial research, and there is constant interest in exploring more and more bacterial genomes. Therefore, there was a demand for rapid and cost-effective approaches to complete genome sequencing.

Studies on genomic diversity indicate that closely related bacteria share a large proportion of their genomic information. Therefore, sequencing an entire genome is essentially a case of duplicated effort in those co-linear regions, when most interest is in the regions of difference from previously sequenced relatives. Therefore, new methods for efficiently detecting genomic variations are required. With such methods, it is possible to reduce the sequencing requirement to a tiny fraction of the genome, a capability that is available in most modern biology laboratories (Gresham *et al.*, 2008). Today, many finished and unfinished genome sequences have been made publicly available, and several databases have been established for bacterial genome sequences analyses, some with an emphasis on comparative genomics, for example the *x*BASE server, http://xbase.bham.ac.uk (Chaudhuri and Pallen, 2006; Chaudhuri *et al.*, 2008).

1.4 Comparative genomics

Since the first bacterial genome was sequenced in 1995, bacterial genome sequencing has been progressing rapidly. At the time of writing, more than 600 bacterial genomes have been completely sequenced and made publicly available and nearly 1800 projects are ongoing (http://www.genomesonline.org). The area of genomics is one of the most dramatic developments in science over the past decade.

After the first two bacterial genomes were sequenced, it was suggested that a comparison of the two may indicate the minimal gene set required for bacterial life (Mushegian and Koonin, 1996). As both strains were non-infectious, the genome sequences provided insights into bacterial biology rather than pathogenicity. On the other hand, as the two bacterial genomes were amongst the smallest and not closely related, the estimate is based on a simple lifestyle. However, it was clear that survival in different environmental niches would require additional genes, for example those associated with pathogenicity. When more genome sequences became available and more genomic variations were identified, it seemed impossible that “one size fits all” in the entire bacterial domain (Pallen and Wren, 2007).

Today, with hundreds of bacterial genomes available, what have we learnt from bacterial genomics? Firstly, continuously expanding genomic data have provided a foundation for various applications, perhaps most significantly for comparative genomics. While analysis of a single genome gives biological insights into any given organism, comparative analysis of multiple genomes provides substantially more information on the physiology and evolution of bacteria, expands better functional annotation for the predicted coding

sequences, and identifies new genes (Fraser *et al.*, 2000). Secondly, since most genome sequence projects have been targeted to pathogenic bacteria, comparative genomics provided a new route for the discovery of bacterial virulence factors. Particularly, *in vitro* culture seems impossible for some obligate intracellular bacterial pathogens, and genomics has provided the effective method for the identification of virulence genes (Raskin *et al.*, 2006).

Comparative genomics sheds light on many genomic variations within closely related bacteria. Although the chromosomal organisation among more closely related bacteria is more conserved, genomic diversity exists and the extent of it was revealed to be far greater than expected even within a single bacterial species (Ohnishi *et al.*, 2002). For example, in *E. coli*, the largest genomes possess over 1 Mb more DNA than the smallest genomes. This variation in genome size reflects the genomic composition, which differs vastly within the same or related species. The complete genome of *E. coli* O157:H7 strain EDL933 revealed 177 strain-specific genes (greater than 50 bp) compared with 234 unique genes (greater than 50 bp) in the laboratory strain *E. coli* K-12 (Perna *et al.*, 2001). This means that the two *E. coli* strains are heterogeneous for more than 25% of their genomes. The identified genomic differences can provide insights into biological function, as well as about the evolution of bacterial pathogens. Particularly, gene gain or loss, and gene context can provide clues to the identity and function of virulence factors. Additionally, genomic comparisons with closely related species can accelerate functional annotation of novel genes and other features such as gene fusions and pseudogenes (Raskin *et al.*, 2006). Increasingly, sequences of closely related genomes are available, meaning that genome

differences are minimal, and comparative genomics allows the identification of single nucleotide polymorphisms (SNPs).

Technologies for comparing bacteria at genomic, transcriptomic and proteomic levels in recent years have been extensively reviewed recently (Binnewies *et al.*, 2006). Here, two commonly used methods are summarised.

DNA Microarrays. DNA microarrays have revolutionised comparative genomics since the technology was first used for comparing several *Mycobacterium bovis* vaccine strains (Behr *et al.*, 1999). Since then, DNA microarray technology has been widely used for comparative analyses of the genome content and genetic variations of different bacterial species, including *Helicobacter pylori* (Salama *et al.*, 2000), *Campylobactor jejuni* (Dorrell *et al.*, 2001), *E. coli* (Ochman and Jones, 2000), *Staphylococcus aureus* (Lindsay *et al.*, 2006) and *Mycobacterium tuberculosis* (Butcher, 2004). Microarrays have become one of the commonest methods to quickly reveal the gene contents of many closely related strains of a species. This approach is referred to as comparative genome hybridisation (CGH).

CGH studies of this kind employ a microarray containing representations of all the genes of a sequenced and annotated reference strain. Labelled DNA from an un-sequenced, but related, experimental strain is hybridised to the array (Schoolnik, 2002a). The resulting hybridised array will identify genes common to both strains and genes that are present in the reference strain but absent in the experimental strain (Schoolnik, 2002b). This technique can therefore provide a rapid assessment of the differences between two genomes. However, the limitation of this method is that it cannot detect genes present in

the experimental strain but absent in the reference strain. Recently, the application of multi-strain microarrays has allowed extensive gene coverage of the species by incorporating a number of strains, and has been used for investigation of the horizontal acquisition or loss of MGEs in related strains of *Staphylococcus aureus* (Witney *et al.*, 2005). However, it is clear from pan-genome studies, such as the *Streptococcus agalactiae* "pan-genome"(Tettelin *et al.*, 2005), that the size of bacterial gene pools often greatly exceeds the sequenced diversity, so these expanded arrays are still unlikely to identify all the genes in the test genome.

Whole-genome PCR scanning (WGPS). WGPS was first used for a systematic PCR analysis of genomic diversity among eight *E. coli* O157 strains (Ohnishi *et al*., 2002). This approach is used for the overall structural comparison of closely related genomes, as determined by whole genome amplification using the Long Range PCR technique, when the sequence of at least one of these related genomes is available. WGPS provides information on gene order and genomic organisation and can rapidly identify regions of a test genome that are co-linear with the reference genome. By comparing the amplified fragments to those of the reference genome, several important features of a new genome could be determined, not only the presence of every target region, but some structural changes, e.g. rearrangement, large insertions and deletions. Particularly, the size and position of these changes can be identified by comparison with the reference genome.

When compared with CGH using DNA microarrays, WGPS has many advantages for comparative genomics among closely related strains. CGH does not provide information about gene position or about genes that are absent in the reference strains but these facts

can be revealed by WGPS. By sequencing the obtained PCR fragments, one can obtain additional genomic information. However, like CGH, one of the drawbacks of WGPS is the difficulty in identifying novel regions within a genome. This is because all of the primers were designed based on a reference strain. Once the PCR corresponding to a certain segment has failed, one cannot discern whether mutations have occurred in the primer binding sites or whether a chromosomal deletion or large insertion or rearrangement has occurred.

1.5 This Thesis

At the start of this work, it was clear that there was a pressing need for methods that allow rapid, efficient and low-cost assessment and analysis of genomic diversity within bacteria. The work presented here includes the development of long PCR-based methods to allow comparative genomics studies to be performed on non-sequenced genomes. Case studies have been performed in three Gram-negative organisms.

1.5.1 *Escherichia coli*

Comparisons among several complete *E. coli* genomes discovered two large cryptic regions, which are present in some *E. coli* strains, but not in others. These two uncharacterised regions contain genes homologous to those from bacterial T3SSs. The first region, ETT2, is similar to *Salmonella* Spi-1, which encodes a non-flagellar T3SS. The second novel gene cluster, Flag-2, is similar to the *Vibrio parahaemolyticus* lateral flagellar system. The aim of this project was to investigate the distribution of both the ETT2 and Flag-2 gene clusters within the full phylogenetic diversity of the *E. coli* species, and then to reconstruct the evolutionary histories of these genomic islands.

1.5.2 *Campylobacter jejuni*

The PCR-based comparative genomic approach was also applied to explore genomic diversity within *Campylobacter jejuni*, comparing the non-sequenced strain M1 with the sequenced reference strain, *C. jejuni* NCTC11168. Long PCR and single-primer PCR were used to identify and obtain sequence information from the most dynamic regions of the M1 chromosome. These approaches represent a shortcut to allow the identification of the core-genome and variable regions in a non-sequenced strain.

1.5.3 *Francisella tularensis*

Another application of the long PCR scanning approach was to guide the finishing process for the genome sequencing of the *Francisella tularensis* isolate, FSC198. Comparison with the complete genome of a closely related strain, *F. tularensis* Schu S4, showed a colinear relationship that facilitated the finishing stage of the FSC198 genome. Finishing the genome to completion allowed the identification of apparent differences between the two genomes, and those regions of discrepancy were subsequently resequenced in both strains. The identified single-nucleotide polymorphisms (SNPs) and variable number tandem repeat regions (VNTRs) shed light on the relationship between the two strains.

The availability of whole-genome sequences can help to identify potential vaccine candidates, an approach called “reverse vaccinology” (Rappuoli and Nabel, 2001). In an additional project, genes located within a previously identified pathogenicity island of *Francisella* were characterised to investigate their potential for vaccine development.

CHAPTER TWO

GENERAL MATERIALS AND METHODS

2.1 Materials

All media was purchased from Oxoid and chemicals from Sigma, unless otherwise stated.

2.2 Media

Luria-Bertani (LB) broth consists of 1% (w/v) tryptone, 0.5% (w/v) yeast extract and 1% (w/v) NaCl. Luria-Bertani Agar (LB-A or LA) was made from LB with the addition of 1.5% (w/v) bacto-agar. LB and LA stock were prepared and autoclaved within the department then stored at room temperature until use. LA was re-melted in a microwave then cooled to 50^{o}C before pouring plates. Antibiotics were added as required at a final concentration of 100 μg/ml ampicillin (Amp), 50 μg/ml kanamycin (Kan) and 25 μg/ml chloramphenicol (Chl) before use. Stock solution of antibiotics were sterilised by filtration through a 0.2 μm filter (Millipore, England). SOC broth used in bacterial transformation was purchased from Sigma.

2.3 Bacterial strains

Commercially available bacterial strains used in this study are listed in Table 2-1. Other bacterial strains are described in the relevant chapters.

Table 2-1 Commercially available bacterial strains used in this study.

Bacterial Strains	Genotype	Source
Top 10	F- *mcrA* Δ(*mrr-hsd*RMS-*mcr*BC) φ80*lac*Z ΔM15 Δ*lac*X74 *recA*1 *ara*D139 Δ(*araleu*) 7697 *galU* *galK rpsL* (StrR) *endA*1 *nupG*	Invitrogen
BL 21 (DE3) pLysS	F-, *ompT*, *hsdS*$_\beta$(r_β-m_β-), *dcm*, *gal*, (DE3), pLysS(Cm^R), *tonA*	Invitrogen
DH 5α	F-, φ80d*lacZ*ΔM15, Δ(*lacZYA*-*argF*)U169, *deoR*, *recA*1, *endA*1, *hsdR*17(rk^-, mk^+), *phoA*, *supE*44, λ^-, *thi*-1, *gyrA*96, *relA*1	Sigma

2.4 Bacterial growth conditions

Agar plates for *E. coli* growth were incubated at 37°C overnight. Liquid cultures were inoculated with a loop of pure growth and then incubated at 37°C in an orbital shaker or shaking water bath at 220 rpm overnight. The overnight culture was used directly for DNA extraction. For mid-logarithmic phase cultures, a small volume of overnight broth culture was inoculated into fresh LB broth with a dilution of one into a thousand, and then incubated in a shaking incubator for 3-4 hours until an OD_{600} of 0.5-0.6 was reached. For long term storage, 750 µl of an overnight broth culture was mixed with 250 µl of 100% (w/v) glycerol in a cryovial and stored at -80°C.

2.5 Transformations

2.5.1 Heat-shock transformation of chemically competent cells

Commercially available chemically competent cells used in this study are listed in Table 2-1. Competent cells were thawed on ice, then 25 µl per transformation was aliquotted into ice-cold 1.5 ml microfuge tubes. The transformation reaction was started by the addition of 1 µl of the desired plasmid to each aliquot followed by gentle mixing. The mixture was left on ice for 30 minutes and then heat-shocked in a 42°C water bath for 40 seconds. 500 µl SOC or LB broth was added into each tube, and the cells were then incubated at 37°C with shaking at 220 rpm for one hour. 50-100 µl of the cells was inoculated onto LB-A plates, containing an appropriate antibiotic if required, and incubated at 37°C overnight.

2.5.2 Electroporation

Electro-competent cells were prepared from a 100 ml mid-logarithmic phase culture. The culture was centrifuged at 3,000 *g* in a bench top centrifuge for 10 minutes, the pellet was washed 3 times in 10 ml ice-cold, sterile 10% (w/v) glycerol and the cells were then re-suspended in 200 µl sterile dH_2O. The cells were used immediately or stored in aliquots at -80^{o}C glycerol.

For each aliquot of 40 µl cells, approximately 200 ng of the plasmid DNA or PCR product was added and mixed gently. The cells were left on ice for up to 30 minutes before being transferred to an ice-cold sterile electroporation cuvette (2mm-gap, Geneflow). Electroporation was performed with a Biorad Gene Pulser under the following conditions: 1.8 kV, 25 µF, and 600 Ω. Following electroporation, cells were immediately washed out of the cuvette with 450 µl SOC or LB broth and transferred to a 1.5 ml microfuge tube. The cells were incubated for 1-2 hours with shaking at 37^{o}C. 50-100 µl of the culture was then plated onto LB-A plates containing the appropriate antibiotic and incubated overnight at 37^{o}C.

2.6 DNA extraction

2.6.1 DNA preparation

Bacterial chromosomal DNA was prepared using the DNeasy Kit (Qiagen), following the manufacturer's recommended protocol. The genomic DNA was aliquotted and stored at -20^{o}C until use.

2.6.2 Plasmid Isolation

Plasmid DNA was isolated using the QIAprep Miniprep kit (Qiagen) according to the manufacturer's instructions.

2.7 DNA Restriction and modification enzymes

2.7.1 Restriction Endonuclease digests

Restriction endonucleases and their buffers were purchased from New England Biolabs (NEB) and digests were carried out according to the manufacturer's instructions. Typically a digest reaction was incubated for 1-2 hours at 37^{o}C in a 50 µl volume, using a buffer supplied by NEB.

2.7.2 Ligation

Ligations were carried out with T4 DNA ligase and the buffer supplied by the manufacturer (Invitrogen, United Kingdom). Ligation reactions contained an approximate 1:1 molar ratio of insert to vector in a 10 µl volume, and were performed at 16^{o}C for up to 8 hours in a PCR machine. A quick ligation was performed using Right-Click 4x DNA Ligation Mix^{TM} (Yorkshire Bioscience). This kit allows the ligation to complete at room temperature in 10 minutes. The reaction requires 100 ng sticky/blunt ended plasmid DNA in a 15 µl volume, with a 1:5 ratio of insert to plasmid. The ligation products were checked by 0.8% agarose gel electrophoresis and visualised under UV.

2.8 Genetic Manipulations by Polymerase Chain Reaction (PCR)

2.8.1 Primer design and synthesis

Primers were designed using the online Primer3 software (Rozen and Skaletsky, 2000). The following guidelines were used for designing primers:

a) primers are 20-25 bases in length

b) GC% is about 50% (10-12 Gs and Cs per primer)

c) melting temperature (Tm) is ideally over 60^{o}C

d) primer hairpins are avoided

e) primers with 3' complementary ends are avoided, as they can result in primer dimerisation.

Primers were usually synthesised by Eurogentec (http://www.eurogentec.com) and supplied at a concentration of 100 μMol. When needed urgently, primers were synthesised by Alta Biosciences within the University of Birmingham.

2.8.2 PCR conditions and reactions

Unless otherwise stated, conventional short PCR was performed using *Taq* DNA polymerase (Invitrogen, United Kingdom). Each PCR was set up in a 200μl sterile PCR tube in a total volume of 20 μl-100 μl. Each 20 μl PCR reaction contained 1U *Taq* polymerase in 1x PCR buffer (supplied by Invitrogen), 20 ng genomic DNA, 8 pmol of each primer, 200 μM each dNTP, and 2 mM $MgCl_2$. PCR conditions were: 30 cycles of 30 seconds at 94^{o}C, 30 seconds at 60^{o}C and 1 minute at 72^{o}C, followed by 7 minutes

extension at 72°C. All PCRs were carried out in a PTC-225 DNA Engine Tetrad thermal cycler (Genetic Research Instrumentation, Braintree, UK).

2.8.3 Colony PCR

Colony PCR was used to screen for appropriately sized inserts using the flanking vector primers. Colony PCR was carried out in a 50 µl reaction volume containing either 5 µl cell lysate or a single colony selected using a toothpick. The cell lysate was prepared by boiling a fresh single colony in 100 µl dH_2O for 5 minutes, followed by centrifugation to obtain a pellet. The supernatant of the cell lysate was used as the template for colony PCR.

2.8.4 Long Range PCR (LR-PCR)

Effective and accurate amplification of long DNA targets (particularly exceeding 5 kb) has expanded the application of PCR in genetic studies. PCR products of up to 35 kb were able to be produced reproducibly using Klen*Taq*1, a mixture of *pfu* and Vent or Deep Vent (Barnes, 1994). The combination of two DNA polymerases, one of which is a proofreading enzyme, reduced the rates of mismatched bases in DNA strands in the PCR reaction (Barnes, 1994; Cheng *et al.*, 1994).

Klen*Taq*1 was kindly provided for this project by Dr Wayne Barnes (Washington University, St. Louis, USA). The buffer used was 10 x KLA, which consists of 500 mM Tris-HCl (pH 9.2), 160 mM ammonium sulphate, 25 mM $MgCl_2$ and 1% Tween 20. LR PCR conditions were: 30 cycles of 10 seconds at 94°C, 30 seconds at 62°C and an extension step of 68°C for 10 minutes plus 1 minute for each kb of the PCR product.

2.9 Agarose gel electrophoresis

DNA was analysed by electrophoresis on 0.5-1% agarose gels depending on the size of the DNA being loaded at, for example, 0.5-0.7% gels showed good resolution of large DNA fragments (more than 5 kb). 50x TAE stock buffer consists of 242 g Tris Base, 57.1 ml glacial acetic acid and 100 ml 0.5 M EDTA (pH 8.0) per litre. Agarose was dissolved in 1x TAE buffer and melted in a microwave. Once the agarose had cooled to about 50^{o}C, ethidium bromide was added to a final volume of 1 μg/ml and the gel was poured. The gel was allowed to set for at least 30 minutes, and then transferred to a horizontal electrophoresis tank containing TAE buffer, with the gel submerged to a depth of 2-5 mm. The sample DNA was mixed with 6x DNA loading buffer [0.25% (w/v) bromophenol blue, 30% (v/v) glycerol in water] and then added onto the gel. DNA electrophoresis was usually performed at 80-100 V for 30 min, but for large fragments (10-15 kb), gels were run at 25 V for more than 10 hours. DNA was visualised on the UV transilluminator of a Biorad gel documentation system.

2.10 PCR purification

PCR products were purified directly using the PCR Purification Kit (Qiagen) for removal of the remaining enzyme and primers following the manufacturer's instructions. Gel extraction was used where multiple bands were visualised by UV. The required DNA band was excised with a clean scalpel and purified from the gel using a QIAquick Gel Extraction Kit (Qiagen) according to the manufacturer's instructions.

2.11 Automated DNA sequencing

Sequencing of plasmid DNA or PCR products was performed by robots in the Functional Genomics Laboratory, School of Biosciences, University of Birmingham (http://www.genomics.bham.ac.uk/). Sequencing reactions consisted of a single primer at a final concentration of 300 nM mixed with 200-500 ng of template DNA in a final volume of 10 µl. The fully automated process carries out the labelling reaction, the cycle sequencing reaction using a Big Dye Terminator Kit (Applied Biosystems) and an MWG primeus HT 96 well Format PCR machine, and the sample purification. Finally, the sequencing products were obtained by capillary electrophoresis on an ABI 3700 DNA analyser with read lengths up to 1000 bp. Genetool Lite™ software and DNAMAN software were used to analyse the sequencing data.

2.12 Protein analysis

2.12.1 Buffers and solutions

A) Resolving buffer for SDS-PAGE gels

4 x resolving buffer contains 1.5 M Tris-HCl and 0.4% (w/v) sodium dodecyl sulphate (SDS) adjusted to pH 8.8.

B) Stacking buffer for SDS-PAGE gels

4 x stacking buffer contains 0.5 M Tris-HCl and 0.4% (w/v) SDS adjusted to pH 6.8.

C) SDS running buffer

The stock of 10x SDS running buffer consists of 30.3 g Tris base, 144 g glycine and 10 g SDS per litre of water. The 1x running buffer contains 50 ml of stock buffer in 450 ml dH_2O.

D) SDS gel loading buffer

2 x SDS gel loading buffer consists of 100 mM Tris-HCl (pH 6.8), 200 mM dithiothreitol, 4% (w/v) SDS, 0.2% (w/v) bromophenol blue and 20% (v/v) glycerol.

E) Western-blot transfer buffer

10 x western-blot transfer buffer contains 5.8 g Tris, 2.9 g glycine and 0.37 g SDS in 200 ml methanol and 800 ml dH_2O.

F) Phosphate Buffered Saline (PBS)

10 x PBS contains 80 g of NaCl, 2 g of KCl, 14.4 g of Na_2HPO_4, and 2.4 g of KH_2PO_4 in 1 litre of dH_2O, adjust the PH to 7.4.

G) Western-blot incubation/blocking buffer

Incubation/blocking buffer was made of 5% (w/v) milk powder in 1x phosphate buffered saline (PBS) and 0.05% Tween 20 (v/v).

H) Staining of SDS-PAGE gels

Coomassie blue R250 stain contains 0.2% (w/v) of coomassie blue R250 in 45% (v/v) absolute ethanol, 45% (v/v) dH_2O and 10% (v/v) glacial acetic acid.

2.12.2 Sodium Dodecyl (lauryl) Sulphate-PolyAcrylamide Gel Electrophoresis (SDS-PAGE)

SDS-PAGE (Laemmli, 1970) has several uses in protein analysis, such as establishing protein size, protein identification, and future blotting applications. The gels were made from Ultrapure Protogel™ (Fisher Scientific) containing 30% (w/v) acrylamide and 0.8% (w/v) NN'-methylenebisacrylamide.

The gels were set up by a 1.5 mm or 1 mm Biorad glass plates. The resolving gel was carefully poured into the glass plates avoiding air bubbles. A gap of 20-30 mm was left at the top of the gel and overlaid with dH_2O to ensure a flat surface and to exclude air. The resolving gel was left for about 50 minutes to set, and then the dH_2O overlay was removed using filter paper. Following that, the stacking gel was poured onto the top of the resolving gel and a 15 lane plastic comb was inserted to make the wells. The stacking gel was allowed to set for 50 minutes. Samples were mixed with 2x SDS loading buffer and heated at 95^{o}C for 5 minutes prior to loading. Molecular weight markers were loaded alongside the samples. For further blotting analyses, a pre-stained protein marker (NEB) was loaded. The tank was assembled with the gel glass plates and filled with SDS running buffer. Gels were run at 170 V until the dye front had reached the bottom of the gel.

2.12.3 Western blot

Western blot was performed by placing a nitrocellulose membrane (Millipore Immobilon-P) on the gel and using electrophoresis to drive the protein bands onto the nitrocellulose membrane. Freshly electrophoresed SDS-PAGE gel, filter paper and a sponger were dipped into transfer buffer at room temperature for 10 minutes prior to blotting. The membrane was pre-soaked in 100% methanol for 30 minutes before soaking in transfer buffer. A “sandwich” was assembled consisting of sponger, filter paper, gel and membrane, filter paper, and sponger, in order, to allow the use of a transblot system (Bio-Rad). The assembly was placed in a tank filled with western-blot transfer buffer. Transfer was either carried out overnight at a low voltage (30 V) or for 90 minutes at 100 V. An ice pack and a magnetic flea were placed in the tank, which was placed on a stirring block, to avoid

overheating. When finished, the membrane was incubated in blocking buffer overnight at 4^{o}C to ensure blocking of all non-specific protein binding sites on the blots.

To detect the desired proteins, two specific antibodies were used. The procedure was as follows:

1) Incubate with primary antibody (at a 1:500 dilution) in 10 ml blocking buffer for 1 hour at room temperature.
2) Wash 3 times for 10 minutes with 1x PBS.
3) Incubate with the secondary antibody (at a 1:500 dilution) in 10 ml fresh blocking buffer for 3 hours at 4^{o}C with gentle shaking.
4) Wash 3 times for 10 minutes with 1x PBS.
5) Add a substrate for the horseradish peroxidase (HRP) linked to the secondary antibody to develop the protein bands to be visualised.
6) Add developing solution that contains 3 ml of the substrate 4-chloro-1-napthol (60 mg dissolved in 20 ml methanol), 10 ml 1x PBS and 20 µl 30% (v/v) H_2O_2. The bands were visualised on the membrane at room temperature.

CHAPTER THREE

DISTRIBUTION AND EVOLUTION OF TWO TYPE III SECRETION GENE CLUSTERS FROM *ESCHERICHIA COLI*

3.1 Introduction

3.1.1 *Escherichia coli*

Escherichia coli is a widely used model organism in various areas of biology. It is one of the most intensively studied and best understood organisms (Donnenberg, 2002). The large body of knowledge about *E. coli* facilitates its use as a model organism. The quotation "all cell biologists have two cells of interest: the one they are studying and *Escherichia coli*" (Fred Neidhardt, 1996) illustrates this organism's special role in biology.

The natural habitat of *E. coli* is the gastrointestinal tract of warm-blooded animals or humans, and most strains are commensal and exist without harming the host (Donnenberg, 2002). However, the species *E. coli* also contains a wide range of pathotypes, which cause a variety of intestinal and extraintestinal diseases, typically classified as: diarrhoeal disease, urinary tract infections (UTI), or bloodstream sepsis and meningitis (Kaper *et al*., 2004). Six well described *E. coli* pathovars are associated with intestinal diseases: enteropathogenic *E. coli* (EPEC), enterohaemorrhagic *E. coli* (EHEC), enterotoxigenic *E. coli* (ETEC), enteroaggregative *E. coli* (EAEC), enteroinvasive *E. coli* (EIEC) and diffusely adherent *E. coli* (DAEC). Two well characterised pathogenic *E. coli* are associated with extraintestinal infections in humans: uropathogenic *E. coli* (UPEC) and neonatal meningitis *E. coli* (NMEC) (Kaper *et al*., 2004). These are often referred to together as extraintestinal pathogenic *E. coli* (ExPEC) (Johnson and Russo, 2002). Furthermore, the *Shigella* spp. is closely related to *E. coli* from a phylogenetic perspective, and is considered as another pathovar within the *E. coli* species (Pupo *et al.*, 2000). There are four recognised *Shigella* species: *S. dysenteriae*, *S. flexneri*, *S. boydii*, and *S. sonnei.*

3.1.2 *E. coli* Genomes

In 1997, the University of Wisconsin completely sequenced the first *E. coli* genome, *E. coli* K-12 strain MG1655 (Blattner *et al.*, 1997). A second *E. coli* K-12 strain, W3110, was sequenced the same year, but the genome as a whole was not completely sequenced and published until more recently (Hayashi *et al.*, 2006). In 2001, the US group and a different Japanese group published the complete sequence of two EHEC strains, *E. coli* O157:H7 EDL933 and Sakai (Hayashi *et al.*, 2001; Perna *et al.*, 2001), respectively.

The complete genome of *E. coli* O157 strain EDL933 was compared with *E. coli* K-12 MG1655 at the whole genome level. The 5.5 Mb chromosome of *E. coli* O157: H7 shares about 4.1 Mb common sequence with the *E. coli* K-12 genome (Perna *et al*., 2001). These conserved regions are referred to as the "backbone" or "core-genome", as they are co-linear (Donnenberg, 2002). The backbone is punctuated by lineage-specific elements (or islands), which account for the genome size difference and many of the phenotypic differences among *E. coli* strains. Genomic comparisons revealed 1.34 Mb of the EDL933 genome and 0.53 Mb of the K-12 genome are heterogeneous (Perna *et al*., 2001). Interestingly, these strain-specific islands are often found at the equivalent loci of the chromosome, indicating the presence of "hot-spots" for the insertion of foreign DNA acquired by horizontal gene transfer (Perna *et al*., 2001).

With the completion of the genome sequence of uropathogenetic *E. coli* strain CFT073, a three-way comparison of *E. coli* strains CFT073, EDL933 and MG1655 became possible (Welch *et al.*, 2002). It became obvious that the conserved core-genome is largely co-linear. Currently, it is estimated that all *E. coli* (including *Shigella*) genomes share about a 3 Mb

conserved core-genome, which represents 65% of the *E. coli* K-12 MG1655 genome (Yang *et al.*, 2005). Apart from this, the remaining genes are highly diverse. However, it is noteworthy that most of the strain-specific islands are not obviously associated with pathogenicity. Genes with an established role in pathogenicity are not limited to the large islands either. For example, only nine large islands (>15 kb) encoded known virulence factors when the O157:H7 EDL933 genome was released (Perna *et al*., 2001).

As genome sequencing has become cheaper and more routine, a range of additional genome sequencing projects has been initiated targeting the *E. coli* species. Genome sequences for almost all the *E. coli* pathotypes, together with all the *Shigella* species, are now publicly available, either as complete sequences or almost-complete drafts. The genome projects and sequencing status can be accessed through the NCBI (National Centre for Biotechnology Information) database (http://www.ncbi.nlm.nih.gov/sites/entrez). The increasing amount of genomes allows additional biological insights to be obtained through comparative genomics. An online database (http://xbase.bham.ac.uk/colibase) has been established in Birmingham with the aim of facilitating this process (Chaudhuri *et al.*, 2004; Chaudhuri and Pallen, 2006).

3.1.3 *E. coli* phylogenetics and diversity

To aid in the study of the variety of *E. coli* strains, the *Escherichia coli* Reference (ECOR) Collection was established by Howard Ochman and Robert Selander in 1984 (Ochman and Selander, 1984). The collection consists of 72 standard reference strains isolated from humans and 16 other mammalian species from various geographical distributions, and includes both pathogenic and non-pathogenic variants (Ochman and Selander, 1984). The

ECOR strains could be separated into five major phylogenetic groups, namely A, B1, B2, D and E (or unclassified), based on analysis of Multi-Locus Enzyme Electrophoresis (MLEE) data using the neighbour-joining algorithm (Herzer *et al.*, 1990). The validity of these groups has been supported in numerous later studies (Lecointre *et al.*, 1998; Escobar-Paramo *et al.*, 2004). More information about the ECOR collection is available from Thomas Whittam's laboratory website (http://foodsafe.msu.edu/whittam).

With the aid of the *E. coli* strain collections, the phylogenetic relationship between *E. coli* pathotypes and the commensal strains could be drawn. A study based on MLEE analysed the genetic distance of *E. coli* pathotypes within the ECOR collections (Donnenberg and Whittam, 2001). The study revealed the existence of two EPEC lineages that are associated with infantile diarrhea, EPEC1 and EPEC2, and two EHEC lineages that are associated with hemorrhagic colitis, EHEC1 and EHEC2. Both EPEC1 and EHEC1 are highly divergent, whereas EPEC2 and EHEC2 are closely related to one another and fall into the B1 group of ECOR (Donnenberg and Whittam, 2001).

Recently, a phylogenetic tree based on multi-locus sequence data (MLST) was constructed using the ECOR collection, plus a second collection of 78 diarrhoeagenic *E. coli* strains (referred to as the DEC collection) (Whittam *et al.*, 1993), and the other representative pathogenic strains of *E. coli/Shigella* (Escobar-Paramo *et al*., 2004). The authors of the study suggested that a specific genetic background is necessary for the expression and maintenance of certain virulence factors (Escobar-Paramo *et al*., 2004). Such analyses may help to determine the positions at which major acquisition of certain virulence factors (VF) occurred. According to this study, the strains belonging to EHEC1 are found in ECOR

group E, while those belonging to EHEC2 are found in group A. For EPEC, EPEC1 strains are in group B2, while EPEC2 strains are in group B1. Many strains with ExPEC virulent determinants are classified as being in group B2 (Escobar-Paramo *et al*., 2004).

3.1.4 Pathogenic *E. coli*

Pathogenic *E. coli* are characterised by the expression of virulence factors, which provide an enhanced ability to cause intestinal or extraintestinal disease (Kaper *et al*., 2004). Many virulence factors are encoded on mobile elements, such as plasmids, bacteriophages and pathogenicity islands, and acquired through horizontal gene transfer (Ochman *et al.*, 2000). Horizontal gene transfer plays an important role in the evolution of pathogenic *E. coli* strains.

However, though lateral gene transfer has the potential to introduce sizable amounts of DNA, bacterial genomes do not appear to be growing ever larger in size (Lawrence *et al.*, 2001). Some commensal *E. coli* can also undergo deletions, point mutations or other DNA rearrangements that can contribute to virulence (Kaper *et al*., 2004). There is increasing evidence that loss of gene function, or genome decay, increases with adaptation to the host (Wren, 2000).

3.1.5 *E. coli* Second T3SS - the ETT2 locus

With the arrival of two complete genome sequences, *E. coli* O157: H7 EDL933 and K-12 MG1655, genome comparisons revealed a second T3SS gene cluster within the *E. coli* species (Perna *et al*., 2001). These genes resemble *Salmonella* SPI-1 and organised in a

cluster, first designated as O-island 115 in Blattner's nomenclature (Perna *et al*., 2001). The gene cluster was 16.9 kb in size with a low GC content and was inserted in the tRNA *glyU* of *E. coli* (Perna *et al*., 2001). It was thought to be horizontally acquired by O157: H7 genomes at the time of discovery. In order to distinguish it from the LEE T3SS, it was named as ETT2 for *E. coli* type three secretion system 2 (Hayashi *et al*., 2001).

Homologous to *Salmonella/Shigella* "*inv-spa*" elements, the ETT2 gene cluster has drawn interest due to its potential role in pathogenicity in *E. coli*. A study by Hartleib *et al.* found that ETT2 was ubiquitously distributed among intestinal pathogenic *E. coli* strains, but not among extra-intestinal, non-pathogenic *E. coli* or other enterics (Hartleib *et al.*, 2003). Later, Makino *et al.* found that the prevalence of ETT2 genes was more common in EHEC strains than the LEE-encoded ETT1. They disclosed that ETT2 genes were present in 78 out of 89 O-antigen serotypes isolated from various pathotypes including EHEC, EPEC2, EAEC, and ETEC (Makino *et al.*, 2003). An incomplete ETT2 segment, which was 8.7 kb in size, was discovered in a EPEC2 strain, B171-8, (O111: NM). This ETT2 segment was sequenced and deposited in GenBank with accession no. AB052736 (Makino *et al*., 2003). Since the non-pathogenic *E. coli* included in both studies were found not to house any ETT2 genes, ETT2 was once considered a new marker for distinguishing between pathogenic and non-pathogenic *E. coli.*

However, these studies did not agree upon the boundaries of the ETT2 gene cluster. Makino *et al*. placed the left boundary at ECs3714 (Sakai genome nomenclature), delineating ETT2 as a 17 kb insertion (Makino *et al*., 2003), whereas Hartleib *et al*. placed the boundary further upstream at ECs3703 (*rmbA*/*yqeH*), making the cluster 29.9 kb in

length (Hartleib *et al*., 2003). The boundaries of the ETT2 gene cluster were later defined through three independent methods, namely homologies to other T3SS genes, G+C content, and genomic comparisons (Figure 3-1) (Ren *et al.*, 2004). The ETT2 island was defined as 27.5 kb in length, with a lower than average GC content, and had inserted into the intergenic region between ECs3702/*yqeG* and the tRNA *glyU* gene. The boundaries of ETT2 extended beyond O-island 115, suggesting that a remnant of ETT2 was retained in *E. coli* K-12.

A comparison of the ETT2 gene cluster from EHEC with the equivalent region of other *Escherichia* or *Shigella* genomes revealed the ETT2 gene locus is present, at least in part, in the majority of genome-sequenced *Escherichia* and *Shigella* strains. Nine of the 12 genome sequences showed evidence that ETT2 genes have been inserted at the same chromosomal site, within the *yqeG-glyU* intergenic region. By genomic comparisons, two genome sequences, EPEC1 strain E2348/69 and UPEC strain CFT073, were identified that lacked the ETT2 island in its entirety, and thus presumably represented the ancestral state for the species (Ren *et al*., 2004). An identical 14.6 kb deletion was found in the two K-12 laboratory strains. Remnants of ETT2 islands were also seen in two of the sequenced *Shigella* genomes. The full complement of ETT2 genes was found in the two genome-sequenced EHEC strains, EDL933 and Sakai, and EAEC strain 042 (Figure 3-1).

3.1.5.1 ETT2 genes

Using such analysis, the ETT2 gene locus was shown to contain 35 genes, namely ECs3703 to ECs3737 (Sakai nomenclature) from EHEC Sakai strain (Table 3-1). Most

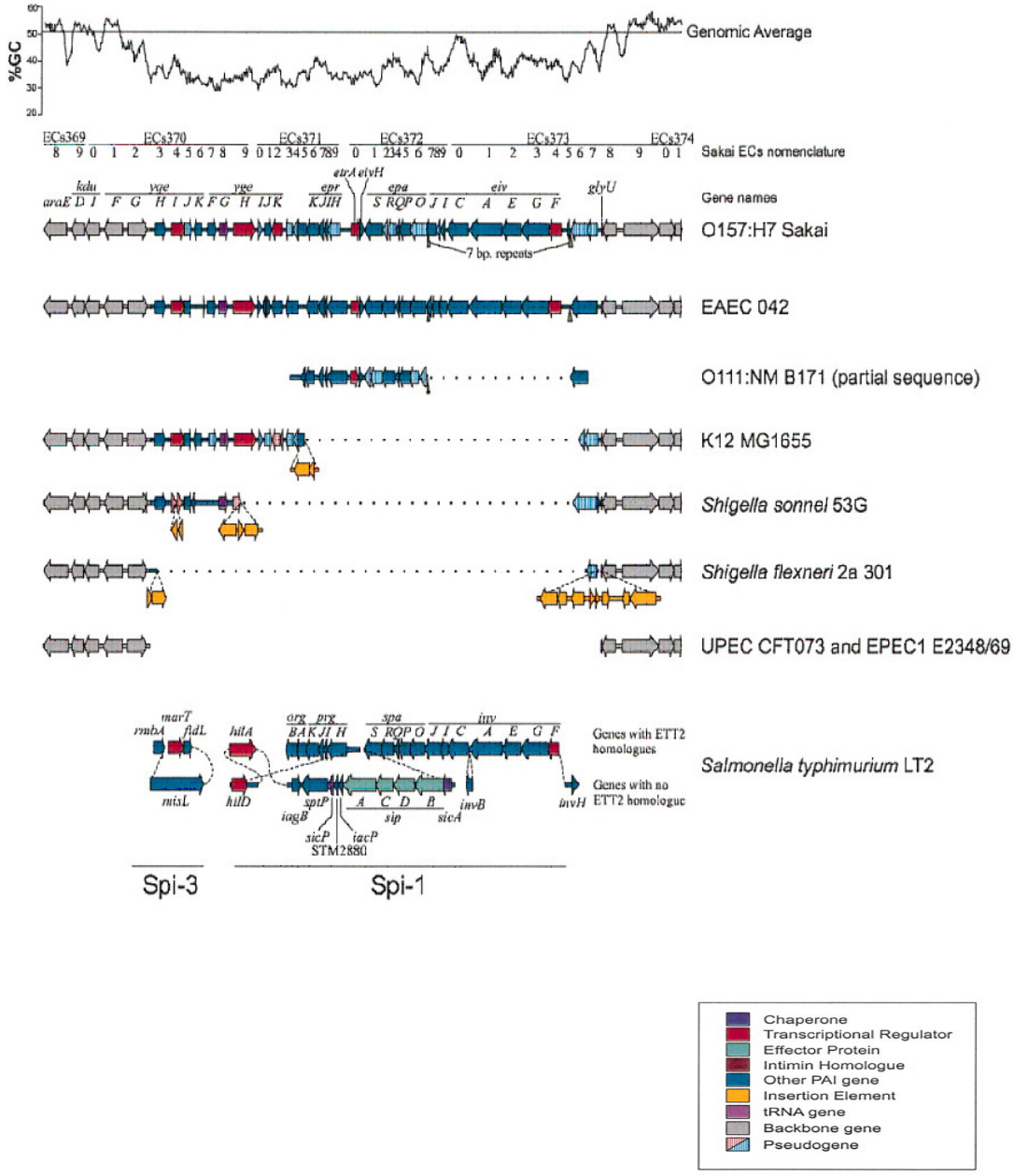

Figure 3-1 Schematic representation of genomic analysis of the ETT2 cluster. Structures of the ETT2 pathogenicity island in a number of *E. coli* and *Shigella* strains, and comparisons with regions of Spi-1 and Spi-3 from *Salmonella enterica* serovar Typhimurium are shown. Homologous genes are vertically aligned. Insertions relative to the complete ETT2 sequence (as seen in Sakai and EAEC 042) are indicated with dashed lines. Dotted lines indicate deletions (Ren *et al.*, 2004).

Table 3-1 Genes within the ETT2 gene cluster (Ren *et al*., 2004).

Gene in Sakai strain	Gene in K-12	Pseudogene(s)	Other homologies
ECs3703	*yqeH*	None	RmbA regulator from Spi-3
ECs3704	*yqeI*	None	MarT from Spi-3
ECs3705	*yqeJ*	Sakai, EDL933, K-12 W3110	FidL from Spi-3
ECs3706	*yqeK*	None	None
ECs3707	*ygeF*	K-12 W3110, K-12 MG1665	None
ECs3708	*ygeG*	None	SicA-like T3SS chaperone
ECs3709	*ygeH*	None	HilA-like T3SS regulator
ECs3710	*ygeI*	Sakai, EDL933, K-12 W3110	None
ECs3711	*b2854*	K-12 W3110, K-12 MG1665	IagB from Spi-1, PilT from *S. enterica*
ECs3712	*ygeK*	K-12 W3110, K-12 MG1665, misidentification of start site in EDL933	SsrB from Spi-2
ECs3713	*b2857*	Sakai, EDL933,	OrgB, MxiN
ECs3714	*b2858*	K-12 W3110, K-12 MG1665	None
ECs3715	*b2859*	K-12 W3110, K-12 MG1665	PrgA, MxiK
ECs3716/*eprH*	Absent	None	PrgK
ECs3717/*eprI*		None	PrgJ, MxiI
ECs3718/*eprJ*		Numerous frame shifts in EDL933	PrgI, MxiH
ECs3719/*eprK*		Sakai, EDL933	PrgH
ECs3720/*etrA*		None	Transcriptional regulator
ECs3721/*epaS*		Last three residues missing in EAEC	SpaS
ECs3722/*epaR2*		Sakai, EDL933	SpaR
ECs3723/*epaR3*			None
ECs3724/*epaQ*		None	SpaQ
ECs3725/*epaP*		None	SpaP
ECs3726/*epaO*		Sakai, EDL933	SpaO
ECs3727/*eivJ*		EAEC or Sakai/EDL933	SpaN/InvJ
ECs3728			None
ECs3729/*eivI*		None	SpaM
ECs3730/*eivC*		None	SpaI/InvC
ECs3731/*eivA*		None	InvA
ECs3732/*eivE*		None	InvE
ECs3733/*eivG*		None	InvG
ECs3734/*eivF*		None	InvF
ECs3735		None	Inner membrane YjdO/YdcX
ECs3736	*b2863*	Different frame shifts in K-12, O157, *S. sonnei, S. flexneri* 2a	Phosphorylase kinase and glucomamylases
ECs3737	*b2862*		

genes are organised in operons, suggesting functional co-ordination. By comparison with other T3SS genes, at least four operons are predicted within one ETT2 locus. Many of these genes have been predicted to be a Type III secretory apparatus, chaperone, regulator or translocator.

Interestingly, three genes from the leftmost extremity of the ETT2 cluster are homologous, not to Spi-1 genes but to the Spi-3 pathogenicity island from *S. enterica*. ECs3703 encodes protein exhibiting similarity to RmbA (39% identical over 190 aa), which is a predicted cytoplasmic protein belonging to the LuxR family of regulatory proteins. ECs3704 has homology with MarT (41% identical over 101 aa), a member of ToxR-like family of regulatory proteins (Blanc-Potard and Groisman, 1997). ToxR is a transmembrane regulatory protein that is required for the synthesis of cholera toxin in *V. cholerae* (Miller and Mekalanos, 1984). ECs3705 exhibits similarity to FidL (43% identical over 139 aa). Both MarT and FidL are predicted to be inner membrane proteins in *S. enterica*. A recent study suggested MarT played a role as a transcriptional activator, and one of its substrates was identified as MisL, an IgA homology transporter protein from *S. enterica* (Tukel *et al.*, 2007). In Spi-3, *misL* is located between *rmbA* and *fidL*. However, ETT2 has no *misL* counterpart. ECs3703, ECs3704 and ECs3705 appear to be organised in one transcriptional unit in the ETT2 locus.

As in *S. typhimurium* Spi-1, the Type III secretory apparatus is encoded by the *inv-spa* gene cluster and the *prgHIJK-orgAB* genes. Comparison of the ETT2 locus with Spi-1 revealed a generally similar gene complement and organisation. The genes ECs3734-ECs3729 and ECs3727-ECs3721 represent homologues to the *invFGEACIJ-spaOPQRS*

genes, respectively, predicted to form one large operon at the right hand end of the ETT2 locus. ECs3734/EivF, the first gene of this predicted operon, is homologous to the *Salmonella* transcriptional regulator InvF. Activation of S*almonella invF* results in transcription of other *inv-spa* genes. *S. typhimurium* InvH is an important outer membrane lipoprotein required for the proper localization of InvG and for the secretion of the virulence factor SipC (Daefler and Russel, 1998). However, an *invH* homologue was found to be absent in the ETT2 locus, although a pseudogene to an *invH* homologue (namely *eivH*) is present in the region between ECs3720 (*etrA*) and ECs3721 (*epaS*) (Makino *et al*., 2003). The co-expressed proteins InvH and InvG have been characterised as part of the functional outer membrane translocation complex in T3SS (Crago and Koronakis, 1998). EivC is homologous to InvC, the ATPase associated with Spi-1, and is thought to provide the energy for the secretion process of T3SS. The ATPases are highly conserved among all T3SSs. The phylogenetic analysis of EivC from ETT2 with its homologues from other TTSSs indicated that ETT2 belongs to the Spi-1/Mxi-Spa group of TTSSs (Ren *et al*., 2004). EivC from ETT2 shares about 40% identity at protein level with EscN from the LEE locus.

Homologues to Spi-1 *prgHIJK-orgAB* have been found as another predicted operon in the ETT2 locus, representing ECs3719-ECs3715 and ECs3713, respectively. All of the six genes in Spi-1 are required for *S. typhimurium* invasion, and their expression is driven by the *prgH* promoter (Klein *et al.*, 2000). ECs3714 has not been shown to have identity with Spi-1 or any other T3SS gene. The *Salmonella prgHIJK* operon encodes components required for formation of the supramolecular type III secretion needle complex (NC) (Kimbrough and Miller, 2000). Co-expression of PrgH and PrgK forms a ring structure

resembling the base of the needle complex, while PrgI forms the needle portion (Lostroh and Lee, 2001). Another protein, PrgJ, was suggested to be involved in the assembly of the needle portion protein PrgI (Sukhan *et al.*, 2003). However, comparisons of ETT2 sequences in both the EHEC O157:H7 genome and in *Salmonella* Spi-1 revealed frame-shift mutations in several important genes required for invasion and secretion: *spaR*, *prgH*, and *orgB*. The mutations in these structure genes would have abolished the function of Type III secretion.

In *Salmonella*, between the *inv-spa* genes and the *prg-org* genes are located the genes for the Spi-1 secreted proteins: SipBCDA and SptP. The Sip protein complex not only comprises a "translocase", through which effectors can pass into the host cell, but SipB and SipC themselves act as effector proteins (Carlson and Jones, 1998). Curiously, the ETT2 locus has none of these homologues. The immediate questions might be where the ETT2 effectors are and why this pathogenicity island has eliminated the effectors during its evolution. Homology searches of ETT2 genes could not find any other effectors, such as *Salmonella avrA* and *Sop* genes. However, a *sicA*-like chaperone gene ECs3708/*ygeG* is present in the ETT2 locus. The functions of the chaperone protein SicA affect the expression of several secreted proteins, not only of SipB and SipC encoded on Spi-1 (Tucker and Galan, 2000), but of SigD and SopE, which are unlinked to Spi-1 (Darwin and Miller, 2000). Activation of the *sicA* gene depends on interaction with the transcriptional factor *invF*. The ETT2 locus contains both SicA-like and EivF/ECs3734 proteins. However, the genetic arrangement of the ETT2 *sicA*-like gene is unlike that in Spi-1, where *sicA* is encoded immediately downstream of *spaS* and is part of the *inv-spa* operon. In addition,

ECs3708 has a relative low homology to *sicA* in Spi-1 (37% identity at the protein level over a 140 aa stretch). For these reasons, the function of ECs3708 in ETT2 is uncertain.

In fact, immediately downstream of the *sicA*-like chaperone of the ETT2 locus is a *hilA*-like T3SS regulator, ECs3709/*ygeH* (29% identity at the protein level over a 403 aa stretch). *Salmonella* HilA is a global transcriptional factor encoded on Spi-1, which co-ordinately regulates the expression of genes encoding the Spi-1 secretory apparatus and many other effectors (Lostroh and Lee, 2001). Besides, activation of InvF also requires HilA. In addition, ETT2 encodes another two regulators, ECs3712 and ECs3720/*etrA*. ECs3712 is homologous to SsrB from Spi-2 (32% identity at the protein level over a 209 aa stretch). SsrB represents the response regulator of a two-component regulatory system, SsrAB, regulating Spi-2 genes. ECs3720/EtrA exerted profound negative effects on gene transcription within the LEE locus of EPEC strains (Zhang *et al.*, 2004). However, although these regulatory genes are present on the ETT2 locus, the signals that trigger ETT2 expression are not clear at this time.

3.1.5.2 Eip gene locus

In an attempt to locate possible ETT2 effectors, homology searches were carried out among the sequenced *Escherichia* and *Shigella* genomes available at the time (Ren *et al*., 2004). A novel 20.9 kb pathogenicity island containing homologues of the *sip* effector genes was found in the EAEC 042 genome (Figure 3-2) (Ren *et al*., 2004). The island was designated as Eip for *E. coli* invasion proteins (Ren *et al*., 2004). Comparative analysis of the Eip locus in genome-sequenced strains revealed that it was not as prevalent as ETT2

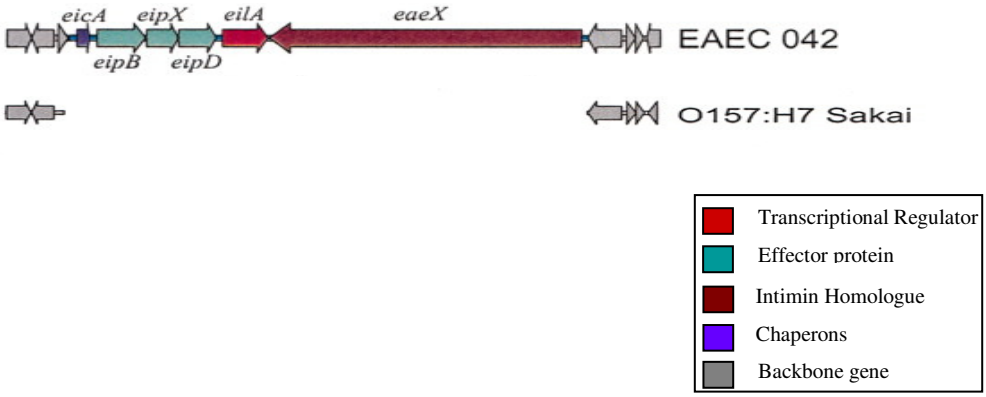

Figure 3-2 Schematic representation of the structure of the Eip island in EAEC strain 042 (produced by Dr. Chaudhuri). The comparison was carried out with the backbone sequence seen in Sakai. *eipB* and *eipD* encode proteins that are homologous to *Salmonella* SipB and SipD, respectively (also see Figure 3-1). Between these two *eip* genes lies a third gene, *eipX*, which shows weak similarity to *espD* and thus may encode an additional secreted translocator protein. In addition, the *eip* island contains genes coding for a novel SicA-like tetratricopeptide repeat chaperone, *eicA*; a novel HilA-like regulator, *eilA*; and an invasin/intimin-like large outer membrane protein, *eaeX* (Ren *et al.*, 2004).

but was only present in the 042 genome (Ren *et al*., 2004). Interestingly, EAEC 042 also contains the most intact ETT2 identified in any *E. coli* genome sequenced so far.

The Eip island is found between the *E. coli* backbone genes, *yicM* and *nlpA*, and is inserted at the *SelC* tRNA gene. This novel island is comprised of six genes arranged in tow predicted operons. Two homologues of *sip* effector genes have been identified: *eipB* encodes a protein similar to SipB (20% identical over 527 aa) and *eipD* encodes a protein similar to SipD (31% identical over 266 aa). Between these two *eip* genes lies a third gene, termed *eipX*, with weak homology to *espD* from the LEE locus. EspD is a secreted translocator protein and plays an important role in forming pore-like structures in the host cell membrane (Ide *et al.*, 2001). Since no translocators have been identified in the ETT2 locus, this finding suggested that the *eip* island might encode an additional translocator protein. In addition, *eipX* was also found to encode a product with low level identity (19%) to IpaC protein of *Shigella flexneri*, while IpaC is similar to SipC effector of *S. enterica* (Sheikh *et al.*, 2006). Therefore, *eipX* was also suggested to encode a SipC-like protein product.

Downstream of *eipB* is a SicA-like chaperone gene, *eicA*, while upstream of *eipD* is a HilA-like regulatory gene, *eilA*. The five genes, *eicA*, *eipB*, *eipX*, *eipD*, and *eilA*, are predicted in an operon (Figure 3-2). In addition, downstream of *eilA*, but in an opposite orientation, is a large gene, *eaeX*, encoding a product similar to invasin/intimin outer membrane protein. Tellingly, the expression of EilA was found to affect the expression of at least seven genes, including five genes in the *eip* locus (*eicA*, *eipB*, *eipX*, *eipD*, and *eaeX*), as well as two genes in the ETT2 locus (*eivF* and *eivA*) (Sheikh *et al.*, 2006). This is

explained by the fact that HilA controls a T3SS in *S. typhimurium*. This finding also demonstrated co-expression of the ETT2 genes and *eip* genes from EAEC strains, even though they are distantly located.

3.1.6 The second flagellar system in *E. coli* - the Flag-2 locus

Bacterial flagella are complex organelles that provide power for bacterial motility and also play a central role in adhesion, biofilm formation, and host invasion. A typical bacterial flagellum comprises of six components: a basal body (including MS ring, P ring, and L ring), a motor, a switch, a hook, a filament, and an export apparatus (Macnab, 2003). Flagellar assembly shares many properties with T3SSs. It initiates from the formation of a ring structure in the membrane, and proceeds to formation of a basal body hook and an extra-cytoplasmic propeller-like filament (Macnab, 2003). The assembly of this complex organelle requires the products of more than 35 genes. Flagellar biosynthesis, assembly and regulation have been most studied in *Salmonella enterica* serovar Typhimurium strain LT2 and *E. coli* K-12 (Macnab, 2003).

A comparison between the *E. coli* K-12 genome and that of *Salmonella enterica* revealed a small but puzzling difference in the flagellar gene repertoires (McClelland *et al.*, 2000). *E. coli* K-12 possesses an additional pair of divergent and promoter-less flagellar genes, *fhiA-mbhA*, which were absent from *S. enterica* (McClelland *et al*., 2000). These two genes are predicted to encode proteins homologous to the flagellar biosynthesis protein FlhA and the flagellar motor protein MotB, respectively. The two genes did not draw much attention at the time they were first noted. However, a few years later, genomic comparison of the K-12 genome with an unfinished genome of EAEC strain 042 revealed something surprising.

A large discrepancy, a gene cluster containing 44 genes in the region of the *fhiA-mbhA* gene pair, was found in the *E. coli* 042 genome but not in the K-12 genome (Figure 3-3) (Ren *et al.*, 2005). More strikingly, this gene cluster apparently encoded a novel and complete flagellar system, in addition to the conventional peritrichous flagellar system (Flag-1). This newly found gene cluster was designated as a second flagellar system in *E. coli*, Flag-2 (Ren *et al*., 2005). It was very surprising to identify two flagellar systems present in one *E. coli* genome.

The *E. coli* 042 Flag-2 gene cluster was compared with the other genome-sequenced *Escherichia/Shigella* and *Salmonella* strains available at the time (Ren *et al*., 2005). The comparison revealed that all other *Escherichia/Shigella* strains only contained two scar-genes *fhiA-mbhA,* the same as in *E. coli* K-12. However, all *S. enterica* strains lack the entire Flag-2, including the counterparts of the scar genes *fhiA* and *mbhA*. The two genes, *fhiA* and *mbhA,* exhibit over 95% identity to, but appear shorter than, their *E. coli* 042 counterparts. This identification allowed the reannotation of *fhiA* and *mbhA* as pseudogenes. Both K-12 strain MG1655 *fhiA* and *mbhA* genes appear to have been truncated such that the first 391 and 189 nucleotides of *E. coli* 042 *lfhA* and *lafU*, respectively (Ren *et al*., 2005). This point of deletion is also found in all other available *Escherichia/Shigella* genome sequences with the exception of *E. coli* 042. It was suggested that the entire Flag-2 locus was originally present in the last common ancestor of the species, but then deletions occurred (Ren *et al*., 2005). The genes *fhiA-mbhA* were the remnants of an ancestral Flag-2 locus.

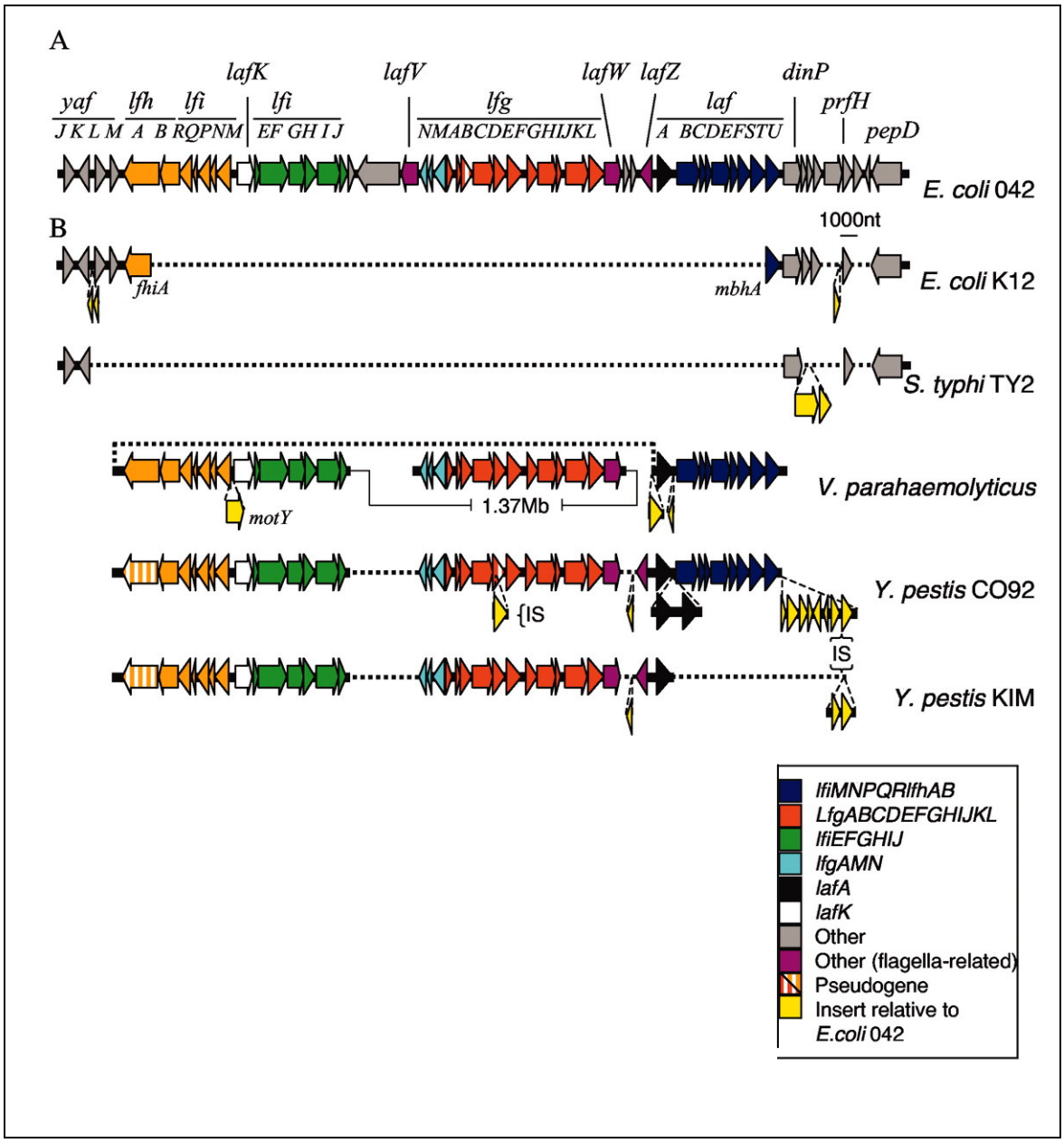

Figure 3-3 A, Schematic representation of the Flag-2 gene cluster in *E. coli* 042 (produced by Dr. Beatson). B, Schematic representation of the Flag-2 gene clusters of other bacteria. The solid black lines indicate the genome fragments in *E. coli* K-12 strain MG1655 (11.4 kb), *S. enterica* serovar *Typhi* Ty2 (4.9 kb), *V. parahaemolyticus* (11.7 kb region 1 and 23.2 kb region 2 from chromosome 2), *Y. pestis* CO92 (44.4 kb), and *Y. pestis* KIM (28.5 kb) that are equivalent to the lateral flagellar cluster of *E. coli* 042. Homologous genes are vertically aligned. Dotted lines indicate the absent genes relative to the *E. coli* 042 Flag-2 gene cluster (Ren *et al.*, 2005).

3.1.6.1 Flag-2 genes

From homology searches and genomic comparisons, *E. coli* 042 Flag-2 genes showed greater similarity to the lateral flagellar genes in *Vibrio parahaemolyticus* (Table 3-2), than to the conventional Flag-1 genes in *E. coli*. *V. parahaemolyticus* possesses dual flagellar systems adapted for movement under different circumstances, where a single polar flagellum propels the bacterium in liquid (swimming) and multiple proton-driven lateral flagella enable translocation over surfaces (swarming) (Stewart and McCarter, 2003). In addition to *V. parahaemolyticus*, *E. coli* 042 Flag-2 gene homologues were discovered in other genomes, including two annotated *Yersinia pestis* genomes, *Chromobacterium violaceum*, *Citrobacter rodentium*, and *Yersinia pseudotuberculosis* (Ren *et al*., 2005). The Flag-2 systems among these species are well conserved and located in the equivalent positions in the closely related genomes. This scenario suggests that the Flag-2 cluster was horizontally acquired by a common ancestor.

The apparent difference between *E. coli* 042 Flag-2 and *V. parahaemolyticus* lateral flagellar genes lies in the genetic organisation (Figure 3-3). The lateral flagellar gene system of *V. parahaemolyticus* contains a total of 38 genes in two distinct genetic regions (Stewart and McCarter, 2003). However, the Flag-2 genes are all organised in a unique chromosomal locus. Region 1 of *V. parahaemolyticus* 1 (~14 kb) includes all of the *flg* genes, encoding many of the structural proteins that are assembled to make the hook basal body structure, while Region 2 (~25 kb) includes the *fli* genes and *laf* genes, encoding the switch, motor, export-assembly, and flagellin genes (Stewart and McCarter, 2003). The majority of Flag-2 structural genes (i.e, *flg* genes counterparts) appear to be intact, as they

Table 3-2 Genes within the *E. coli* Flag-2 gene cluster.

Gene in *E .coli* 042	*E. coli* K-12	*V. parahaemolyticus*	Predicted function
lfhA/Ec042-0245	*fhiA*	$flhA_L$	Export, assembly
lfhB/Ec042-0246		$flhB_L$	Export, assembly
lfiR/Ec042-0247		$fliR_L$	Export, assembly
lfiQ/Ec042-0248		$fliQ_L$	Export, assembly
lfiP/Ec042-0249		$fliP_L$	Export, assembly
lfiN/Ec042-0250		$fliN_L$	Switch (C ring)
lfiM/Ec042-0251		$fliM_L$	Switch (C ring)
lafK/Ec042-0252		*lafK*	Regulatory
lfiE/Ec042-0253		$fliE_L$	Basal body component
lfiF/Ec042-0254		$fliF_L$	M ring
lfiG/Ec042-0255		$fliG_L$	Switch (C ring)
lfiH/Ec042-0256		$fliH_L$	Export, assembly
lfiI/Ec042-0257		$fliI_L$	Export, assembly
lfiJ/Ec042-0258		$fliJ_L$	Export, assembly
Ec042-0259			Cytidylyl transferase
Ec042-0260			Glycosyl transferase
lafV/Ec042-0261			Lysine-*N*-methylase
lfgN/Ec042-0262		$flgN_L$	Chaperone
lfgM/Ec042-0263		$flgM_L$	Anti σ^{28}
lfgA/Ec042-0264		$flgA_L$	P-ring addition
lfgB/Ec042-0265		$flgB_L$	Rod
lfgC/Ec042-0266		$flgC_L$	Rod
lfgD/Ec042-0267		$flgD_L$	Rod
lfgE/Ec042-0268		$flgE_L$	Hook
lfgF/Ec042-0269		$flgF_L$	Rod
lfgG/Ec042-0270		$flgG_L$	Rod
lfgH/Ec042-0271		$flgH_L$	L ring
lfgI/Ec042-0272		$flgI_L$	P ring
lfgJ/Ec042-0273		$flgJ_L$	Peptidoglycan hydrolase
lfgK/Ec042-0274		$flgK_L$	Hook-associated protein 1
lfgL/Ec042-0275		$flgL_L$	Hook-associated protein 3
lafW/Ec042-0276		VPA0275	Possible hook-associated protein
Ec042-0277			Unknown (COG4683)
Ec042-0278			Regulator
lafZ/Ec042-0279			Transmembrane regulator
lafA/Ec042-0280		*lafA*	Flagellin
lafB/Ec042-0281		*lafB*	Hook-associated protein 2
lafC/Ec042-0282		*lafC*	Chaperone
lafD/Ec042-0283		*lafD*	Chaperone
lafE/Ec042-0284		*lafE*	Hook length control
lafF/Ec042-0285		*lafF*	Unknown
lafS/Ec042-0286		*lafS*	σ^{28}
lafT/Ec042-0287		*lafT*	H^+ motor protein A
lafU/Ec042-0288	*mbhA*	*lafU*	H^+ motor protein B

code for proteins with functional counterparts in other organisms. Only one gene, *lfgC*, was identified as a pseudogene in *E. coli* 042. *lfgC* encodes a FlgC-like proximal rod protein and so is likely to be essential for the production of Flag-2 flagella by *E. coli* 042. The frameshift that occurs in the *lfgC* gene might cause the inactivation of Flag-2 in *E. coli* 042. Furthermore, genomic rearrangements exist between the two lateral flagellar systems of *E. coli* 042 and *V. parahaemolyticus*. The homologues to the *flg* genes of *V. parahaemolyticus* are inserted in between the counterparts of the *fli* genes and the *laf* genes in *E. coli* 042 (Figure 3-3).

The majority of *E. coli* 042 Flag-2 protein sequences exhibit 25 to 58% amino acid identity with their orthologs in the *V. parahaemolyticus* lateral flagellar system (Ren *et al*., 2005). Although *E. coli* 042 Flag-2 contains positional orthologs of all *V. parahaemolyticus* lateral flagellar genes, one exception is *motY*, which encodes a motor component. MotY is thought to localise to the outer membrane for flagella rotation, and is most commonly found in sodium-driven polar motility systems (Okabe *et al.*, 2002). $MotY_L$ in *V. parahaemolyticus* was suggested to have a function in mobility, but not in the production of flagella (Stewart and McCarter, 2003). On the other hand, the Flag-2 locus obviously contains additional genes relative to the lateral flagellar system of *V. parahaemolyticus*. Two additional CDSs (Ec042-0259 and Ec042-0260) are found between *lfiJ* and *lfgN* in the *E. coli* 042 Flag-2 cluster but are absent in the *V. parahaemolyticus* genome (Figure 3-3). Homology searches of these two non-flagellar gene products found that they are often clustered together, functioning as part of capsular polysaccharide biosynthesis (Ren *et al*., 2005). In addition, two further predicted coding sequences with no counterparts in the *V. parahaemolyticus* lateral flagellar system are found between *lfgL* and *lafA* in the *E. coli*

042 Flag-2 cluster, and are referred to as Ec042-0277 and Ec042-0278. Ec042-0277 encodes a protein with unknown function (COG4683), while Ec042-0278 encodes a product predicted to function as a transcriptional regulator. However, the transcription cascade of this gene is not clear.

The dual flagellar systems of *V. parahaemolyticus* can display different expression patterns during a change in environment. The lateral flagella are expressed only when the cell senses a surface environment, where LafK is responsible for lateral flagellar gene expression. In their study, Kim and McCater found that the polar flagellar gene, *flaK*, could be substituted by the lateral flagellar regulator LafK, but not the other way round (Kim and McCarter, 2004). Regulation by LafK in *V. parahaemolyticus* has been demonstrated to be dependent on interaction with *rpoN* (encoding σ^{54}) (Stewart and McCarter, 2003). RpoN has not previously been shown to be required for *E. coli* flagellar systems, whereas the FlhDC master operon regulates the peritrichous flagella of many swarming bacteria including *Escherichia coli* and *Salmonella enterica* serovar Typhimurium (Fraser and Hughes, 1999). However, the study of Flag-2 LafK suggests that the Flag-2 system is RpoN-dependent in *E. coli* 042 (Ren *et al*., 2005).

In the case of the two flagellar systems and the two non-flagellar T3SSs, one might ask whether any of the conserved components from one system might interact with and complement mutations in components of one of the other systems. If this happens, the components from other T3SSs can compensate for the defective genes in ETT2 or Flag-2, so that a functional second T3SS would be possible. Although one cannot entirely discount this possibility, two lines of argument count against it.

Firstly, there is an energy cost to producing these large multi-protein organelles. It is therefore unlikely that a bacterial cell would express two such systems at the same time. In addition, it also seems likely that each system switches on during a distinct environmental condition. For example, bacteria can develop distinct organelles of locomotion to adapt to different circumstances such as liquid and viscosity. In fact, as Zhang *et al.* have shown, regulatory components of the ETT2 system are able to inhibit synthesis of the LEE-encoded system (Zhang *et al.*, 2004). Thus, it is very unlikely that components of one system interact with components of another system.

Secondly, although the components of one system are homologous to their equivalents in another system, their homologies are relatively low. For example, even the most conserved component of these systems, the ATPase, shows only 50% identity at the protein level between the homologues from the two flagellar systems and only 44% identity is found between EscN and EivC. If protein-protein interactions in these systems are imagined as predominantly dependent on congruence between the molecular surfaces of interacting partners, even minor changes in sequence are likely to alter the surface properties of the proteins and hence their capacity for interactions. Thus, it seems unlikely, although not impossible, that proteins from one system can interact with homologues of their usual partners in another system.

3.1.7 Aims

Comparative genomics is a powerful tool in revealing bacterial genomic diversity. Using this approach, two cryptic gene clusters, ETT2 and Flag-2, were discovered. Both ETT2 and Flag-2 were surprisingly found to encode bacterial T3SSs (either non-flagellar or

flagellar T3SS). Although ETT2 has been investigated previously (Hartleib *et al*., 2003; Makino *et al*., 2003), only fragments of the gene cluster have been surveyed.

Three aims have been addressed in this project:

- To develop a rapid and low-cost PCR-based approach used for efficient genomic comparisons of two cryptic gene clusters.
- To understand how widespread the distribution of ETT2 and Flag-2 gene clusters are among *E. coli* strains drawn from the ECOR collection and representatives of selected pathotypes.
- To understand how the ETT2 and Flag-2 clusters have evolved.

3.2 Materials and Methods

3.2.1 Bacterial strains

The 72 ECOR strains were kindly provided by Thomas Whittam and stored in 10% glycerol at -80^{o}C. A description of the ECOR strains is available from Thomas Whittam's website (http://foodsafe.msu.edu/whittam/ECOR). Representatives of other pathotypes, including NMEC strain *E. coli* RS218, EAEC strain 042, ETEC strain H10407, EAEC strain EAEC25, UPEC strain CFT073, and *E. coli* strain K-12 and EPEC strain E2348/69 were kindly provided by Dr Henderson (University of Birmingham, UK), while an isogenic non-toxigenic derivative of the *E. coli* O157:H7 Sakai strain was a kind gift from Chihiro Sasakawa (University of Tokyo, Japan). All ECOR strains were incubated in LB agar/broth culture at 37^{o}C, while other pathogenic *E. coli* strains were cultured in LB plus appropriate antibiotics. In addition, four strains from *Escherichia* spp. other than *E. coli* (*Escherichia blattae*, *Escherichia fergunsonii*, *Escherichia hermannii*, and *Escherichia vulneris*) were purchased from the America Type Culture Collection. Genomic DNA was extracted using the DNeasy Kit (Qiagen, UK) and stored at -20^{o}C until required.

3.2.2 Primer design

Primers for the ETT2 study were designed using Primer3 software (Rozen and Skaletsky, 2000), which is incorporated into the *x*BASE server (http://xbase.bham.ac.uk/colibase). Primers were designed with an optimal length of 22 bp, optimal Tm of 60^{o}C, and optimal G+C content of 50%. A series of primers was designed to amplify each ~5 kb fragment that overlapped the regions of interest, with adjacent pairs overlapping by about 200 bp

(Table 3-3). Two additional pairs of primers were designed to identify two specific ETT2 genotypes seen in sequenced genomes, one pair covering the 8.7 kb deletion seen in the EPEC2 B171-8 sequence (primers named B171-8-like) and the other pair to identify an unoccupied *yqeG-glyU* intergenic region as seen in the UPEC CFT073 sequence (primers named No ETT2) (Table 3-3). These two pairs of primers, centered on the insertion-deletion (in-del) sites, were aimed to amplify ~200 bp and ~600 bp products, respectively. Short PCRs using the B171-8-Like and No ETT2 primer pairs were employed to identify the same ETT2 deletion patterns as seen for *E. coli* strain B171-8 and UPEC strain CFT073, respectively. For comparative purposes, a similar experiment was performed to investigate the distribution of the LEE cluster among all these *E. coli* strains. Eight pairs of PCR primers were designed to equally cover the whole LEE cluster with short overlaps of about 200 bp. However, only short PCRs were performed to examine the overlap regions.

Similarly, in the study of Flag-2, long overlapping PCR primers were designed to amplify eight ~5 kb fragments (primers named Flag 1-8) spanning the whole ~35 kb Flag-2 cluster, with each fragment overlapping its neighbours by a few hundred base pairs (see Table 3-4). One pair of primers (named fhiA-mbhA F/R) was designed to span the *E. coli* K-12 *fhiA-mbhA* remnant of Flag-2. Another two pairs of primers (named fhiA-flanking and mbhA-flanking) were designed to amplify the flanking regions of the EAEC 042 Flag-2 cluster. These primers were designed to amplify from within the genes of *fhiA* and *mbhA* to their respective flanking sites within the Flag-2 gene cluster, with products of ~600 bp and ~1,000 bp in size, respectively.

Table 3-3 Primers used to detect ETT2 and LEE gene clusters.

Primers	Sequence of primer (5'-3')	Gene name	Start Position
ETT2-1-F	GACCCAGCGCACCTGAGTAAGT	ECs3693	3697422
ETT2-1-R	AAGAGCGCAGTGTTTTGCCTGT	ECs3697	3702420
ETT2-2-F	GTGTGTTACCTCCGGGTCATCC	ECs3696	3701930
ETT2-2-R	CGCCGGACGATTTAAAGATGAG	ECs3701	3707421
ETT2-3-F	CGCACTGTGGATGCTCTGTCTT	ECs3700	3706496
ETT2-3-R	CGACTCATGGATTTGCACCAGA	ECs3706	3712367
ETT2-4-F	AATGACCAGGGACGAGCAAATC	ECs3705	3711838
ETT2-4-R	TATCCATTGCAAAACCCGCATT	ECs3711	3716991
ETT2-5-F	ATGTGCCTAACCCGCTCAAAAA	ECs3709	3715986
ETT2-5-R	ACCGACCCTGATCTGGTTGTAA	Intergenic region	3721489
ETT2-6-F	GGGAAATTATCAGCAAGCCATGA	ECs3719	3720991
ETT2-6-R	GCAGAAGAGAGTGGCAGCTGGT	ECs3726	3726481
ETT2-7-F	AGCGCGCCATTTACACGTATCT	ECs3726	3725748
ETT2-7-R	TGCACTTGATGCGAGTTGTTCA	ECs3732	3731740
ETT2-8-F	GGTGGGCAATGGAATTATGAGC	ECs3731	3731190
ETT2-8-R	AAACAGCGGCAGAAACCCACTA	Intergenic region	3736743
ETT2-9-F	TCGGTCACCTTTTTGCCAATCT	ECs3736	3736360
ETT2-9-R	TCCCGTTAATGGTGCATTCGAT	ECs3741	3741865
ETT2-10-F	AATTACGCCTGGCATTTGGTGT	ECs3740	3741579
ETT2-10-R	TCAGGCGAACGGTATCGTCATA	ECs3744	3746363
B171-8-Like-F	AGACCAGCTGCCACTCTCTTCT	ECs3726	3726458
B171-8-Like-R	GCTTGATTTAGGGGGAGAATCC	ECs3736	3736081
No ETT2-F	CCTGATCGTGGGTATCCTGT	ECs3702	3709433
No ETT2-R	GCTTGCATTTCCAGATTCGT	ECs3738	3737548
LEE-1-F	TTAAGGCATCGATGTGTCCTTC	ECs4602	4631996
LEE-1-R	GACATCTTGTTCTGCGCCATTA	ECs4595	4626190
LEE-2-F	TTTTAAACTGCAGCGACCTTACC	Intergenic region	4626387
LEE-2-R	GCCTGAGGATCTGTTTTTGCTT	ECs4586	4619398
LEE-3-F	CGGAACTCATCGAAAGGTGTTT	ECs4588	4619953
LEE-3-R	CAAAACAAACAAAAACGGAACG	ECs4579	4614954
LEE-4-F	GCATTATACGCACCAACTGCAT	ECs4579	4615154
LEE-4-R	CGACATCTTGCAACAATGAACA	ECs4572	4610159
LEE-5-F	CTGAATGACCGATGGTGCTAAG	ECs4573	4610825
LEE-5-R	CCCCATCGTGTACTCACCAATA	ECs4568	4605820
LEE-6-F	ACTTCCGCGATCAAGGTAAAAA	ECs4568	4606277
LEE-6-R	ACCAGGATTCGACTGCAGCTTA	ECs4559	4598150
LEE-7-F	CGACGATTTGGTCGTTGAATAA	ECs4559	4599009
LEE-7-R	CTCCCATGCCATAACCAATTTT	ECs4533	4580425
LEE-8-F	TTATCGGTCTCAGCACCCCTAT	ECs4533	4580516
LEE-8-R	CTTTCGCTCAATGATGTTCCTG	ECs4531	4575520

Table 3-4 Primers used to detect Flag-2 gene cluster.

Primers	Sequence of primer (5'-3')	Gene name	Start Position
fhiA-mbhA-F (FhiA-F)	GTTGATCGCCAGAATCATCATC	fhiA	286134
fhiA-mbhA-R (MbhA-R)	ATATTGCGGTTCTGGTCGTCTT	mbhA	322732
fhiA-flanking-F (LfhB-F)	TGAAAGTCAGGTGGAAGTGGTC	Ec042-0246	286987
fhiA-flanking-R (LfhA-R)	CCGATGGTCATCAGCACATACT	fhiA	286048
mbhA-flanking-F (Lafu-F)	GGATGAGACGGGCTGATTTTAT	mbhA	322507
mbhA-flanking-R (Lafu-R)	GGGACGTTTTTAGGCGTCTTTA	Ec042-0287	321991
Flg1-F	TTCAAACATATTGCGGTTCTGG	mbhA	322739
Flg1-R	GGCGTCACCATGACTTTTACC	Ec042-0281	317749
Flg2-F	CGTCCTGAATTTTGCTCATCTG	Ec042-0281	318241
Flg2-R	AGCAATGGAAGCTACCCTCAAG	Ec042-0275	312750
Flg3-F	GTCATCCTCAGACAGCATCACC	Ec042-0276	313333
Flg3-R	CGACAACCTGTATCTGGAAACC	Ec042-0270	307812
Flg4-F	GAGGAAACCGAGTTGTCGTTCT	Ec042-0271	308794
Flg4-R	GGTCGCAATCTGTGAGGAAATA	Ec042-0264	303332
Flg5-F	GCTGATTTGCAGATTCAGAAAGG	Ec042-0265	303937
Flg5-R	TGACAGCAAATAACGCAGTTCC	Ec042-0260	298459
Flg6-F	TCTGCCGGAAAATATTCAATCC	Ec042-0260	298912
Flg6-R	TTATTCCGCTGTGGAAAGATGA	Ec042-0254	293922
Flg7-F	GTTGAGGATCCCCTGCAACAT	Ec042-0255	294527
Flg7-R	TCATCAGAATCAGCACCTGGAT	Ec042-0249	289546
Flg8-F	CCGGGGATATTTTACCCATCTC	Ec042-0251	290205
Flg8-R	CCCCCATAATCTTCAACTCCAG	fhiA	284693

3.2.3 Tiling-path PCR scanning

In this project, a new PCR-based approach, termed Tiling-path PCR scanning (TP-PCR), was developed (Figure 3-4). TP-PCR is essentially a cut-down version of whole-genome PCR scanning, which consists of a series of inter-locking PCRs to construct a complete tiling path through the region of interest. This PCR approach is a combination of long and short PCRs. Long PCRs were performed initially to investigate the region in question. In the case of any negative long PCR result by a given primer pair, the relevant short overlaps were further surveyed by PCR using the same primer. If both long and short PCRs appeared negative, a deletion-scanning long PCR using the primers from the flanking regions was performed. The PCR-based nature of this method allows us to investigate the presence of ETT2 and Flag-2 in non-sequenced *E. coli* strains, including well-characterised and phylogenetically diverse strains drawn from the ECOR collection, and other representatives of selected pathotypes and examples of non-*coli Escherichia* strains.

For ETT2, Long PCRs were performed with a KlenTaq mixture (DNA Polymerase Technology Inc., St. Louis, Mo.) in the buffer supplied by the manufacturer. Each 20 µl Long PCR mixture contained 20 ng of genomic DNA as a template, 8 pmol of each primer, and a 250 µM concentration of each deoxynucleoside triphosphate (dNTP). Long PCR conditions were 30 cycles of 10 seconds at 94°C, 30 seconds at 62°C, and 10 minutes at 68°C, followed by a 10 minutes extension at 68°C. For short PCRs, each 20 µl reaction mixture contained 1U of *Taq* polymerase (Invitrogen, United Kingdom) in the buffer supplied by the manufacturer, 20 ng of genomic DNA, and a 250 µM concentration of each dNTP. Short PCR conditions were 30 cycles of 30 seconds at 94°C, 30 seconds at 62°C, and 30 minutes at 72°C, followed by a 7 minutes extension at 72°C. Supplementary Long

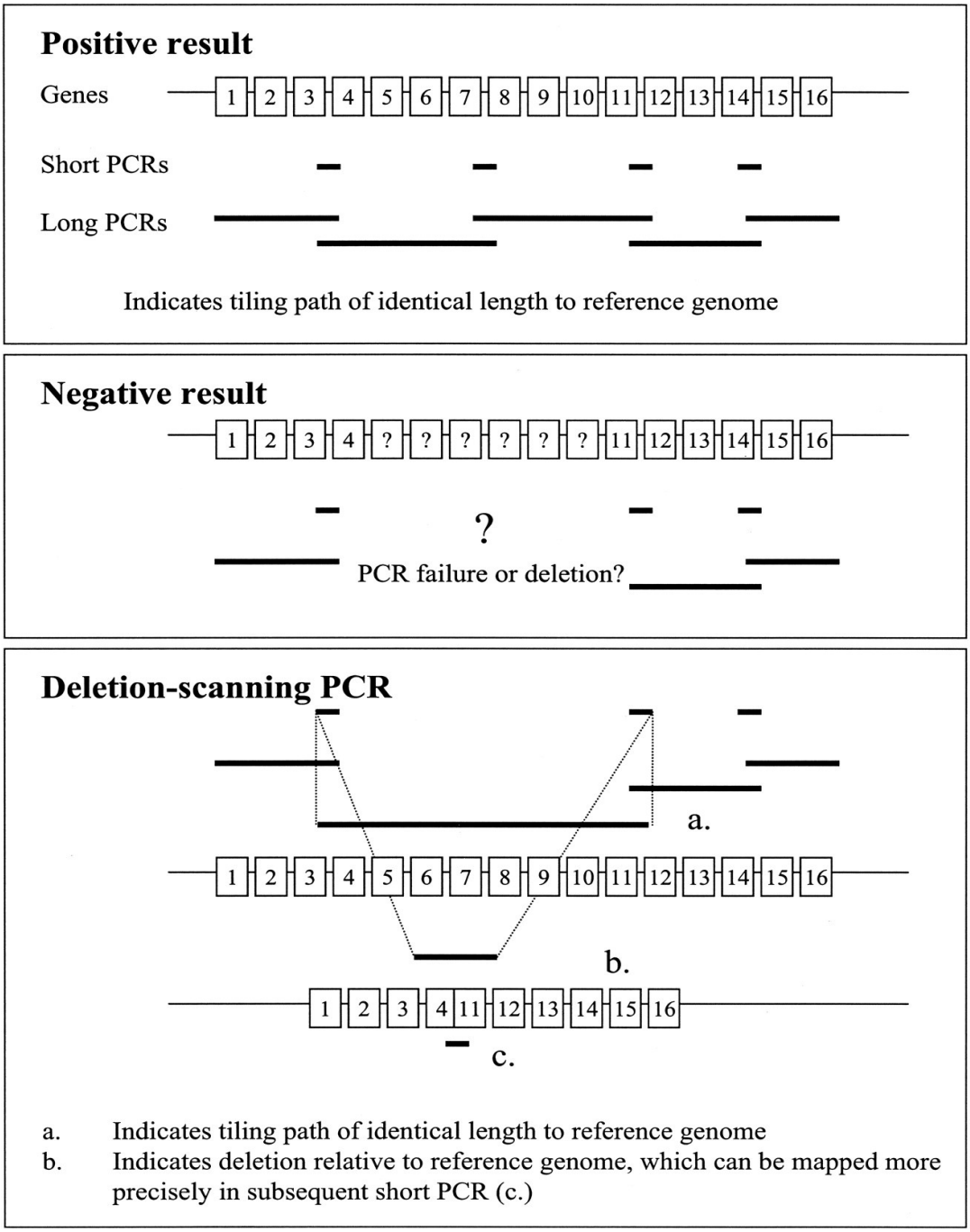

Figure 3-4 The strategy of Tiling-Path PCR (figure prepared by Professor Pallen).

PCRs were performed in a few cases, using TaKaRa LA *Taq* (Cambrex BioScience Ltd.) in the buffer supplied by the manufacturer. For these supplementary PCRs, each 20 µl reaction mixture contained 48 ng of template DNA, 4 pmol of each primer, a 200 µM concentration of each dNTP, and 1 U of TaKaRa LA Taq, and the reaction conditions were 30 cycles of 20 seconds at 96°C and 10 minutes at 69°C, with a 10 minutes extension at 72°C. All PCRs were carried out in a PTC-225 DNA Engine Tetrad thermal cycler (Genetic Research Instrumentation, Braintree, United Kingdom). Long PCR fragments were analysed by electrophoresis using a 0.7% agarose gel, while short products were analyzed using a 1.2% gel.

For the Flag-2 gene cluster, a three-stage PCR strategy was employed to scan isolates for the cluster. Initially, both primers of *fhiA-mbhA* were applied to the 72 *E. coli* strains in a conventional short PCR, aiming to detect the K-12-like *fhiA-mbhA* genotype. Short PCRs were performed using *Taq* polymerase (Invitrogen, United Kingdom) in the buffer supplied by the manufacturer. Second, the primer pairs of *fhiA*-flanking and *mbhA*-flanking were applied to all strains to detect the pairs of genes at the ends of the 042 Flag-2 gene cluster. Finally, tiling-path PCR was used to obtain a complete tiling path through the gene cluster in all Flag-2-positive tested strains. Long PCRs were performed by using TaKaRa LA *Taq* (Cambrex Bio Science, United Kingdom) in the buffer supplied by the manufacturer. Any negative results obtained by the long PCR were followed up by deletion-scanning PCRs as described above. Long and short PCR conditions were the same as those in the ETT2 study previously described.

3.3 Results

3.3.1 The ETT2 gene cluster

3.3.1.1 TP-PCR is used to identify genomic diversity

In this project, tiling-path PCR proved to be an efficient tool for studying the equivalent regions from many related strains. A set of ten long PCR primers was applied to all 72 ECOR strains, the laboratory strain K-12 and another 7 pathotypic *E. coli* strains. This PCR method, combining long and short PCRs, constructed a complete tiling path across the whole ETT2 region. In the majority of *E. coli* strains, some ETT2 genes were found but there appeared to be deletions relative to the complete gene cluster from EAEC 042. From the first round of long PCR scanning, the size and position of the deletions could be identified (Figure 3-5). In order to eliminate other possibilities, such as the primers annealing to the wrong sites, insertions or rearrangements at the region of interest or false-negative long PCRs caused by PCR conditions and primer point mutations etc, the deletion-scanning PCRs surveyed the deleted region using flanking primers. If the deletion-scanning PCR works, the strategy is particularly powerful in that the presence of a deletion can be confirmed by a positive PCR result, rather than a negative result, which could just indicate that the PCR failed for some reason. Deletion sizes were calculated by subtracting the deletion-scanning PCR product size from the expected product size from the Sakai strain. Small insertions, deletions or the original false-negative PCRs could be indicated, as well as the extent of these changes. However, this approach cannot resolve deletions followed by insertions, e.g., of IS elements. Using this method, the complete tiling path through the ETT2 gene cluster for 68 of 72 ECOR strains and for all of the representative pathotype strains has been constructed (Table 3-5).

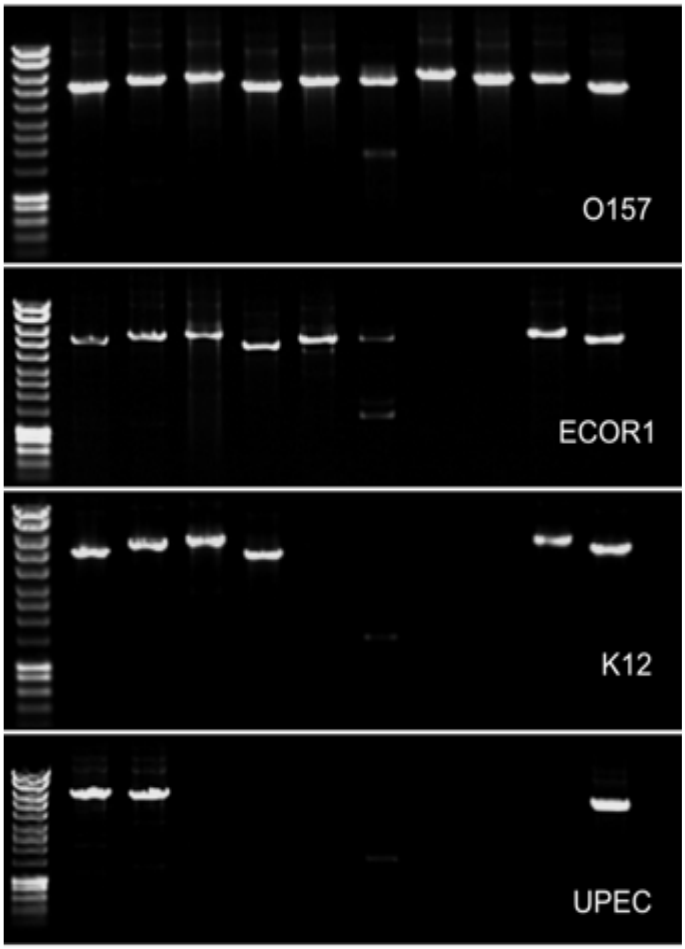

Figure 3-5 Gel images illustrating PCR results for the ETT2 gene cluster. Strains, from top to bottom: O157 Sakai strain (complete ETT2 gene set), ECOR1 (a B171-8-like strain with an 8.7 kb deletion), K-12 (14.6 kb deletion), and CFT073 (UPEC, with no ETT2, and 27.5 kb deletion). Lanes, from left to right: molecular weight markers (HyperLadder I; Bioline, UK), ~5 kb amplicons obtained with ETT2 TP-PCR primer pairs 1 to 10 (see Table 3-3 for details), negative control (DNA, no primers) (Ren *et al.*, 2004).

Table 3-5 Distribution of ETT2 gene cluster PCR fragments among *E. coli* strains.

Strain	Category	ETT2 LPCR pattern [a]	indel-specific PCR results [b]	Conclusions from deletion-spanning PCRs [c]
E2348/69	EPEC	++-------+	EPEC1-like	ETT2 absent
CFT073	UPEC			
RS218	NMEC			
ECOR 23	ECOR A			
ECOR 51, 52, 54, 55, 56, 57, 59, 61, 62, 63, 64, 65, 66	ECOR B2			
ECOR 4	ECOR A		EPEC1-like, with 2.4 kb insertion	ETT2 absent and something else inserted next to *glyU* or ~25kb deletion in ETT2 between 3F and 9R
H10407	ETEC	++++++--++	EPEC2-like	8.7 kb deletion as in EPEC2 strain B171-8 sequence
EAEC25	EAEC			
ECOR 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 15, 18, 22, 24, 25	ECOR A			
ECOR 26, 27, 28, 29, 30, 32, 33, 34, 45, 58, 67, 68	ECOR B1			
ECOR 39	ECOR D			
ECOR 45	ECOR B1			
ECOR 14	ECOR A		EPEC2-like, with 1 kb insertion	8.7 kb deletion as in B171-8 sequence, with ~1 kb insertion
ECOR 13, 20, 21	ECOR A	++-------+	Negative	~23 kb deletion between 3F and 9R
ECOR 19				~30 kb deletion between 2R and 10F
ECOR 53	ECOR B2			~30 kb deletion between 3F and 9R, i.e. ETT2 absent
ECOR 60				
ECOR 69	ECOR B1			Unresolved deletion between 2R and 10F
ECOR 16	ECOR A	++-+-----+		Unresolved deletion between 2R and 4F ~18 kb deletion between 4R and 9R
ECOR 17	ECOR A	+++------+		22 kb deletion between 4F and 9R
ECOR 38	ECOR D	++-----+++		~21 kb deletion between 3F and 7R
ECOR 35	ECOR D	+++-++----		No deletion between 3R and 5F, Unresolved deletion after 6R
ECOR 31	ECOR E	++----++++		~15 kb deletion between 2R and 7F
K-12	Laboratory	++++----++		18 kb deletion as in K-12 sequence

ECOR 36	ECOR D	+++-++--+-		No deletion between 3R and 5F, 8 kb deletion between 7F and 8R Unresolved deletion after 9R
ECOR 40	ECOR D	+++-+--+++		No deletion between 3R and 5F 6 kb deletion between 6F and 7R
ECOR 41				No deletion between 3R and 5F 4 kb deletion between 6F and 7R
ECOR 70	ECOR B1	++++++---+		~5 kb deletion between 6R and 9R
ECOR 72	ECOR B1	++++++--++		~5 kb deletion between 7F and 8R
ECOR 71				~8 kb deletion between 6R and 9F
ECOR 42, 43	ECOR E			
ECOR 37	ECOR E	+++++-++++		~3 kb deletion between 6R and 7F
ECOR 48	ECOR D	+++-++++-+		No deletion between 3R and 5F, No deletion between 8R and 10F i.e. **EHEC1-like**
ECOR 44, 46, 47, 49, 50		+++-++++++		No deletion between 3R and 5F i.e. **EHEC1-like**
O157	EHEC	++++++++++		no deletions detected, i.e **EHEC1-like**
042	EAEC	++++++++++		no deletions detected, i.e **EHEC1-like**

[a] The ++----++ notation indicates the Long PCR results from ten long PCR reactions; "+" represents the positive PCRs in that region, and "-"represents the negative PCRs.

[b] An ~200 bp PCR was used to detect a deletion point identical to that seen in B171-8 (EPEC2-Like), while an ~600 bp PCR was used to detect a CFT078-like vacant ETT2 insertion point (EPEC1-Like). Negative means that both indel-specific PCRs failed to give a product.

[c] When the indel-specific PCRs failed to resolve a deletion, deletion-scanning long PCRs were performed with forward and reverse primers from the sets flanking the deletion. The full ETT2 gene clusters are indicated in bold.

Four strains, ECOR-16, -35, -36, and -69, remained unresolved by PCR alone, after using various combinations of primers. It seems likely that large insertions or rearrangements might have occurred in these strains, but to be certain, more detailed information is required, either from DNA sequencing or other fragment analysis.

3.3.1.2 The distribution of the ETT2 locus is largely congruent with *E. coli* known phylogeny

The pattern of the results from the ETT2 TP-PCR was very striking. Firstly, unlike the previous study (Makino *et al*., 2003), the ETT2 gene cluster was surveyed by sampling a wide diversity of examples from the *E. coli* species, including both pathogenic and commensal strains. Surprisingly, the ETT2 locus is present in the majority of these *E. coli* strains, either complete or in part. Secondly, using the known phylogeny ECOR strains as a backbone, the distribution of ETT2 was surprisingly found to be non-random, but instead largely congruent with the major phylogenetic divisions (A, B1, B2, D and E), as deduced by multilocus enzyme electrophoresis (MLEE) and other methods (Lecointre *et al*., 1998) (Figure 3-6). The majority of *E. coli* strains fell into only three predominant ETT2 patterns, i.e. divided into three ETT2 genotypes: "no ETT2" genotype, 8.7 kb deletion genotype, and full ETT2 genotype (Figure 3-6). It is very interesting to contemplate how these ETT2 genotypes arose. The good correlation of the ETT2 locus with the ECOR phylogenetic divisions hints that the ETT2 cluster could be a new phylogenetic marker.

The 8.7 kb deletion was seen in the already sequenced ETT2 cluster from an EPEC2 strain, B171-8 (Makino *et al*., 2003). This deletion was centred on 7 bp repeats (CC/ATCATT), suggesting a mechanism for the deletion. Surprisingly, this 8.7 kb deletion represents the

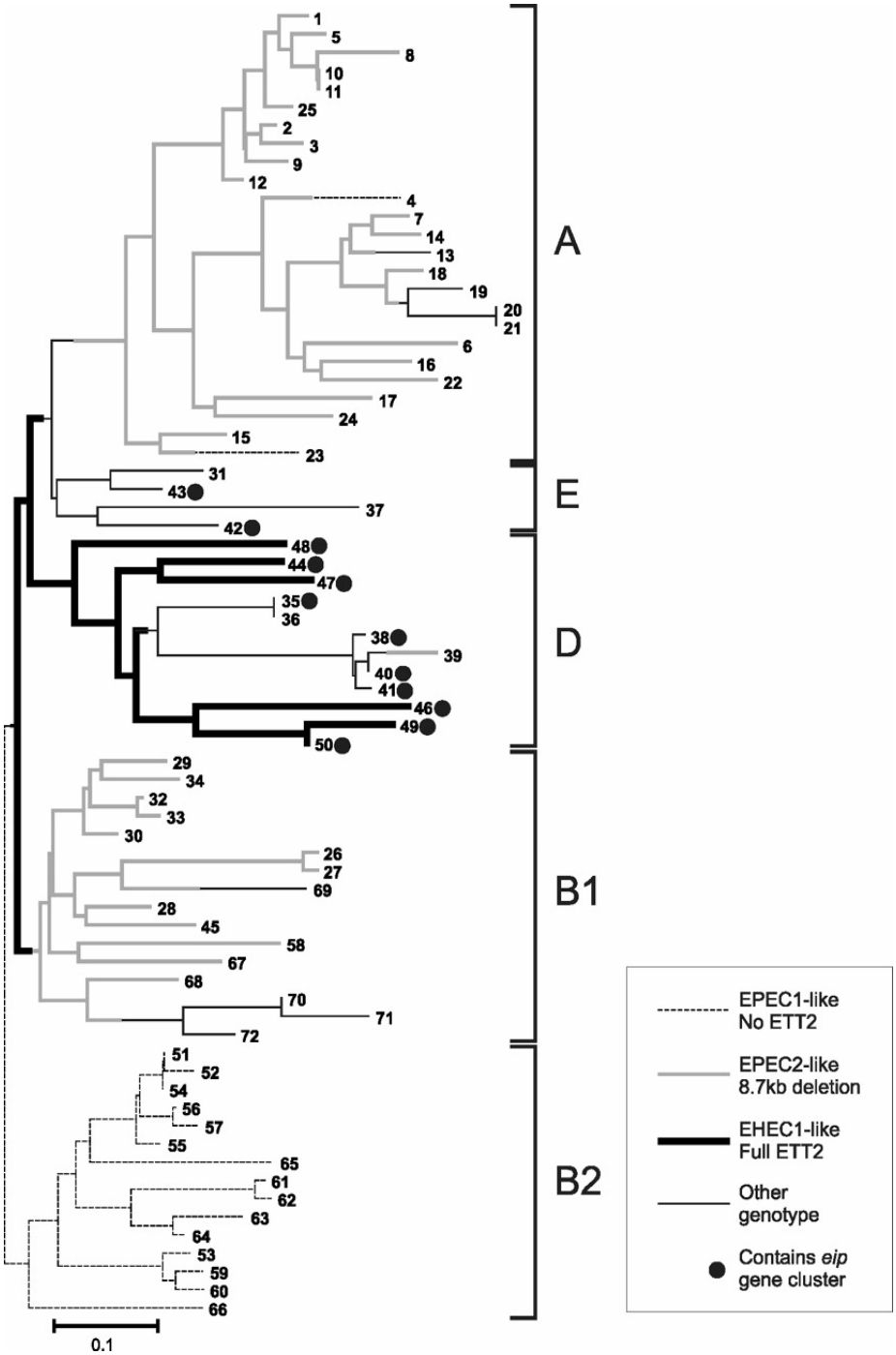

Figure 3-6 TP-PCR results superimposed on the phylogenetic structure of *E. coli*. The tree was constructed by Dr Chaudhuri, obtained by neighbor-joining analysis of the ECOR MLEE data (available at http://foodsafe.msu.edu/whittam/ecor). Branches containing one of the three most common genotypes are highlighted by bold, dotted or grey lines. Filled circles indicate strains with *eip* clusters (Ren *et al.*, 2004).

most common ETT2-genotype, including the majority of the A and B1 ECOR groups. In order to confirm whether the deletion was identical (to within a few base pairs) to that in B171-8, short PCRs of ~ 200 bp across the deletion site were performed on all 72 ECOR strains and other representative pathogenic *E. coli* strains. The results showed that 17 of 25 ECOR group A strains, 12 of 16 ECOR group B1 strains, and the pathogenic strains H10407 (ETEC) and EAEC25 (EAEC) have identical deletion points (Figure 3-7). These strains are classified as an EPEC2-like genotype.

The "no ETT2" genotype accounts for the second largest group, mainly found in ECOR group B2 strains. In these strains, only those long PCRs annealing to the *E. coli* backbone regions worked. The absences were confirmed by performing a short 600bp PCR across the ETT2 insertion point (ECs3702-ECs3738), as seen in UPEC strain CFT073. 13 of 15 ECOR group B2 strains, two strains (ECOR 4 and 23) from ECOR group A, and the pathogenic strains CFT073 (UPEC) and RS218 (NMEC) showed identical deletion points (Figure 3-7). A complementary experiment examined the ETT2 region from the sequenced EPEC 1 strain E2348/69. The TP-PCR pattern was the same as seen for UPEC CFT073, also with the same deletion points (data not shown). In all these strains the entire ETT2 region is absent, and they are classified as an EPEC1-like genotype.

Apparently intact ETT2 gene clusters were found in 6 of 10 ECOR group D strains, to which EAEC 042 was shown to be most closely related in a previous study (Escobar-Paramo *et al*., 2004). Within these six ECOR group D strains, four are commensals, suggesting that the intact ETT2 is not necessarily associated with a virulent phenotype.

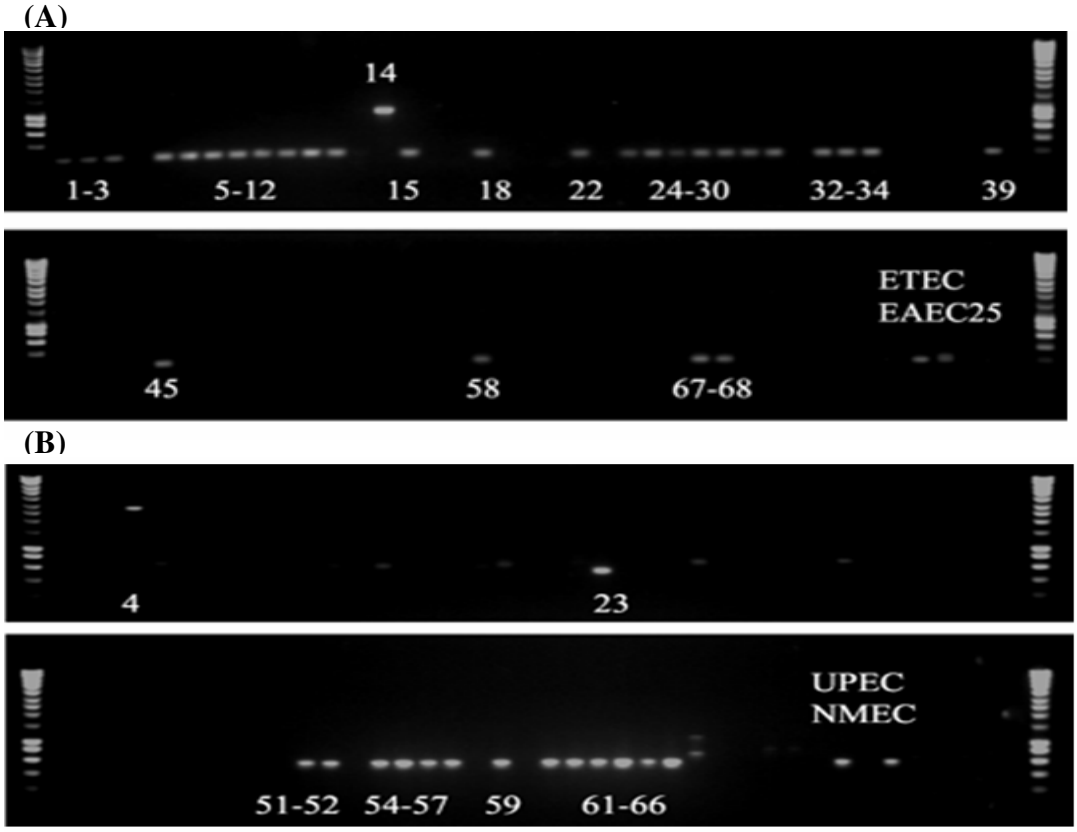

Figure 3-7 Indel-specific short PCRs for detecting (A) EPEC2- like, and (B) EPEC1-like genotype strains. The first two gels show the results for ECOR strains from 200 bp PCRs across the deletion seen in strain B171-8, using primer pair "B171-8 like" (Table 3-3). The second two gels show the results for ECOR strains from 600 bp PCRs across the ETT2 insertion site with primers "no ETT2" (see Table 3-3 for details). Positive results are labeled with ECOR strain numbers or pathotypes.

Nearly intact ETT2 gene sets were found in the remaining ECOR group D strains and ECOR group E strains. The two sequenced EHEC O157:H7 strains (EHEC 1) have been previously shown to be closely related to ECOR group E (Escobar-Paramo *et al*., 2004). The strains containing an intact ETT2 are named EHEC1-like genotype. Interestingly, the distribution of the *eip* gene cluster was found to correlate well with those EHEC1-like strains. The 13 *eip*-positive ECOR strains encompassed 11 of the 12 group D strains and 2 of 5 E strains (Figure 3-6).

In addition to the ECOR collection, both the ETT2 and *eip* gene clusters were surveyed in a collection of 43 freshly collected local blood culture isolates of *E. coli* and 36 freshly collected local urine contaminants from patients with no laboratory evidence of urinary tract infections (presumably representing commensal strains of *E. coli*). A series of short PCRs using the overlapping primers were performed (A. Fivian, BSc project). ETT2 was detected in 16 of the 43 (37%) bloodstream isolates and 10 of the 36 (28%) commensal isolates (data not shown). These results indicated the high prevalence of ETT2 in human *E. coli*. The *eip* gene cluster was present in 7 of the 15 ETT2-positive bloodstream isolates, and 6 of the 10 ETT2-positive commensal isolates.

Surprisingly, the *eip* cluster was present in all six ECOR strains with EHEC 1-like genotypes but absent in the two genome-sequenced O157:H7 strains. From the studies on both the ECOR strains and the clinical strains, the *eip* cluster was not found in any isolate that lacked the ETT2 gene set. Furthermore, both the ETT2 and *eip* gene sets have no obvious link to bacterial virulence. However, the distributions of the ETT2 and *eip* gene clusters are correlated, suggesting a functional relationship between the two regions.

3.3.1.3 The ETT2 locus was acquired after the divergence of ECOR group B2

The distribution of the ETT2 locus was found to correspond well with the phylogeny as determined in previous studies (Escobar-Paramo *et al*., 2004) (see Figure 3-6). The "known" phylogeny has been determined by MLST, amongst other methods. Based on several lines of evidence (Lecointre *et al*., 1998), the ECOR B2 group is thought to have been the first to diverge from the other *E. coli* lineages, followed by the divergence of group D. Consistent with the MLST-derived branching pattern, the ancestral status was assigned to the "no ETT2" genotype - all 15 ECOR strains belonging to the B2 group, EPEC1 strain E2348/69, which is known to be extremely divergent from other groups, and UPEC strain CFT073, which is closely associated with the B2 group. Except for two ECOR group B2 strains (ECOR-53 and -60), all these strains were confirmed to have the same deletion points using a short PCR (Figure 3-7), suggesting the ETT2 locus was inserted into the lineage only once.

Both EPEC and EHEC exist as two lineages, which have been predicted by previous studies based on the MLEE data (Donnenberg and Whittam, 2001). Although EPEC1 and EHEC1 are highly divergent, EPEC1 appears to have diverged first. The authors suggested that EHEC1 strains diverged from other *E. coli* strains after the divergence of the EPEC1 strains. In this study, the EPEC1 strain falls into the "no-ETT2" genotype, which is closely affiliated with the B2 group, whereas EHEC1 strain falls into the complete the ETT2 genotype, which is close to the D group. These results were in line with the fact that the B2 group is the deepest branching group followed by the D group, as predicted by MLST (Lecointre *et al*., 1998). Using an MLST phylogenetic tree as a backbone, the ETT2 locus was suggested to enter an ancestral *E. coli* strain sometime after the divergence of the B2

group (and the EPEC1 strain) but before the divergence of the D group (and the EHEC1 strain).

The presence of the ETT2 genes was examined in four non-*coli Escherichia* spp. (*E. blattae*, *E. fergunsonii*, *E. hermannii*, and *E. vulneris*) by performing overlapping short PCRs. These species are most closely related to *E. coli*. The results showed that all four *Escherichia* spp. possessed no ETT2 (data not shown), suggesting that the ETT2 gene loci has never entered into other species rather than *E. coli*. Furthermore, the ETT2 gene cluster was always found at the same chromosomal location, within the *yqeG-glyU* intergenic region. In contrast, LEE can be inserted into *selC*, *pheV*, or *pheU* (Jores *et al.*, 2004). These results support the idea of a single insertion event subsequent to the divergence of group B2. Contrary to the previous claims (Makino *et al*., 2003), the ETT2 gene cluster was not an insertion in EHEC strains but a deletion in *E. coli* K-12 strains, with a remnant persisting in the *E. coli* K-12 region.

The whole ETT2 locus entered into the species *E. coli* only once, but the accumulated deletions occurred afterwards. The deletions were possible because that the ETT2 locus does not provide any selective advantage, and the genetic decay of this region was forced. However, in a study on ETT2 regulators, the authors surprisingly found that two regulatory genes, ECs3720 or *etrA* and ECs3734 or *eivF*, from the ETT2 cluster in EHEC O157:H7 have negative regulation effects (Zhang *et al.*, 2004). The deletion of these two genes leads to greatly increased secretion of proteins encoded by the LEE cluster and to increased adhesion to human intestinal cells (Zhang *et al*., 2004). This suggested that ETT2 may have retained a regulatory role and cross talk to the other T3SS's. Therefore, ETT2 may

have had an ancestral function that has been lost in some or all strains. It appears that the effects of regulatory genes can outlive widespread decay of other genes in a functionally coherent gene cluster in order to increase the bacterial infection (Zhang *et al*., 2004).

3.3.1.4 The ETT2 locus was not a functional T3SS

Several findings from this study suggest that the ETT2 locus does not encode a fully functional virulence-related T3SS in *E. coli* strains. The ETT2 genes were found equally distributed among pathogenic and commensal strains, in that 50 of 72 ECOR strains contain ETT2-associated genes. For comparison, the ECOR collection for fragments of the LEE cluster was surveyed by using five short PCRs. Only two strains were positive for LEE fragments, namely ECOR25, an ECOR group A strain from a healthy dog (four of five fragments were positive), and ECOR37, an ECOR group E strain from a healthy marmoset (all five fragments were positive) (data not shown). Unlike ETT2, the LEE cluster was mostly absent from the genomes of commensal *E. coli* strains.

In addition, many ETT2 genes carry mutations and deletions, raising the possibility that a functional T3SS has been degraded. Of the sequenced *E. coli* genomes, an intact ETT2 gene cluster was only seen in EAEC strain 042. A comparison of the 042 ETT2 genes with their homologues in Spi-1 could not identify any inactivating frameshift mutations. In both EHEC strain EDL933 and Sakai strains, though the complete ETT2 gene clusters were present, pseudogenes and frameshift mutations were apparent that would prevent the expression of a functional ETT2-encoded T3SS (see Table 3-1). The *eip* locus is also absent from two of EHEC strains. EAEC strain 042 may retain the ability to encode a functional T3SS, but its *in vivo* role is as yet unknown.

3.3.2 The Flag-2 gene cluster

Similar to the study on the ETT2 cluster, the distribution of the Flag-2 gene cluster was surveyed among the ECOR collection and another seven representatives of *E. coli* strains. In this study, three rounds of PCR were employed aimed at identifying the strains containing the Flag-2 cluster. Firstly, PCR across the *fhiA-mbhA* boundary was performed on all strains. A positive PCR result was obtained for 57 strains (~80%) from the ECOR collection, the *E. coli* K-12 strain and five pathogenic *E. coli* (*E. coli* RS218, *E. coli* CFT073, ETEC strain H10407, EAEC strain 25 and *E. coli* O157:H7 Sakai strain). This means they all possessed the same two-gene scar seen in K-12 (Figure 3-8A). Fifteen ECOR strains (~20%) and EAEC strain 042 gave a negative result with this PCR (ECOR-1, -3, -4, -5, -12, -17, -24, -35, -36, -48, -49, -50, -64, -65, and -67). A negative PCR result can occur for several reasons, such as point mutations in the primers, poor quality primers, or genetic changes like a large insertion or rearrangement etc. There is also the possibility that a full length Flag-2 gene cluster is harboured at this site.

In order to know whether the Flag-2 genes were present in this region, a second round of PCR was performed that targeted both ends of the full Flag-2 cluster. Two pairs of primers were designed to obtain two PCR products, spanning from either *fhiA* or *mbhA* to their flanking genes within the Flag-2 locus. Very interestingly, these PCRs provided complementary results to the first round: the 57 ECOR strains and another 6 representative *E. coli* strains with the K-12 genotype showed negative results as expected, whilst the other 15 ECOR strains and EAEC strain 042 gave positive results (Figure 3-8 B). The results suggest that 15 ECOR strains might possess the full Flag-2 gene cluster like EAEC 042, although they do not eliminate the possibility of an internal deletion in the Flag-2 cluster.

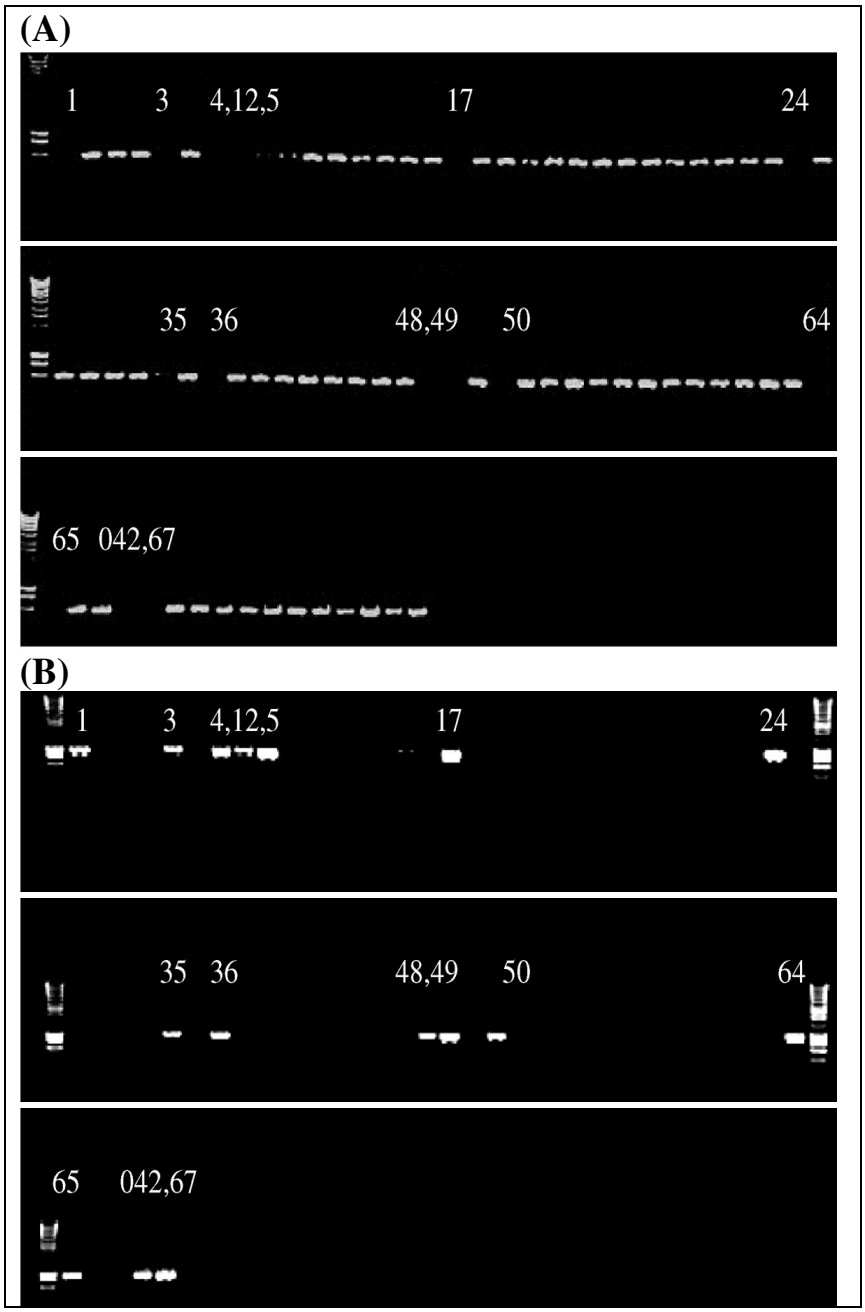

Figure 3-8 Gel images illustrating the identification of the Flag-2 genes of *E. coli*. (A) PCR scanning with the primer pair fhiA-mbhA. A 600 bp PCR product indicates that *fhiA* and *mbhA* are fused, as in *E. coli* K-12. Negative results suggest the presence of intervening sequence between *fhiA* and *mbhA*. (B) PCR scanning with primers fhiA-flanking. An ~1,000 bp PCR product indicates the presence of a full-length *fhiA* flanking gene within the Flag-2 locus, as in *E. coli* 042. PCR was carried out on all 72 ECOR strains plus *E. coli* RS218, EAEC strain 042, *E. coli* CFT073, ETEC strain H10407, EAEC strain 25, *E. coli* K-12, and *E. coli* O157:H7 Sakai. PCR mixtures were loaded on a 1.0% agarose gel alongside HyperLadder I MW markers (Bioline, United Kingdom). Lanes corresponding to Flag-2-positive strains are labeled according to ECOR strain number or pathotype.

Next, interest was focused on the 15 ECOR strains to find out whether they have an EAEC 042-like genotype. Eight pairs of long PCR primers were used to survey the ~35 kilobase cluster. Using TP-PCR, a complete tiling path through the entire Flag-2 locus was constructed for these 15 ECOR strains and EAEC strain 042 (Table 3-6). Most long PCRs were positive for the 15 ECOR strains. According to the long PCR pattern, the ECOR strains that were found to harbour the Flag-2 gene cluster could be divided into four types: type 1, including ECOR-4, -49, and -50; type 2, including ECOR-1, -3, -5, -12, -17, -24, -64, -65, -67; type 3, including ECOR-35 and -36; and type 4, including ECOR-48 (Table 3-6). Although none of them were positive for all eight long PCRs, the following deletion-scanning PCR using the flanking primers showed that actually they all possess the complete Flag-2 region (Figure 3-9). Unlike the ETT2 gene cluster, no large scale insertions, deletions, or rearrangements occurred in any of the Flag-2 clusters from these 15 ECOR strains compared to the 042 genotype. The existence of four types of long PCR patterns suggested they might have some point changes compared to the EAEC 042 genome, like SNPs at the primer sites, which would produce negative long PCRs.

Fifteen ECOR strains, accounting for 20% of the ECOR collection, contain the full Flag-2 gene cluster. The majority of *E. coli* strains have lost almost all the Flag-2 genes, but possess an identical fusion gene (*fhiA-mbhA*) between the remnants of the counterparts from the ends of the Flag-2 cluster. Although most strains lack the Flag-2 cluster, a sizable minority (about one fifth) of *E. coli* strains has retained the Flag-2 cluster. It was assumed that the Flag-2 cluster was present in the last common ancestor of all *E. coli* strains but had undergone a single deletion. However, unlike the ETT2 gene cluster, the distribution of Flag-2 is not congruent with the established phylogeny of *E. coli* strains. Flag-2 gene

Table 3-6 *E. coli* strains from the ECOR collection that possess an apparently intact Flag-2 gene cluster.

Isolate	Group	H antigen*	Flag-2 long PCR pattern	ETT2 genotype
ECOR-4	A	HN	++-+++++	Absent
ECOR-49	D	NM		Complete
ECOR-50	D	HN		Complete
ECOR-1	A	HN	+--+++++	Partial
ECOR-3	A	NM		Partial
ECOR-5	A	NM		Partial
ECOR-12	A	H32		Partial
ECOR-17	A	NM		Partial
ECOR-24	A	NM		Partial
ECOR-64	B2	NM		Absent
ECOR-65	B2	H10		Absent
ECOR-67	B1	H43		Partial
ECOR-35	D	NM	+---++++	Partial
ECOR-36	D	H25		Partial
ECOR-48	D	HM	++--++++	Complete

*HN, non-typeable with standard antisera; NM, non-motile strain; HM indicates a form of non-typeable motile strain in which multiple H antisera reacted.

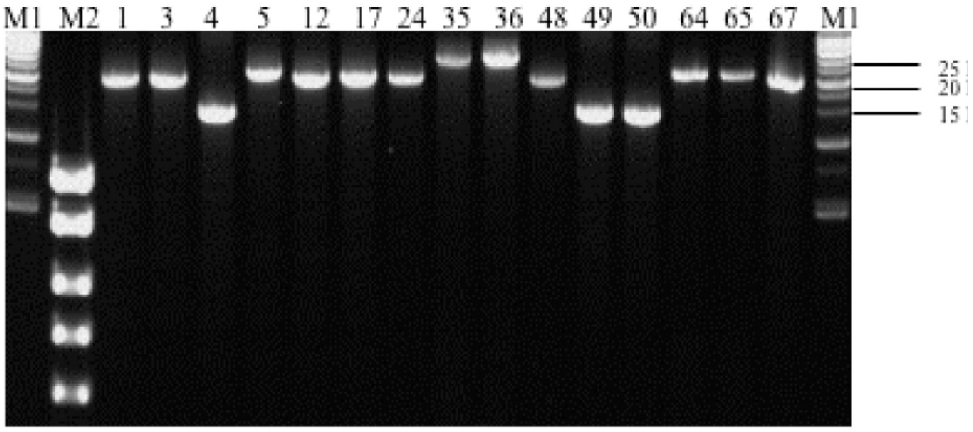

Figure 3-9 Deletion-scanning PCR of 15 ECOR strains for the presence of the Flag-2 gene cluster. Lanes M1 and M2 contained a high molecular weight DNA marker (Invitrogen) and the positions of the 15, 20, and 25 kb markers are indicated on the right. The ECOR strain numbers (see text and Table 3-6) are indicated at the top of the gel.

clusters are more common in ECOR group A but are randomly distributed throughout all four groups. It remains a mystery why the Flag-2 cluster was purged out from many *E. coli* strains. Additionally, short PCRs using overlapping primers were performed on the four non-*coli Escherichia* spp. and showed that all four of them possessed the K-12-like genotype. The presence of the Flag-2 cluster at identical sites in *E. coli* and its close relative non-*coli Escherichia* spp. and *Citrobacter rodentium*, combined with its absence from *S. enterica*, suggests that it was acquired by horizontal gene transfer after the former three species diverged from *Salmonella*.

3.4 Discussion

3.4.1 The need to maintain an energetic program of genome sequencing

The studies on both the ETT2 and Flag-2 gene clusters presented here emphasise the importance of maintaining an energetic program of genome sequencing within a taxonomic group. ETT2 was discovered in the second published *E. coli* genome (Perna *et al*., 2001), but its presence in whole or part in the majority of genome-sequenced *Escherichia/Shigella* strains, including a remnant in the initial K-12 genome, was not immediately apparent until multiple sequences were available. A fragment of sequence from the EPEC2 strain O111: NM B171-8, which was not subjected to complete genome sequencing until recently, suggested the presence of an 8.7 kb deletion within ETT2, centred on a 7 bp repeat. This deletion pattern has been found across ECOR groups A and B1 and represents the most prevalent form of ETT2. Even more strikingly, Flag-2 was not discovered in the first 10 *Escherichia/Shigella* genome sequences. Without a comparative viewpoint, the presence of the *fhiA* and *mbhA* genes in K-12 was puzzling. It was not until sequencing of the 042 genome that their presence as remnants of a complete flagellar system was determined.

3.4.2 A single strain like K-12 is not the archetype for a whole species

The study on the ETT2 cluster revealed that an identical 14.6 kb deletion within ETT2 was present in both sequenced *E. coli* K-12 strains. Initially, there was an implicit assumption that the non-pathogenic strain K-12 represented the ancestral state, and that the genomes of pathogenic *E. coli* would consist of the K-12 genome with additional pathogenicity determinants. This assumption led to several misunderstandings about the origins of the

ETT2 gene cluster. Though *E. coli* K-12 has long been used as a model strain and possesses one of the smallest *E. coli* genomes, it is unsafe to conclude that it represents the ancestral state for the whole species or a minimal *E. coli* genome. In fact, remnants of both islands studied in this chapter, an ~12 kb fragment of ETT2 and the two gene scar representing the edges of Flag-2, were found in both *E. coli* K-12 genomes.

Prior to this study, it was possible that ETT2 was formed by a single insertion with subsequent gene loss or by assembly of the largest ETT2 gene cluster from the smaller gene clusters by gene acquisition. The latter hypothesis now seems less plausible. Firstly, conservation in gene order between ETT2 and Spi-1 and other T3SS genes was observed. This is inconsistent with multiple independent gene acquisitions, which would be unlikely to keep the same gene order in several generations. Secondly, it was observed that the indels often had the same deletion boundaries (Figure 3-7). These boundaries are often marked by truncated genes, suggesting deletion rather than insertion. This observation is contrary to the initial hypothesis that the ETT2 gene cluster is an insertion into the *glyU* tRNA site in *E. coli* O157:H7 relative to *E. coli* K-12.

3.4.3 Tiling-Path PCR is an effective technique for comparative genomics

In this study, tiling-path PCR was developed and used to investigate two large pathogenic islands (ETT2 and Flag-2, approx. ~50 kb and ~35 kb in size, respectively). These methods allow the study of the distribution of entire pathogenicity islands by performing several long inter-locking PCRs. Previous studies (Hartleib *et al*., 2003; Makino *et al*., 2003) have

given a misleading picture since they only examined discontinued fragments of the region of interest. Tiling-path PCR can provide information about the full length of the island.

The advent and development of long range PCR makes the study of large genetic regions more effective and less expensive(Barnes, 1994). Tiling-path PCR can be considered a reduced form of whole-genome PCR scanning, pioneered by Tetsuya Hayashi to study *E. coli* O157 genomic diversity (Ohnishi *et al*., 2002). In whole-genome PCR analysis, hundreds of primers aim to amplify about 10-20 kilobase fragments overlapping with the adjacent fragment at both ends by a few hundred base pairs. Recently, software like GenoFrag (Ben Zakour *et al.*, 2004) was developed to facilitate the design of primers optimised for whole-genome PCR scanning (also see Chapter 5). With such software packages, the primers are automatically obtained when the user inputs several fixed parameters, such as the primer length, G+C%, primer Tm, and the lengths of overlaps. If most of these primers are designed to anneal to the regions on the bacterial chromosomal backbone, they are easily applied to all the strains within the same species. This could greatly reduce the cost and time for comparisons of large numbers of strains within a species.

The TP-PCR approach provides information on gene order and genomic organisation and can rapidly identify regions of a test genome that are co-linear with the reference genome. However, once a certain segment failed to yield a PCR amplicon, one could not discern whether a chromosomal deletion or large insertion or rearrangement had occurred. Under these circumstances, the use of a single-primer PCR walking approach (Karlyshev *et al.*, 2000) could be beneficial, as it allows direct sequencing from the known chromosome into

the unknown region (refer to Chapter 4). These methods will be of use in investigating the gene content of large strain collections whilst the complete genome sequencing of such collections remains impractical or economically unrealistic.

3.4.4 Sampling the full range of phylogenetic diversity within a species

One would not have expected that the distribution of the ETT2 locus would be so well organised in the *E. coli* strains. The surprise came from two important aspects of this project: first, sampling a wide diversity of samples from the species *E. coli*, including both pathogenic and commensal strains; second, a complete tiling-path was constructed across this region.

From the studies on ETT2 and Flag-2, the need to adopt a comparative approach using the full phylogenetic diversity of strains was emphasised. In the previous study on ETT2, the authors sampled certain serotypes from pathogenic *E. coli* only (Makino *et al*., 2003). Here, the study sampled seventy nine well-validated and phylogenetically diverse strains, including both pathogenic representatives and commensal strains. With such a survey, there is less risk of misunderstanding the prevalence of a new gene cluster among the whole species. Investigations of this nature may help to shed light on the evolutionary processes of new genomic islands relative to the phylogeny of the genomic backbone. The distribution of ETT2 in the majority of commensal *E. coli* strains casts doubt upon its previous implication as a virulence factor.

The complete Flag-2 cluster contains genes encoding products sufficient for flagellar motility. However, with the exception of one strain of *E. coli* 042, most genome-sequenced *Escherichia/Shigella* strains only possess two gene scar, and all *Salmonella* strains have none of the genes at all. When the Flag-2 cluster was investigated in the whole range of phylogenetic strains, the entire gene cluster was found in another fifteen ECOR strains. The presence of Flag-2-like gene clusters in fifteen ECOR strains, *Yersinia pestis*, *Yersinia pseudotuberculosis*, and *Chromobacterium violaceum* suggests that the coexistence of two flagellar systems within the same species is more common than previously suspected. This poses a question about the possible function of the Flag-2 gene cluster in *E. coli*. However, a phenotype associated with the Flag-2 gene cluster has not yet been demonstrated in any of these strains.

CHAPTER FOUR

IDENTIFICATION OF GENOMIC DIVERSITY BETWEEN TWO *CAMPYLOBACTER JEJUNI* STRAINS

4.1 Introduction

Since *Campylobacter jejuni* was recognised as a human pathogen in the 1970s, it has become the leading bacterial cause of food-borne gastroenteritis in many industrialised countries (Sahin *et al.*, 2002). The bacteria can penetrate through the epithelial layer by first attaching to and then invading the epithelial cells and then causes diarrhoea (Nachamkin *et al.*, 1998; Hofreuter *et al.*, 2006). Infections caused by *C. jejuni* vary from mild, non-inflammatory, self-limiting diarrhoea to severe, inflammatory, bloody diarrhoea lasting for several weeks (Wassenaar and Blaser, 1999). In addition, *C. jejuni* is associated with the development of the neurological disorder Guillain-Barré syndrome (GBS) (Nachamkin *et al*., 1998). The consumption of contaminated meat, particularly poultry, accounts for the majority of human infections caused by *C. jejuni* (Wassenaar and Blaser, 1999). *C. jejuni* can be rapidly transmitted between adjacent animals due to its low infectious dose. Efforts to reduce campylobacters in the food chain have been largely concentrated at the poultry farm level (Newell *et al.*, 2001).

Because of the lack of genetic information, the virulence mechanisms of *C. jejuni* have been poorly understood for more than 20 years. The first complete genome sequence of a *Campylobacter jejuni* strain, NCTC 11168, was released to the public in 2000 (Parkhill *et al.*, 2000). *C. jejuni* 11168 has a relatively small genome of 1,641,481 base pairs in size, and a relatively low GC content of 30.6%. The genome contains 1654 predicted coding sequences (CDS), of which the average size is nearly 1 kb. *C. jejuni* is a very dense genome as 94.3% of genes encode for proteins. Recently, re-annotation of the whole genome reduced the total number of CDS from 1654 to 1643 (Gundogdu *et al.*, 2007). Re-

annotation was carried out using a new annotation database, where more currently available *Campylobacter* genomes are incorporated.

The protein coding genes of the *Campylobacter* genome were rarely organised into operons or clusters, however, some large regions with a very low GC content were discovered. Three such large gene clusters were identified to code for important surface structures in *C. jejuni* strains. These included CDSs Cj1135-Cj1148 representing genes encoding the lipooligosaccharide (LOS), CDSs Cj1412-1442 corresponding to the genes encoding the encapsular polysaccharide (EP), and CDSs Cj1305-Cj1345, which code for flagellar modification (Parkhill *et al*., 2000). Most of the hypervariable sequences in the genome are coincident with the genetic loci for LOS biosynthesis, EP biosynthesis and flagellar modification, and much research effort has been focused on them. The lower-than-average GC content of these regions of the *C. jejuni* genome suggested these genes were acquired through horizontal gene transfer. The "hot-spots" for the frequent integration of diverse genetic materials were suggested in the *C. jejuni* genomes (Hofreuter *et al*., 2006).

Recently, more complete genomes from the species have become publicly available (Table 4-1). Genome comparison between *C. jejuni* and its closely related genome *Helicobacter pylori* indicated that the three genetic loci relating to surface structures are unique in *C. jejuni* but absent in *H. pylori* (Parkhill *et al*., 2000). Comparison between *C. jejuni* NCTC 11168 and the further two complete *C. jejuni* genomes, *C. jejuni* strain RM1221 (Fouts *et al.*, 2005) and *C. jejuni* subsp. *jejuni* 81-176 (Hofreuter *et al*., 2006), showed that LOS, flagellin, and the capsule represent hypervariable regions in all three *C. jejuni* genomes.

Table 4-1 Current *C. jejuni* genomes sequenced or in progress.

Campylobacter jejuni strains	Sequence status	Length (Mbs)	Proteins	Number of RNAs	GenBank Accession number
Campylobacter jejuni 11168	complete	1.6	1634	56	AL111168
Campylobacter jejuni RM1221	complete	1.8	1838	56	CP000025
Campylobacter jejuni subsp. *doylei* 269.97	complete	1.8	1731	61	CP000768
Campylobacter jejuni subsp. *jejuni* 81-176	complete	1.6	1653	54	CP000538
Campylobacter jejuni subsp. *jejuni* 84-25	Assembly	1.67162	1748	46	AANT00000000
Campylobacter jejuni subsp. *jejuni* CF93-6	Assembly	1.6763	1757	46	AANJ00000000
Campylobacter jejuni subsp. *jejuni* CG8486	Assembly	1.6	1425	NA	AASY00000000
Campylobacter jejuni subsp. *jejuni* HB93-13	Assembly	1.69479	1710	NA	AANQ00000000

NA Currently not annotated

The variation lies in sequence divergence among the predicted gene products, and also the presence/absence of specific genes within these loci (Hofreuter *et al*., 2006).

One of the most surprising features of the *C. jejuni* genome is that it contains almost no insertion sequences or phage-associated sequences and very few repeat sequences (Parkhill *et al*., 2000). Therefore, the genome sequencing revealed very few mechanisms to generate genetic diversity. However, the diversity of *C. jejuni* strains has been demonstrated at both the genotypic and phenotypic levels (Wassenaar and Blaser, 1999). *C. jejuni* also contains a wide variety of clones within the species (Wassenaar and Blaser, 1999). In order to know more about *C. jejuni* genetic diversity, whole-genome microarray comparison was used shortly after the genome was finished (Dorrell *et al.*, 2001).

Primers for a whole-genome PCR microarray were designed to represent each CDS of the published *C. jejuni* NCTC 11168 genome from 2000. Up to 30 NCTC 11168 loci were discovered to be either absent or highly divergent among 11 *C. jejuni* strains from diverse origins (Dorrell *et al*., 2001). Many of these divergent regions were characterised as being involved in the biosynthesis of surface structures including flagella, LOS and the capsule (Dorrell *et al.*, 2001). In contrast, about 80% of *C. jejuni* NCTC 11168 genes are conserved, the so-called "core genes", mainly representing those for housekeeping functions such as metabolic, biosynthetic, cellular, and regulatory processes (Dorrell *et al*., 2001).

However, as noted before, there are several drawbacks to the use of whole genome DNA microarrays for comparative genomics: in particular, point mutations, small deletions, gene rearrangements, and novel genes will not be detected. We speculated that whole genome

PCR scanning techniques might overcome this limitation, building on the success of similar methods in the study of two *E. coli* gene clusters, ETT2 and Flag-2 (see chapter 3). In addition, a modified version of single primer PCR is also developed to identify the novel genes absent from the reference strain. This study aims to use PCR-based methods to analyse a novel genome, of the *C. jejuni* strain M1, in comparison with the reference strain, NCTC 11168. The *C. jejuni* strain M1 was first isolated from an abattoir by Diane Newell and colleagues (Veterinary Laboratories Agency, Weybridge, UK).

4.2 Materials and Methods

4.2.1 Bacterial strains

C. jejuni strain M1 was a gift from Professor Duncan Maskell (University of Cambridge). The strain was stored at -80^{o}C in 1% (w/v) proteose peptone water containing 10% (v/v) glycerol until required. The strain was routinely grown from frozen on 10% (v/v) blood agar plates, kindly provided by Professor Charles Penn (University of Birmingham). To prepare DNA, a small amount of material from the frozen culture was scraped onto blood agar plates and incubated overnight at 42^{o}C in a microaerobic atmosphere [85% (v/v) N_2, 7.5% (v/v) CO_2, 7.5% (v/v) O_2]. Genomic DNA was extracted using a DNAeasy Kit (Qiagen, UK) and stored in aliquots at -20°C.

4.2.2 PCR Primers

The PCR primers used in this study were kindly provided by Professor Brendan Wren (London School of Hygiene and Tropical Medicine). A total of 1731 pairs of primers were originally used for a whole genome DNA microarray. They were designed to represent all the initially identified CDSs from the *C. jejuni* NCTC 11168 genome project (Parkhill *et al.*, 2000). The primers were provided at a concentration of 50 µMol and stored in 96-well plates. Unfortunately, two plates containing both the forward and reverse primers corresponding to CDSs Cj0106 to Cj0202 were not available from Professor Wren's stock. The PCR scanning for these genes was therefore not included in this project. The missing primers were not resynthesised, because of a competing project from Sanger Institute (UK).

4.2.3 Tiling-Path PCR

Two adjacent genes were combined into each PCR amplicon in order to construct a complete tiling-path through the whole genome (Figure 4-1). In other words, in generating every PCR product, the left primer was from the upstream gene and the right primer was from the opposite strand of the adjacent downstream gene. That means that the tiling-path was constructed by amplifying every gene once as the downstream gene in the first PCR, then secondly as the upstream gene in the next PCR. PCR products ranged from 500 bp to 3 kb in size. All PCRs were amplified using *Taq* polymerase (Invitrogen, UK) according to the manufacturer's instructions. The PCR master reagent (containing *Taq*, 10x buffer, Mg^{2+}, dNTPs, and dH_2O) was prepared by hand, but the mixing of the master reagent with each individual primer was automatically performed using a robot (Functional Genomics Laboratory, University of Birmingham). This enabled the processing of 384 (4 x 96) PCRs a day. The PCR conditions were: 30 cycles of 30 seconds at 94°C, 30 seconds at 55°C and 1 minute at 72°C, followed by a 7 minute extension at 72°C.

Gel electrophoresis and PCR purification for the large number of PCRs in this project were facilitated by high-throughput methods. Gel electrophoresis was performed using a 96-well horizontal gel tank, and the PCR products were loaded using a multichannel pipette. The QIAgen MinElute 96 UF PCR purification kit was used for high-throughput PCR clean up.

4.2.4 Long Single-Primer PCR

Any unsuccessful PCR was followed up by a deletion-scanning PCR (DS-PCR) using the flanking primers to confirm the presence of a deletion or genome-specific island (also see

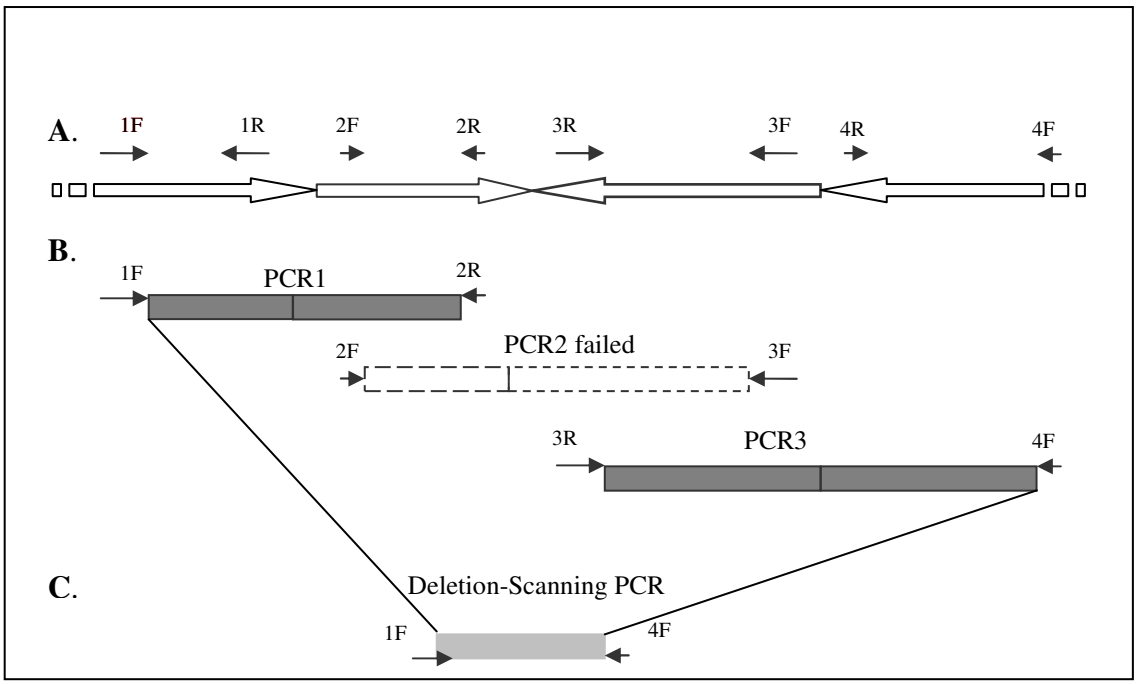

Figure 4-1 The whole-genome tiling-path PCR strategy used to analyse the *C. jejuni* M1 strain. A. Each gene is illustrated by a block arrow. Primers are labelled as F and R relative to each gene, in which the genome sequenced strand primer was called F and the other primer R. B. Two adjacent genes are combined into one PCR amplicon as illustrated. C. Deletion-scanning PCR was used to confirm the region corresponding to the failed PCR. For this purpose, one could also use the primers 1F+3F or 2F+4F in this example.

Chapter 3). Single-primer PCR approaches have been demonstrated to allow chromosomal walking from the known region into the unknown chromosome (Karlyshev *et al*., 2000). A modified Single-Primer PCR (SP-PCR) protocol was developed using long PCR techniques to give products of 5-10 kb in length (Figure 4-2). Long SP-PCR was performed using TaKaRa LA *Taq* polymerase (Cambrex Bio Science Wokingham, Ltd) according to the manufacturer's instructions using the buffer supplied. Long SP-PCR reaction conditions were: an initial 2 minutes hot start at 96ºC followed by three steps: the first 30 cycles were 20 seconds at 96ºC, 30 seconds at 60ºC and 10 minutes at 72ºC, then the following 2 cycles were: 20 seconds at 96ºC, 30 seconds at 30ºC and 10 minutes at 72ºC, and the final 30 cycles were: 20 seconds at 96ºC, 30 seconds at 60ºC and 10 minutes at 72ºC. This was followed by 10 minutes extension at 72ºC.

4.2.5 Sequencing of long SP-PCR products

For SP-PCR sequencing, the gel purified long SP-PCR product was cloned into a pCR® 2.1 vector (TA Cloning® Kit, Invitrogen, UK) and grown on X-gal LB plates (Figure 4-2). To select the correct clone for sequencing, firstly, three white colonies from each cloning were selected for plasmid DNA prep (QIAprep spin miniprep kit, Qiagen, UK). Secondly, PCR was performed on each plasmid DNA using two primers. The forward one is the primer used for SP-PCR, while the reverse primers were designed using the opposite strand to anneal around a few hundred base pairs downstream of the SP-PCR primer binding site (Table 4-2). Once PCR confirmed an insertion in that region, sequencing was subsequently carried out from both directions using the vector primers (M13 forward and reverse primers). New primers can be designed from the derived sequences successively until the gap is closed.

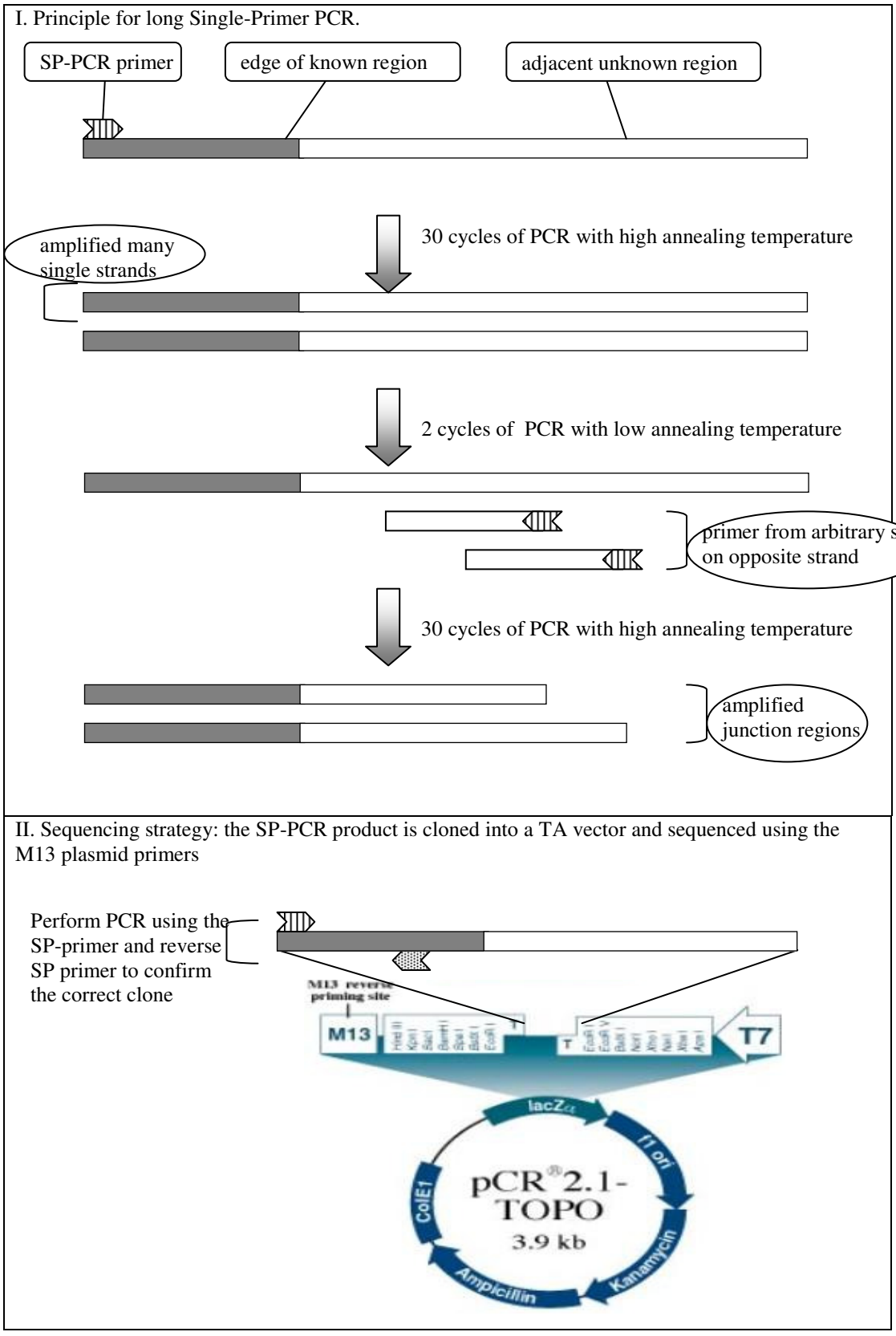

Figure 4-2 Long single-primer PCR and the sequencing strategy used to analyse the large genomic discrepancies between two *Campylobacter* genomes.

Table 4-2 Primers used for long single-primer PCR in the *C. jejuni* study.

Regions of differences (RD)	SP-PCR originating from	SP-primer	Reverse SP-primer
RD-1	Cj0028	ATCGGGTTTAAATTGGCATGAA	TTCCAGCAGCACCTTTATGG
	Cj0034c	GCCATAGTTTTGGCAAGTTTGA	TCCGCTTCGACTTCTTTTTGA
RD-2	Cj0054c	TTGTGAAAAAGCACCTTCGTTG	Not required*
	Cj0059c	GCGATAATGCTTAGCGAAGAGC	TTTGAACGCCGAAGTTACCA
RD-3	Cj0289c	TTGGAACAAAGGCCACGATATT	ATTTGGCGCTTCAGATCAAT
	Cj0300c	ATGCGTGCCAATTTACAAGATG	Not required
RD-4	Cj0304c	TGCGTTTTAAATCAAGCCATTG	Not required
	Cj0309c	GTGAAATTTTTGGCGTGGTGAT	ATTACCCCACCAGCGGTTC
RD-5	Cj0479	TGGTAATTGCAGGAAAAGGTGA	CAATCACAGCCGCATTTTT
	Cj0487	GGAGTTTCTTGCGAAAAACCAT	GGAAGTAATGCACACGAGCAA
RD-6	Cj0564	TTGCGAATGCATGGAGTGTTAT	Not required
	Cj-0571	AAAGCCTTGATCGAGTTTTGGA	Not required
RD-7	Cj0615	GTGCTTTAGTGGGTGGGCTTAC	TCGCCTATAGCCCCTGCTAA
	Cj0619	AGCAACAATGCGAGCAAAAATA	ATGGCATGGGGAGTATTTGC
RD-8	Cj0742	TGAAAATGATGGTGATTTTTCTGG	TATCCGCACTTCCCTCACCT
	Cj0757	CGAAAAAGAAGCGTGGAATTTT	AAAAACTGCGCATCCACCTT
RD-9	Cj0775c	GGCAAAATCACTTCTTTGTCCA	CATGATGGCGCTTTAAGTGA
	Cj0778	TATCATCGCCTAAAGCCTTTGC	TGGTGCAGGACAAGTTGGAC
RD-10	Cj1024c	GCTTCTTGGCTGAGCTGTTTTT	AAATGGCGAATTTCGTTCAG
	Cj1030c	TGGCCTTAGGCTTTGGTTTTAG	ACAGGAGGAAGCTCGCTAGG
RD-11	Cj1049c	TATCTTGCCACCACAAAGGCTA	TTTTTGGCGATTTTTGGTTT
	Cj-1055c	GCCCCTTTAAGCATGTGGCTAT	CAGCCAAATTAAGCCCAAAA
RD-12	Cj1134	GCAAAATCAAAACACCACCAAA	TTTGTCTTATGCCGCCTTTT
	Cj1146c	TGATTTGCTTTGCTTTGATGCT	TCTTCGCCATAACTCAAACG
RD-13	Cj1295	ATGCAAAACACGATGAAAATGG	GCACAAAGCCTGCTTTAACA
	Cj1312	ATATCAGTTCCACCCATGCAAA	TGCCTATGCAAAAGCAAGTG
RD-14	Cj1316c	TTTTGGATCACGCTTGATATGG	GGTGTTTGTGATGCTTGTCG
	Cj1342c	AATTCTTCATCGCCTCGTTTTT	TTTTACCGGAATTCCCATGC
RD-15	Cj1548c	ACTGCTACGCTTGAGCCTTCTT	GCAGGGGTTGGTTGTATGGT
	Cj1562	AAAGGTTTGTTTCCTTTTGTTCA	CATCGACTTCATTTGCAAGC
RD-16	Cj1676	GCAAAATTCCTGGTACACTTGG	TCCTGCAAAGTCGTTTTTCG
	Cj1680c	TTTTAGCCCTAAGGATGCAAAA	AACATTAGCCACACCCTCTTT
RD-17	Cj1721c	AACCCAAATATCCACCTGCAAC	TTGGGTTTAAAAGCGGGTTA
	Cj0001	AAGCTTGCATTTCCAACTGCTT	ATGAAGTGCAAAGCGGAAAT

*Not required: the sequencing data was obtained before the reverse SP-primers were designed.

4.3 Results

4.3.1 Co-linear regions between *C. jejuni* M1 and *C. jejuni* NCTC 11168

In total, 1559 PCRs were carried out on genomic DNA of *C. jejuni* M1 in this project. The high throughput techniques of robotic PCR, gel electrophoresis, and purification were demonstrated to have been used to improve productivity for a project of this size. The first round PCR produced over 80% positive PCR products, using genomic DNA from *C. jejuni* M1, with the expected sizes calculated from *C. jejuni* NCTC 11168.

When a single PCR or a few continuous PCRs failed during the initial scanning, further PCR studies were undertaken using various combinations of primers flanking that region. One example is shown in Figure 4-3. The deletion-scanning PCRs were shown to be a useful tool in discovering any deletions or insertions. Using this method, nearly 40% (113 genes) out of the initial negative PCRs (304 genes) were confirmed. Deletion-scanning PCR showed 10 regions with obvious insertions, with the largest insertion being up to 10 kb, and 5 regions with deletions ranging from 1 to 5 kb (Figure 4-4).

Concurrent to this work, ten large islands in the *C. jejuni* M1 genome were discovered by the Sanger Institute, by the construction and sequencing of BAC (bacterial artificial chromosome) clones (Parkhill, personal communication). Interestingly, seven of the ten regions corresponded with the regions identified by the author's genome-tiling PCR. In addition, several insertions and deletions in regions not covered by the BAC clones were identified by deletion-scanning PCR.

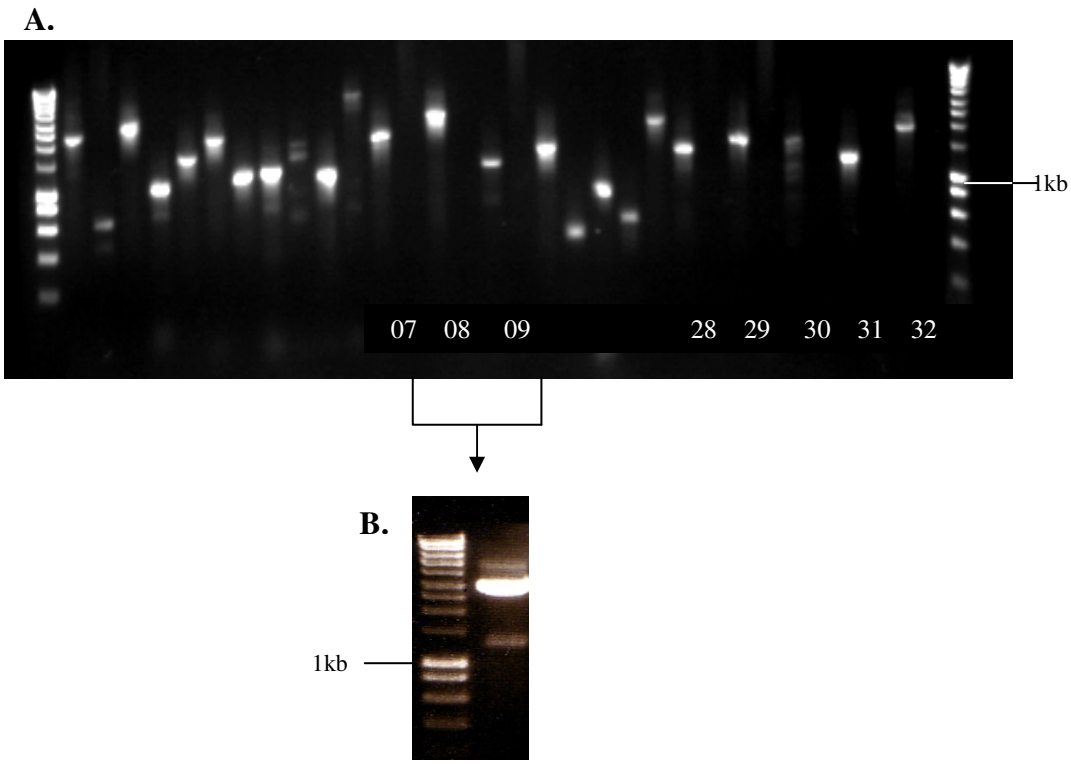

Figure 4-3 A. Gel images illustrating the Tiling-Path PCR analysis of the *C. jejuni* M1 strain. The PCR amplicons represent the CDSs from Cj0001 to Cj0032. PCR products were loaded on a 1.0% agarose gel with Hyper Ladder I MW markers with the l kb bands marked (Bioline, United Kingdom). Lanes corresponding to the negative amplicons are labelled according to CDS numbers (in abbreviation). B. Deletion-scanning PCR confirmed the initial negative results. Using the Cj0007 forward primer and Cj0009 reverse primer confirmed a ~4 kb PCR product.

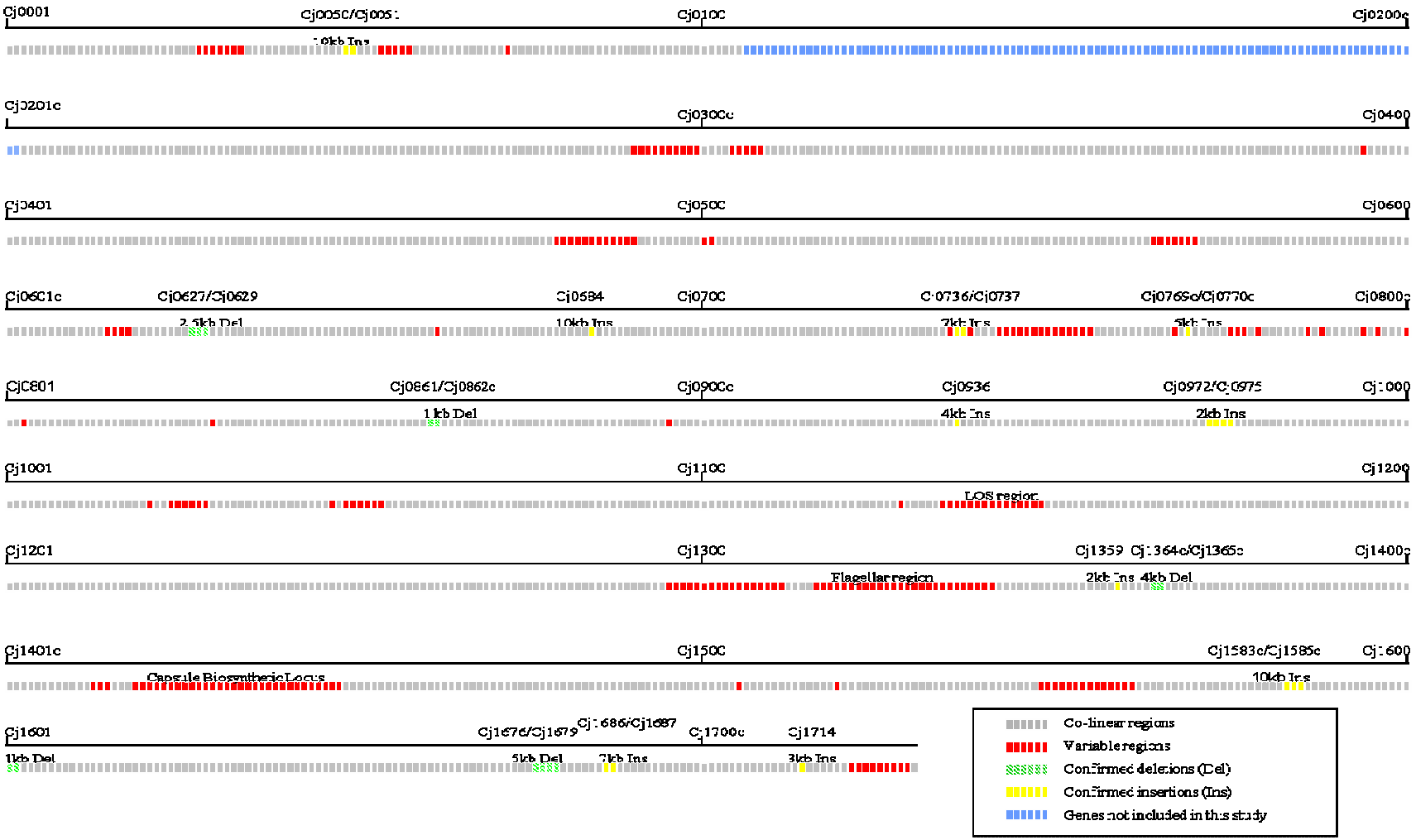

Figure 4-4 Whole-genome tiling-PCR analysis of the *C. jejuni* M1 strain. Each rectangle represents the equivalent gene in the M1 genome to the coding sequence in *C. jejuni* NCTC 11168 genome. The upper lines depict the positions in the genome.

At this stage, twenty two single genes failed to be amplified from the *C. jejuni* M1 strain. In addition, seventeen large discrepancy regions were identified in the M1 genome when compared to the genome of *C. jejuni* NCTC 11168 (Figure 4-4). The regions which were successfully amplified from the genomic DNA of *C. jejuni* M1 were inferred to be genetically co-linear to *C. jejuni* NCTC 11168. By the final stage, 1368 genes from *C. jejuni* M1 were confirmed to be co-linear to *C. jejuni* NCTC 11168, which accounts for 83.2% of the *C. jejuni* NCTC 11168 genome.

4.3.2 Identification of genomic discrepancy between two genomes

The seventeen regions that showed obvious discrepancy between the two genomes were named regions of differences (RD) in this study. In some of these regions, up to 30 continuous primer pairs did not amplify. The previously identified variable regions, including genes encoding the capsular biosynthesis locus, flagellar region and lipooligosaccharide locus (LOS), were also within the RDs of the M1 strain. These regions correspond to the large regions of low G+C content (Parkhill *et al*., 2000). In order to further understand these variable regions, long single-primer PCR was used to survey the unknown regions from the known chromosome. 34 primers for long single-primer PCR were designed from both extremes of every region. All long single-primer PCRs worked and products up to 10 kb in size were obtained (data not shown).

For effective sequencing, PCR products were cloned into a TA vector with X-gal selection. Three white colonies from each cloning procedure were chosen and plasmid DNA was extracted for sequencing. Using this approach, it is possible to obtain sequence data from both strands using the vector primers. Five plasmid insert DNAs, which represent

chromosomal walking fragments originating from Cj0054c, Cj0300c, Cj0304c, Cj0564 and Cj0571, provided sequence data. However, we were unable to obtain sequence data from the rest of the plasmids. Unfortunately, this project could not be finished due to funding constraints and because of a competing project in the Sanger Institute. However, the work presented here demonstrates a series of easy, quick PCR approaches to identify large genomic discrepancies between an uncharacterised genome and a sequenced reference. From the data generated in this project, it is possible to construct a representation of the *C. jejuni* M1 genome consisting of 34 contigs (Table 4-3).

Table 4-3 Assembled *C. jejuni* M1 genomic contigs relative to *C. jejuni* NCTC 11168.

Contig	Successful scanning PCRs, suggesting colinear regions	Comments*
1.	Cj0001-Cj0028	**Cj0028-Cj0034c PCR failed.** SP-PCR from Cj0028 to Cj0034c worked. No direct sequence available.
2.	Cj0034c-Cj0054c Cj0059c-Cj0072c	**Cj0054c-Cj0058 PCR failed.** SP-PCR and sequence suggest Cj0055c, 0056c, 0057 absent **Cj0072c PCR failed**. Pseudogene.
3.	Cj0073c-Cj0290c Cj0300c-Cj0304c Cj0309c-Cj0394c	**Cj0290c-Cj0299 PCR failed.** SP-PCR and sequence from Cj0300c suggest Cj0296c, 0297c, 0298c, 0299 absent. No direct sequence from Cj0290c. Known molybdenum transport region in 81-176. **Cj0304c-Cj0308c PCR failed.** SP-PCR and sequence from Cj0304c suggest Cj0305c, 0306c and 0307 absent. **Cj0394c PCR failed.**
4.	Cj0395c-Cj0479 Cj0491-Cj0500 Cj0501-Cj0564	**Cj0479-Cj0490 PCR failed.** SP-PCR and sequence showed Cj0479 close to Cj0491, suggesting Cj0480-Cj0490 are missing. This is supported by *H. pylori* comparison. **Cj0500-Cj0501 PCR failed.** Deletion-scanning PCR showed >10kb insertion = BAC Region 1. **Cj0564-Cj0570 PCR failed**. SP-PCR from Cj0564 and Cj0570 worked, but no direct sequence obtained. Known region of difference in RM1221 and 81-176.
5.	Cj0570-Cj0615 Cj0619-Cj0627 Cj0630c-Cj0662c	**Cj0615-Cj0618 PCR failed.** SP-PCR and sequence from both directions suggest the deletion of Cj0617, 0618. **Cj0627-Cj0629 PCR failed.** Deletion-scanning PCR showed ~2.5kb deletion. **Cj0662c PCR failed.**
6.	Cj0663c-Cj0684 Cj0685c-Cj0736 Cj0740-Cj0742	**Cj0684 PCR failed**. Deletion-scanning PCR showed ~10kb insertion within Cj0684. **Cj0736-Cj0739 PCR failed**. Deletion-scanning PCR showed 7kb insertion = BAC Region 2. **Cj0742-Cj0755 PCR failed**. SP-PCR no sequence obtained.
7.	Cj0757-Cj0767c	**Cj0767c PCR failed.**
8.	Cj0768c-Cj0769c Cj0770c-Cj0775c	**Cj0769c-Cj0770c PCR failed**. Deletion-scanning PCR showed ~5kb insertion = BAC Region 9. **Cj0775c-Cj0777 PCR failed.** SP-PCR and sequence showed no difference in this region but gap not closed.
9.	Cj0778-Cj0779	**Cj0779 PCR failed.**
10.	Cj0780-Cj0786	**Cj0786 PCR failed.**
11.	Cj0787-Cj0788	**Cj0788 PCR failed.**
12.	Cj0788-Cj0794	**Cj0794 PCR failed.**
13.	Cj0795cCj0796c	**Cj0796c PCR failed.**
14.	Cj0797c-Cj0800c	**Cj0800c PCR failed.**
15.	Cj0800c-Cj0803	**Cj0803 PCR failed.**
16.	Cj0803-Cj0830	**Cj0830 PCR failed.**

17.	Cj0831c-Cj0862c Cj0863c-Cj0895c	**Cj0862c PCR failed**. Deletion-scanning PCR showed ~1.5kb small deletion. **Cj0895c PCR failed.**
18.	Cj0896c-Cj0936 Cj0937-Cj0972 Cj0972-Cj1021c	**Cj0936 PCR failed.** Deletion-scanning PCR showed ~4kb insertion within Cj0936 = BAC Region 8. Confirmed by 81-176 comparison. **Cj0972-Cj0975 PCR failed.** Deletion-scanning PCR showed ~2kb insertion = BAC Region 3. **Cj1021c PCR failed.**
19.	Cj1022c-Cj1024c	**Cj1024c-Cj1029c PCR failed.** SP-PCR and sequence showed Cj1024c close to Cj1028c and Cj1029c; and Cj1030c close to Cj1025c and Cj1026c. Possible rearrangement, but gap not closed.
20.	Cj1030c-Cj1047c	**Cj1047c PCR failed.**
21.	Cj1048c-Cj1049c	**Cj1049c-Cj1054c PCR failed.** SP-PCR and TA clone worked, no direct sequence obtained.
22.	Cj1055c-Cj1128c	**Cj1128c PCR failed.**
23.	Cj1129c-Cj1134	**Cj1134-Cj1148 PCR failed.** Known LOS region.
24.	Cj1149c-Cj1295	**Cj1295-Cj1311 PCR failed.** SP-PCR and sequence from Cj1312 generated ~400 bp novel sequence. Gap not closed.
25.	Cj1312-Cj1316c	**Cj1316c-Cj1341c PCR failed.** Known Flagellar region.
26.	Cj1342c-Cj1359 Cj1360c-Cj1364c Cj1366c-Cj1413c	**Cj1359 PCR failed.** Deletion-scanning PCR showed ~2kb insertion within Cj1359 = BAC Region 4. **Cj1364c-Cj1365c PCR failed.** Deletion-scanning PCR showed ~4kb deletion. **Cj1413c-Cj1415c PCR failed.** Known Capsule Biosynthetic Locus.
27.	Cj1416c-Cj1419c	**Cj1419c-Cj1448c PCR failed.** Known Capsule Biosynthetic Locus.
28.	Cj1449c-Cj1505c	**Cj1505c PCR failed.**
29.	Cj1506c-Cj1519	**Cj1519 PCR failed.**
30.	Cj1519-Cj1548c	**Cj1548c-1561 PCR failed**. SP-PCR and sequence from Cj1548c found novel sequence.
31.	Cj1562-Cj1583c Cj1586-Cj1601 Cj1601-Cj1676 Cj1680c-Cj1686c	**Cj1583c-Cj1585c PCR failed.** Deletion-scanning PCR showed >10kb insertion = BAC Region 10. Confirmed by 81-176 comparison. **Cj1601-Cj1602 PCR failed.** Deletion-scanning PCR showed ~1.5kb deletion. **Cj1676-Cj1679 PCR failed.** Deletion-scanning PCR showed ~5kb deletion. This is confirmed by RM1221 comparison. **Cj1686c-Cj1687 PCR failed.** Deletion-scanning PCR showed ~7kb insertion. Confirmed by 81-176 comparison.
32.	Cj1688c-Cj1714	**Cj1714 PCR failed.** Deletion-scanning PCR showed ~3kb insertion.
33.	Cj1714-Cj1721c	**Cj1721c-Cj1729c PCR failed.** SP-PCR, sequence not available.
34.	Cj1731c-Cj0001	

*All failed PCRs or clusters used for assembling M1contigs are marked in bold. Regions which were identified in the Sanger BAC sequencing project are indicated in red. Regions corresponding to known variable regions in other strains of the species are indicated in blue.

4.4 Discussion

Campylobacter jejuni is an important food-borne human pathogen. However, understanding of the genetics, physiology and virulence of this organism remains poor. The existing control and prevention in the food chain are focused at the level of the poultry farm (Newell *et al*., 2001). The virulence mechanisms of this organism and the sources and transmission routes of human campylobacteriosis are not fully understood. All of these have greatly hindered the design of disease prevention strategies. *C. jejuni* represents a variety of genotypically diverse species (Wassenaar and Blaser, 1999). Therefore, it is necessary to distinguish between *C. jejuni* strains from different sources, for a further understanding of its sources of infection and transmission routes.

The availability of the *C. jejuni* NCTC 11168 complete genome sequence has not only shed light on the genetic features of this species, but also provided the basis for the application of post-genomic techniques. The genome sequence identified several large gene clusters with an unusually low GC content, suggesting they had been acquired by horizontal gene transfer. However, the sequence of the genome revealed few mechanisms by which genetic diversity can be generated. Because of this, there is interest in investigating the mechanisms of genotypic diversity by characterising more *C. jejuni* strains.

The approach described here can be used for whole genome comparison amongst closely related bacteria, where at least one has been completely sequenced. This approach is based on existing PCR techniques, together with the newly modified long single-primer PCR. This allowed the easy identification of co-linear regions, as well as the detection and

characterisation of any large difference. The two *Campylobacter* strains, *C. jejuni* M1 and NCTC 11168, were found to be closely related, sharing at least 83.5% of the *C. jejuni* NCTC 11168 genes.

Use of a single-primer PCR chromosome walking approach has been reported previously (Karlyshev *et al*., 2000). The published approach also allowed direct sequencing of the PCR products with a nested primer. However, this approach was restricted to fragments of a few hundred bp and it was unclear whether it would scale-up to long-PCR conditions (Karlyshev *et al*., 2000). In this project, a modified version of the SP-PCR protocol, using the conditions and enzymes used for long-PCR, was devised. The results demonstrated that it was possible to obtain single-primer PCR products up to 10 kilobases in length. This greatly extends the utility of the approach for chromosome walking. Furthermore, these PCR amplicons could be cloned into selected plasmids and sequenced with the vector primers, which allowed sequencing from both orientations. However, good-quality sequence is still restricted to about 1 kilobase. In other words, sequence read-length was limited by the sequencing reaction rather than the SP-PCR.

From the data obtained in this study, the genome of *C. jejuni* M1 has been assembled into 34 contigs. The next target is to close the gaps, which would reveal all the genetic differences between the two genomes. Though full gap closure was not possible within the limitations of this project, the sequence data and PCR results obtained have provided some useful information for further research. For example, the genes Cj0054c-Cj0058 were not successfully amplified by genome-tiling PCR. When sequence was obtained originating from both Cj0054c and Cj0058, it showed a gap of about 150 bp between Cj0054c and

Cj0058, suggesting the deletion of genes Cj0055c, Cj0056c, and Cj0057. In another case, sequence was obtained from both ends of the region between Cj1024c and Cj1030c. The sequence originating from Cj1024c corresponded to the genes Cj1028c and Cj1029c, while the sequence originating from Cj1030c corresponded to Cj1025c and Cj1026c. This suggested the possibility of a rearrangement within this region. However, further studies are required to confirm these rearrangements in the *C. jejuni* M1 genome.

In summary, the advantage of using genome-tiling PCR is that all the insertions/deletions are indicated by their size and position and these results have provided at least preliminary information on the genome organisation of a novel strain, *C. jejuni* M1.

CHAPTER FIVE

GENOME SEQUENCING OF THE EUROPEAN *FRANCISELLA TULARENSIS* SUBSPECIES *TULARENSIS* ISOLATE FSC198

5.1 Introduction

5.1.1 The bacterium *Francisella tularensis* and the disease tularaemia

Francisella tularensis is a small gram-negative bacterium and a facultative intracellular pathogen. It causes a zoonotic and infectious disease, tularaemia, among humans and animals. It was isolated in 1911 from rodents suffering from a plague-like disease in Tulare County, CA (USA) (Gliatto *et al.*, 1994).The disease has not been demonstrated to be directly transmissible between humans, whereas rodents, hares and rabbits are important sources of human infections (Ellis *et al.*, 2002). The true reservoir of the bacterium in the environment is as yet unknown.

The most common infectious route is a bite from anthropod vectors, such as ticks, flies, and mosquitoes, which have been previously infected. The infection can be acquired through the skin or mucous membrane, but the most severe form of the disease is caused by inhalation of an aerosol containing bacteria. Indeed, inhalation of as few as 10 colony-forming units (CFU) is sufficient to cause disease in humans (Riley *et al.*, 1995). The respiratory form is the most dangerous tularaemia and there is up to 30% mortality in the absence of antibiotic therapy (Dennis *et al.*, 2001). Outbreaks are rare but can involve a large number of cases. For example, almost 700 cases of respiratory tularaemia were reported in one outbreak in Sweden in 1966-1967 (Dahlstrand *et al.*, 1971).

Given its infectious nature, ease of dissemination and the capacity to cause severe illness and death, this organism has been considered as a potential biological weapon since the 1950s. It is known to have been developed as a bioweapon by Japan, the United States and

the USSR (Christopher *et al.*, 2005). Moreover, in the post-Cold War era, it is included among the top six agents showing a potential threat to public health if used as a bioterrorism weapon (Harris, 1992). Unfortunately, there is no licensed vaccine available so far (Titball and Oyston, 2003). Though the live vaccine strain (LVS) has been demonstrated to provide protective immunity against tularaemia, the mechanism of attenuation and protection is not known (Tarnvik, 1989; Sandstrom, 1994). This has led to the withdrawal of the UK and USA licensing for vaccine use of this strain. The need to develop a safe and licensable vaccine to protect against tularaemia, particularly the respiratory form of the disease, has led to increased interest in this organism in recent years.

Francisella is classified as a member of the γ-subgroup of the proteobacteria (Forsman *et al.*, 1990). The genus *Francisella* contains two well-recognised species, *F. tularensis* and *F. philomiragia*, in addition to a number of tick endosymbionts. *F. philomiragia* is an opportunistic pathogen, and virulent only in immunosuppressed or near-drowned individuals (Hollis *et al.*, 1989). *F. tularensis* can be divided into four recognised subspecies: *tularensis* (formerly known as *F. tularensis* type A)*, holarctica* (formerly *F. tularensis* type B)*, mediasiatica* and *novicida.* The classification sometimes includes an additional subspecies as *holarctica* variants found in Japan (Svensson *et al.*, 2005) are intermediate to *F. tularensis* subsp*. tularensis* and the other *F. tularensis* subsp*. holarctica* isolates. These subspecies are distinct from each other in both virulence and geographical distribution. Two subspecies, *tularensis* and *holarctica*, cause disease in humans: the former causes diseases with high mortality whereas the latter is highly infectious but shows moderate virulence. The LVS is an attenuated variant of *F. tularensis* subsp*. holarctica*. *F.*

tularensis subsp. *novicida* causes diseases in mice similar to those caused by subsp. *tularensis* in human (Anthony *et al.*, 1994).

F. tularensis is widely distributed in the Northern hemisphere, and the different subspecies show restricted geographical ranges. *F. tularensis* subsp. *tularensis* strains are predominantly found in North America, while *F. tularensis* subsp. *holarctica* strains are found in Europe and Asia, and to a lesser extent in North America. *F. tularensis* subsp. *mediasiatica* is primarily found in central Asia and the former USSR, while *F. tularensis* subsp. *novicida* was thought to be restricted to North America, but there has been a report of isolates found in Australia (Whipp *et al.*, 2003).

5.1.2 Complete genome sequences of *Francisella* strains

The molecular basis of *Francisella* infection has been poorly understood for many years due to a lack of genetic tools. Prior to this project, only two complete genome sequences from *Francisella* isolates were available. The first was from a highly virulent *F. tularensis* subsp. *tularensis* strain Schu S4 (*Francisella* strain collection number FSC237). This strain was originally obtained from an ulcer from a case of tularaemia in Ohio in 1941. Since then, it has been adopted widely for use in laboratory studies. The second genome sequence was obtained from the LVS. This strain was originally obtained through multiple passage of a virulent *F. tularensis* subsp. *holartica* isolate. The LVS genome was completed at Lawrence Livermore National Laboratory in 2006 (https://maple.lsd.ornl.gov/microbial/ftul_lvs/).

The genome of *F. tularensis* strain Schu S4 is small (about 1.9 Mb), with an overall G+C content of 32.9%. There are 1,804 predicted CDSs (including pseudogenes), of which 302 are unique to *F. tularensis* (Larsson *et al.*, 2005). The genome is characterised by a high proportion of inactivated genes, which have been degraded by insertions, deletions, and substitution mutations. More than 10% of the CDSs in the Schu S4 genome are pseudogenes or gene fragments, among which 14% are due to disruption by IS elements. In addition, a large number of IS elements scattered throughout the genome is another feature of the *F. tularensis* genome. The presence of a large number of IS elements and pseudogenes also disrupts more than half of the predicted metabolic pathways.

The *F. tularensis* Schu S4 genome sequence revealed two identical regions of 33.9 kb in size. The origin of this duplicated region is not clear since no known homologues were identified. The 25 genes encoded within the duplicated region represent a possible pathogenicity island, showing the potential to contribute to virulence (Larsson *et al.*, 2005). Homologues of some of the genes within the pathogenicity island have been suggested to encode a new secretion system in a recent study (Pukatzki *et al.*, 2006).

5.1.3 The European subspecies tularensis strains

Although the usual assumption is that *Francisella tularensis* subspecies *tularensis* is confined to North America, several isolates from Europe were reported in the 1980s to be as highly pathogenic as subspecies *tularensis* strains. These strains were first recorded during a survey of 155 *Francisella* strains isolated over the years 1978-1996 from small mammals, fleas, ticks and mites in Slovakia (Gurycova, 1998). Seventeen strains isolated from mites and fleas were found to have the ability to ferment glycerol and glucose, were

positive for citrulline uridase, and sensitive to erythromycin (Gurycova, 1998). These biochemical properties are typical of subspecies *tularensis* but not of subspecies *holarctica*. Two of the strains isolated from mites, originally designated as SE219 and SE221, exhibited a high pathogenicity for domestic rabbits, which is considered the most important property of *F. tularensis* subsp. *tularensis,* in contrast to other subspecies of *F. tularensis*. Therefore, these 17 strains (15 strains from fleas and 2 strains from mites) were classified as *F. tularensis* subsp. *tularensis.*

Interestingly, the subspecies *tularensis* strains were isolated initially from four species of fleas and two years later from two species of mites within the region of the Danube river basin near Bratislava (Gurycova, 1998). Over the following two years, isolates of *F. tularensis* subsp. *tularensis* were recovered repeatedly from fleas and mites captured in this region. The reason for the presence of these strains in the region is unknown. The isolates SE219 and SE221 were deposited in the *Francisella* strain collection (FOI, Umea, Sweden) and are known as FSC198 and FSC199 respectively. A further isolate, strain Sev-23, was obtained during a later survey from *Ixodes* spp. ticks in South East Austria in 1990, and again identified as *F. tularensis* subsp. *tularensis* (D. Gurycova, unpublished). These findings of subspecies *tularensis* strains in Europe could have serious epizootic implications.

Though the four subspecies of *F. tularensis* are distinct in virulence, biotypes and geographical distribution, they are highly similar in gene content and have greater than 95% DNA sequence identity (Nano *et al.*, 2004). Recently, a study using variable-number tandem repeats (VNTRs) at multiple loci was performed to assess the genetic relationships

among 192 *F. tularensis* isolates, representing all the recognised subspecies (Johansson *et al.*, 2004). VNTRs, also known as short sequence repeats (SSR) or micro-satellites, have been used as high-speed molecular clocks for monitoring microbial genome evolution (van Belkum, 1999a). VNTR loci normally have multiple copies of a small repeat unit of about 10-30 bp. The number of repeats present at a given genomic site is highly variable between strains, because these unstable repeats lead to slippage during DNA replication. Some experiments have shown the promise to discriminate individual strains from those species with little genomic variation, e.g., *Mycobacterium tuberculosis* and *Bacillus anthracis* (Keim *et al.*, 2000; Mazars *et al.*, 2001).

Johansson *et al*. applied twenty-five informative VNTR markers throughout the whole physical map of *F. tularensis* Schu S4. The 192 *F. tularensis* isolates were identified as 120 different genotypes (Figure 5-1). VNTR analysis revealed a great difference in genetic diversity between the subspecies *tularensis* and *holarctica* (Johansson *et al.*, 2004). *F. tularensis* subsp. *tularensis* is highly diverse, and is enough to be divided into two genetically distinct groups, arbitrarily designated: clade A.I and A.II. Clade A.II was significantly less diverse than clade A.I. Clade A.I was found predominantly in the mid-West of America and included 29 North American isolates (Johansson *et al.*, 2004). Strikingly, two Slovakian strains are grouped into the clade A.I sub-population, and are most closely related to the laboratory strain Schu S4 (Figure 5-1).

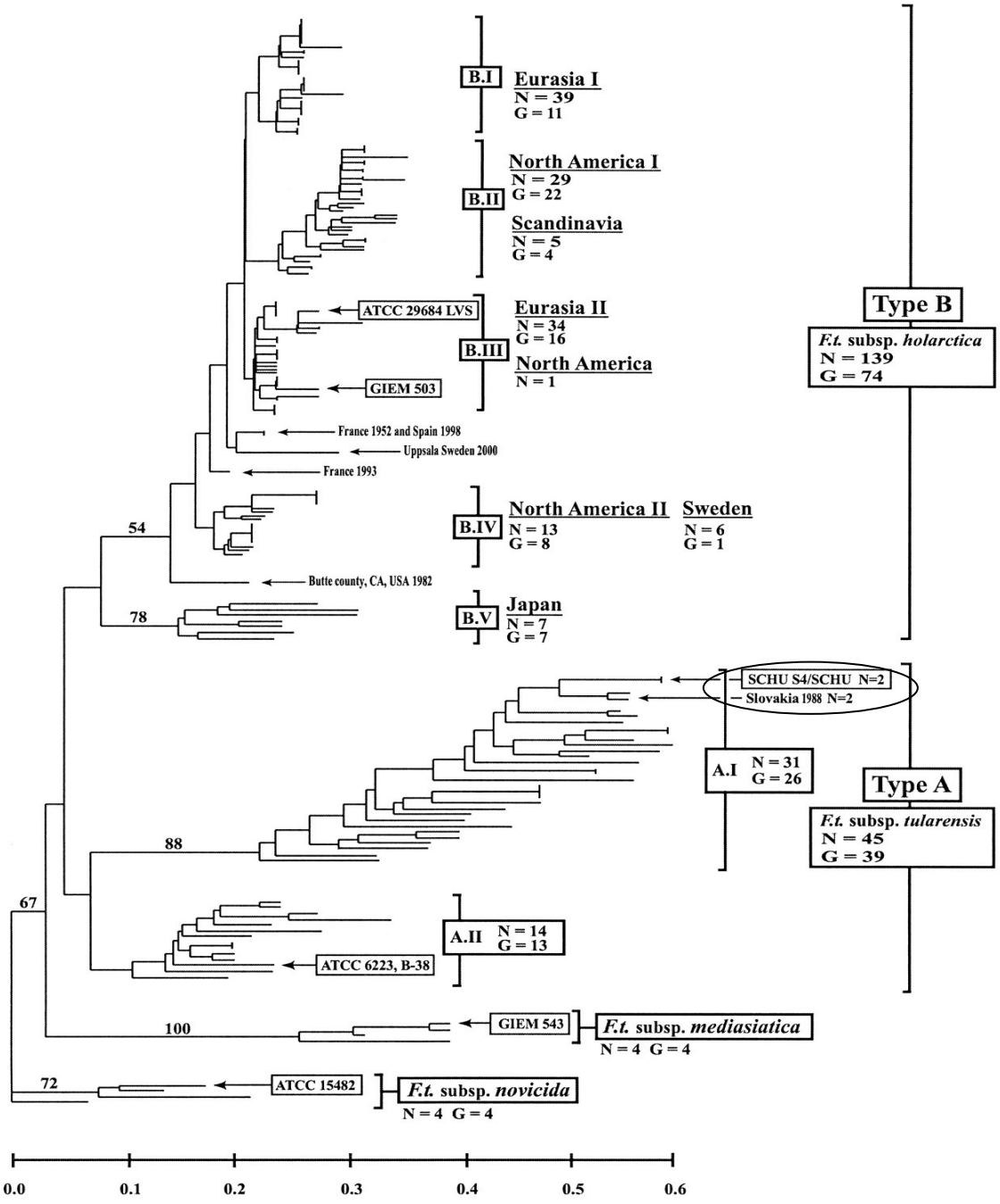

Figure 5-1 Genetic relationships among global *F. tularensis* isolates based on allelic differences at 25 variable-number tandem repeat (VNTR) markers (reproduced from Johansson *et al*., 2004). The tree was constructed using the neighbor-joining algorithm and rooted using *F. tularensis* subsp. *novicida* isolates. Two Slovakian strains were grouped into the Clade A.I sub-population, belonging to subsp. *tularensis*. The close relationship between Schu S4 and the two Slovakian strains is highlighted by oval.

5.1.4 Aims

The complete genome sequence of subspecies *tularensis* strain Schu S4 shed light on the genetic content of *F. tularensis* (Larsson *et al.*, 2005). A second genome sequence from subsp. *holarctica* strain LVS is also available (NCBI accession number NC_007880; GenBank accession number AM233362). The collection of more complete *Francisella* genomes will help in the understanding of this organism, and will aid in the determination of the function of unknown genes. Because of the lack of any licensed vaccine for this etiological agent, *F. tularensis* is still a threat to public health, and therefore, there is an increasing need to discover vaccine candidates. The genome sequencing and comparative genomic studies can benefit the potential development of a defined vaccine strain.

Since the 1980s, several strains belonging to subspecies *tularensis* have been isolated from Europe. However, little is known about their origin and they remain an enigma (Gurycova, 1998). The VNTR study revealed that these European subsp. *tularensis* strains were most closely related to the Schu S4 laboratory strain (Johansson *et al.*, 2004). However, any explanation for their close relationship remained speculative because no other genetic information on the European subsp. *tularensis* strains was available. Therefore, the aim of this project was to examine their close relationship by obtaining the complete sequence of one of the genomes, and one of the Slovakian isolates, FSC198, was chosen as the sequencing target.

5.2 Materials and Methods

5.2.1 Bacterial strains

The genomic DNA of *F. tularensis* strain FSC198 was kindly provided by the Health Protection Agency (HPA; Porton Down, UK). The genomic DNA from *F. tularensis* subsp. *tularensis* Schu S4 strain and subsp. *novicida* U112 strain were kindly provided by the Defence Science and Technology Laboratory (Porton Down, UK).

5.2.2 Shotgun sequencing and genome assembly

Shotgun sequencing data was prepared and obtained from the HPA. The shotgun sequencing library was obtained by randomly shearing the FSC198 DNA into 1-2 kb fragments, which were then ligated into the pLEXX AK double-insert vector (Chaudhuri *et al.*, 2007). The shotgun library resulted in sequencing reads at a level of 10x theoretical coverage. Initial assembly of the sequence gave 68 contigs but included numerous misassemblies due to IS elements (R. Chaudhuri, personal communication). At this stage, PCR and sequencing strategies were then applied to complete the final genome assembly and the genome finishing (see Figure 5-2).

5.2.3 Whole genome PCR scanning

As the previous study showed, FSC198 is a close relative of Schu S4 (Johansson *et al.*, 2004). To investigate any difference in genome organisation between the two strains, whole genome PCR scanning was applied to DNA from FSC198 using primers designed for Schu S4. One hundred and nineteen primer pairs were automatically designed using the

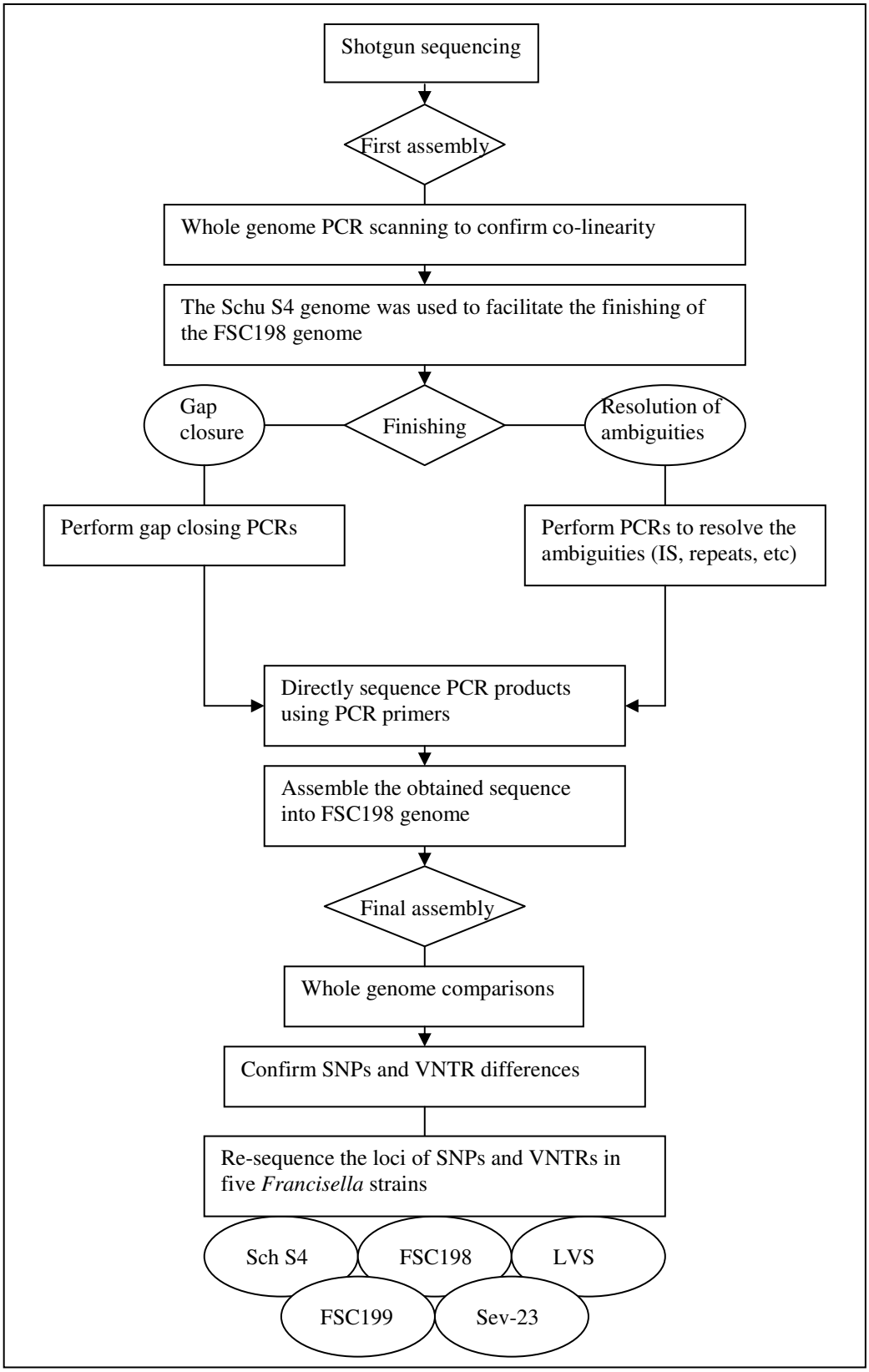

Figure 5-2 The strategy used to obtain the complete sequence of the *F. tularensis* subspecies *tularensis* FSC198 genome.

program GenoFrag (Ben Zakour *et al.*, 2004) (primers are listed in Appendix I). The primers were designed to produce ~17 kb fragments with overlaps of a few hundred base pairs.

Long PCRs were performed using TaKaRa LA Taq (Cambrex Bio Science, Wokingham, UK) as previously described (Ren *et al.*, 2004). Regions were assumed to be colinear to the equivalent region of Schu S4 if long PCR amplification was successful. Any large differences, such as deletions or insertions, could be revealed by long PCR scanning.

5.2.4 Gap closure

The genome finishing stage consisted of gap closure and resolution of ambiguous sequence. Gap closing PCRs were performed using primers designed based on the Schu S4 genome, and followed by direct sequencing with both PCR primers. Twenty five pairs of PCR primers were designed to yield 1-1.5kb products, spanning each gap (Table 5-1). The sequencing results obtained were assembled into the FSC198 genome.

5.2.5 Resolution of sequence ambiguities

To resolve the ambiguous positions, 353 PCRs, including 9 long PCRs, were employed (primers are listed in Appendix II). The regions representing the multiple copies of IS and large repeats were included to ensure that the sequences of each repeat region were independent. The large repeated regions included two copies of the 33.9 kb pathogenicity island within the genome. Large repeats were resolved by long PCR to amplify each specific repeat, and then short PCR products were obtained from within the long PCR

Table 5-1 Primers used for Gap closure. Gene names and nucleotide position numbers of each primer are specified relative to the *F. tularensis* Schu S4 genome.

Primer Name	Sequence of primer (5′-3′)	Gene name	Start Position
GAP1-F	GGGTGGTAGAACAGTAGGTGCT	tufA	151181
GAP1-R	TAGGACTCGAACAACTTCACCA	nusG	152302
GAP2-F	TAACCACCACGAACGTGATTAG	rpsA	199799
GAP2-R	GTCGCCACCATATAAAACAACA	ddlB	200808
GAP3-F	TGATAGCCAAACATCTGTATTTC	Intergenic region	462253
GAP3-R	CACTGTTATCAAGCATATTTGGTG	FTT0447c	463351
GAP4-F	TGTGTCTTTCGTGAGATAAAGC	coaD	598922
GAP4-R	AGTCGCACCATTATCCTGAAT	fopA1	600006
GAP5-F	GTGGTACGCTCAATGTCAAAAG	lon	646585
GAP5-R	AGCAACATATGCTCTACTACCAA	FTT0628	647573
GAP6-F	GAGGATTTACTTCAAAGAGGGCT	FTT0663	682060
GAP6-R	TGGCAATCCATATAGTAAAGGC	hdc	683079
GAP7-F	GCATCAGCTTTAGAAATGTTTGG	ispD	729862
GAP7-R	TGTACTCTATTGGTCTAAATAGCGA	sdaC1	730932
GAP8-F	CAGTGATATAGAGCATCAAAAGA	Intergenic region	744147
GAP8-R	TGATACAAAGATACGTACAAGTTTT	Intergenic region	745292
GAP9-F	TTTTGAGCATGATTAGAAAATA	Intergenic region	780378
GAP9-R	GAAATAACTAGAAACTTACCCATTCAA	recF	781521
GAP10-F	TTCTAAGCTAGTATTATAAATTTGTTG	yjjV	781849
GAP10-R	ACGGTCCTTGCTTTTCAATAA	glyS	783010
GAP11-F	TCAAAGTTGCACCTAGTTTAAT	FTT0796	815587
GAP11-R	TGTTTGATTAATAACAGATTCTAAA	FTT0798	816743
GAP12-F	ATGCAACAACATCATAGCATCC	ans	822295
GAP12-R	CAACTTTGCAAATGATTGAT	Intergenic region	823450
GAP13-F	ATCAAAAAGCCAGTTTAGCTGC	FTT0868c	879125
GAP13-R	TCTTGCCACATATACTGAAATCG	FTT0869	880126
GAP14-F	TTCGGCATTATTAGATTCGCTT	FTT0881c	889368
GAP14-R	TTCTCAAGTCTATGTACAGCTTCCTT	FTT0882	890457
GAP15-F	GGCAAATATTAAAAATTATGGCAAA	FTT0944	955092
GAP15-R	AAAATTGCTGTTTCTAATTCAA	FTT0944	956195
GAP16-F	GGTAGCATCGCAACAATACTCA	FTT0944	956274
GAP16-R	TTTTCTATTTGTTTGATTAGTTCCA	FTT0945	957357
GAP17-F	TTTCATCAGCATTTGCATCAAT	FTT0981	992870
GAP17-R	GAAGCAATAATTATAACTACAAAAGA	FTT0982	994003
GAP18-F	CGACTTAATTCCTTGTAGCCCA	FTT1007c	1020110
GAP18-R	TTTCATCAATATAAAGTTGAGCTG	FTT1009	1021271
GAP19-F	AAAAACTCAAAACTCCAAAAA	FTT1089	1099430
GAP19-R	AGGAGAGTTTTGGCATTCCTC	FTT1091	1100615
GAP20-F	TTGTGAAACTGGAATTGTTTTAGG	dedA2	1239896
GAP20-R	CATTGTTACTAAAGAAAAATGCCAAA	dedA1	1241017
GAP21-F	TGGATCTTCTAAGAGGGAAAAA	FTT1319	1347030
GAP21-R	CAAGCATGTGAGATATGCCAGT	FTT1321	1348270
GAP22-F	TGCAGACCCAGCAATATCTAAA	FTT1666c	1734994
GAP22-R	TTGTTATAGCCGTGTGAAAAAT	FTT1667	1736061
GAP23-F	TTTAAGTCCTTGTGGCTTCTCA	Intergenic region	1759181
GAP23-R	CCATTAAGGTTGTCAATTTGGT	FTT1691	1760609
GAP24-F	AAAATTTAGCAAACAATCAA	Intergenic region	1875891
GAP24-R	CCTATGTTAGGTTTGAGAGTTAAT	Intergenic region	1876899
GAP25-F	CCTATTTTATTCAAATCAGCGA	Intergenic region	1886410
GAP25-R	CCAGCTAGACCGAATAAGGTTG	trpG2	1887442

product. These short PCR products were sequenced directly on both strands using the PCR primers. All the sequencing data obtained were assembled into FSC 198 genome.

5.2.6 Genome comparison

Upon final assembly, whole genome comparison of Schu S4 and FSC198 was performed. The comparisons revealed a number of Single Nucleotide Polymorphisms (SNPs). DNA fragments including these SNPs were all PCR amplified and re-sequenced from both the Schu S4 and FSC198 strains to determine whether the apparent differences represented true SNPs or errors in one of the genome sequences. In addition, to examine the diversity among European subspecies *tularensis* isolates, the confirmed SNPs and the VNTR differences identified in the previous study (Johansson *et al.*, 2004) were also sequenced in strains FSC 199 and Sev-23. These sequence data were provided by the HPA.

5.3 Results

5.3.1 No large differences are revealed by whole genome PCR scanning

The initial assembly from the original HPA shotgun sequences resulted in 30 large contigs (R. Chaudhuri, unpublished observations). Alignments with the Schu S4 genome indicated an extremely close relationship between FSC198 and Schu S4. 119 primer pairs were designed to span the whole genome of *F. tularensis* Schu S4, and these primers were used to perform long PCR scanning on the genomic DNA of *F. tularensis* FSC198. All but two of these PCRs were successful and yielded products of the expected size (data not shown). Given the initial assembly and data from the previous VNTR study (Johansson *et al.*, 2004), it was concluded that there were no large differences between the two genomes. Based on this observation, the Schu S4 genome sequence was used to facilitate the finishing of the FSC198 genome.

5.3.2 Genome finishing

To finish the genome, the focus was on gap closure, removing low quality regions and resolving multiple copies of IS elements and large repeats. Gap closure was performed by PCR and subsequent sequencing using the PCR primers. Each PCR generated a product no longer than 1.5 kb. Because the first ~100 base pairs of the sequencing reads are not reliable, the annealing sites of PCR primers were selected to be at least 200 base pairs from the gap. The fidelity of the gap sequences was confirmed by sequencing both strands of the PCR products. The 25 gap regions were all closed by PCR and sequencing in this project.

5.3.3 SNPs and VNTRs

As the project progressed, it became increasingly clear that the genomes of FSC198 and Schu S4 were virtually identical. Whole genome sequencing here also led to an approach for detecting SNPs. Upon final assembly, the two genomes were found to be different at forty apparent SNPs. Most of these apparent SNPs modified the sequence of the encoded proteins. Re-sequencing the SNP loci in both genomes suggested that the majority of the discrepancies were due to errors in the published Schu S4 sequence rather than genuine differences. These were found to correspond to regions of poor sequence coverage from the Schu S4 sequencing project (M. Forsman, personal communication). Only eight SNPs were confirmed finally (Table 5-2). All the SNPs are non-synonymous, suggesting that adaptive evolution is responsible. Similarly, VNTRs were different at five loci, including three loci identified in Johansson's study (Johansson *et al.*, 2004). However, only three were confirmed following re-sequencing of these loci in both genomes (Table 5-3). Again, the results revealed that sequence errors appeared in the published Schu S4 genome.

To further investigate the diversity of the European isolates of *F. tularensis* subspecies *tularensis*, the eight true SNPs identified in this study, and the twenty five VNTR loci identified previously (Johansson *et al.*, 2004), were sequenced in the other available European strains, FSC 199 and Sev-23. Three of the eight SNPs that distinguish FSC198 from Schu S4 were also identified in FSC199 and Sev-23 (Table 5-2). This suggests that all three European isolates share a common ancestor after their divergence from the Schu S4 strain. When sequencing the twenty five VNTR loci, the same three loci identified in FSC198 were also variable in FSC199 and Sev-23 (Table 5-3), but there were differences between the three strains. This suggests that the three European isolates are distinct from

Table 5-2 Single nucleotide differences identified between the FSC 198 genome and the published Schu S4 sequence and confirmed by re-sequencing (reproduced from Dr. Chaudhuri). The bases at the equivalent loci in strains FSC 199, Sev-23 and LVS are also shown. The SNPs relative to Schu S4 that are conserved in all three European strains are highlighted in red.

Locus	Schu S4 coordinate	FSC 198 coordinate	Schu S4 base	FSC 198 base	FSC 199 base	Sev-23 base	LVS base	Effect in FSC 198
S1	390291	390243	C	T	T	T	C	FTF0387 substitution A→V
S2	621878	621830	C	T	C	C	C	FTF0602c substitution V→I
S3	639511	639463	C	A	C	C	C	FTF0620 substitution A→E
S4	701628	701580	G	T	T	T	G	SthA substitution Q→K
S5	911511	911463	C	T	T	T	C	FTF0903 substitution R→C
S6	1007564	1007516	G	C	G	G	G	YbhO substitution E→D
S7	1008149	1008101	G	A	G	G	G	YbhO substitution M→I
S8	1134418	1134369	G	A	G	G	G	MetN substitution M→I

Table 5-3 Variable number tandem repeat (VNTR) differences between the FSC 198 genome and the published Schu S4 sequence confirmed by resequencing (reproduced from Dr. Chaudhuri). The number of tandem repeats at each locus for strains FSC 199, Sev-23 and LVS are also shown. VNTR loci are numbered as in Johansson's study (Johansson *et al.*, 2004).

Locus	Schu S4 coordinate	FSC 198 coordinate	Repeat Unit	Schu S4 copy no.	FSC 198 copy no.	FSC 199 copy no.	Sev-23 copy no.	LVS copy no.
M8	8266	8266	aagatatttgtagaaa	4	5	5	5	2
M3	308635	308650	aataaggat	21	14	25	28	13
M10	1283659	1283610	aagatatttgtagaaa	18	11	11	10	2

each other, but more closely related to each other than to Schu S4.

Additionally, the eight identified SNPs and three VNTR differences were investigated in another subspecies of *F. tularensis*, the subsp. *holarctica* strain LVS. Interestingly, all eight SNPs showed identical residues between LVS and Schu S4, but the VNTRs are distinct from each other (Tables 5-2 and 5-3). The SNP results suggest that the Schu S4 genome has the ancestral genotype at these positions, and that the differences in the European isolates are due to substitutions that have occurred in that lineage.

5.3.4 Complete genome sequence of FSC198

The complete genome sequence of *F. tularensis* subspecies *tularensis* strain FSC 198 is 1,892,616 bp in length, with a GC content of 32.36% (see Figure 5-3) (Chaudhuri *et al.*, 2007). An online comparative genomics database (http://xbase.bham.ac.uk/ftbase/) has been developed, allowing effective comparison among the up-to-date sequenced *Francisella* genomes. Genome sequencing shows that the European isolate FSC198 is almost identical to the US laboratory strain Schu S4.

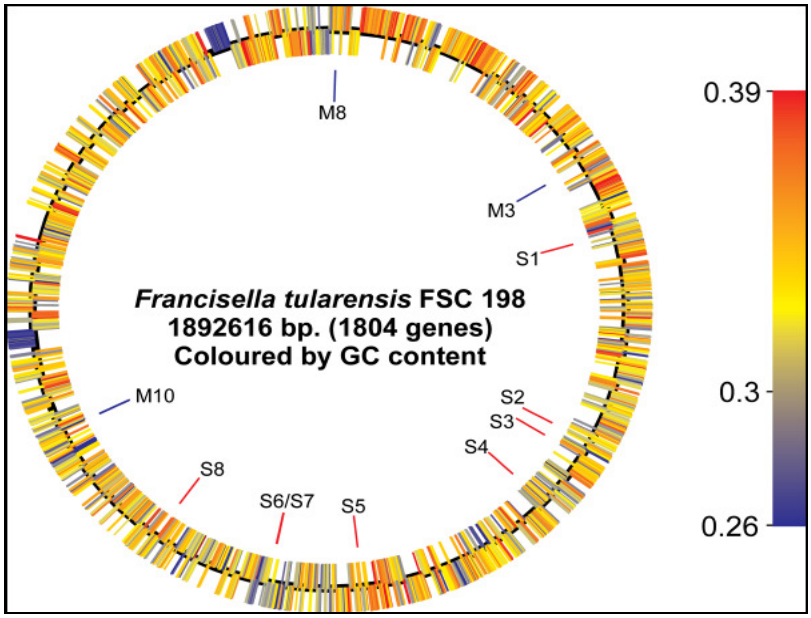

Figure 5-3 Circular representation of the complete genome sequence of FSC198 (reproduced from http://xbase.bham.ac.uk). Predicted coding sequences are coloured according to their GC content. The inner circle indicates the positions of SNPs (red) and VNTR differences (blue) relative to the published Schu S4 genome sequence. SNP and VNTR loci are numbered as in Tables 5-2 and 5-3, respectively (Chaudhuri *et al.*, 2007).

5.4 Discussion

Interest in the bacterium *Francisella tularensis* has been greatly increased in recent years to counter its potential application as a bioterrorist agent. The subspecies *tularensis* is particularly troublesome because of its ability to cause incapacitating, potentially fatal disease and to spread via aerosolization. It has the potential to threaten the public health system in the event of an outbreak. *F. tularensis* was traditionally considered endemic only within the Northern Hemisphere, with *F. tularensis* subsp. *tularensis* restricted to North America. However, there have been reports of several isolates belonging to *F. tularensis* subsp. *tularensis* from Central Europe (Gurycova, 1998) and an isolate of *F. tularensis* subsp. *novicida* from Australia (Whipp *et al.*, 2003). These indicate that either the geographical distribution of *F. tularensis* is more pervasive than originally thought or that the pathogen is spreading. The European isolates of *F. tularensis* subsp. *tularensis* have remained an enigma since first being isolated in the 1980s. There is considerable interest in discovering how these highly virulent subspecies *tularensis* strains were introduced into Europe. For this purpose, one of the isolates, FSC198, has been investigated by whole genome sequencing.

The genome sequencing reported here revealed that the genomes of FSC198 and Schu S4 are almost identical. Only eight SNPs and three VNTRs were confirmed that distinguished the two strains. Both SNPs and VNTRs can be used as molecular markers for efficient genotyping. SNPs represent the smallest-scale variation in bacterial genomes and can occur, in theory, anywhere in a genome. For this reason, it is very unlikely that the same change happens in two separate lineages in a short space of time. In contrast, VNTR changes can happen at only a small number of sites due to variations in repeat copy. However, these

sites can change frequently and allow more variations than a single base change (there are only four bases).Therefore, both SNPs and VNTRs are useful in distinguishing between the two genomes and characterising their genetic distance. Particularly, they have vital value to those species with little genetic variation, e. g. *Mycobacterium tuberculosis*, *Yersinia pestis* and *Bacillus anthracis* (Keim *et al.*, 2000; Mazars *et al.*, 2001;Gutacker *et al.*, 2002; Touchman *et al.*, 2007). Furthermore, if two genomes share some SNPs and VNTRs, it implies a shared history and close relatedness between them.

VNTRs as a previous study has shown (Johansson *et al.*, 2004), are extremely useful and powerful tool in estimating evolutionary relationships between world-wide populations of *F. tularensis*. Here in this study, we have shown the value of using whole-genome sequencing to detect SNPs in addition to VNTRs. As illustrated in Table 5-2, three European subsp. *tularensis* strains, FSC198, FSC199 and Sev 23, possess three identical SNPs, which differ from Schu S4. This result suggests that the European strains are more similar to each other than to Schu S4. Therefore, it is reasonable to conclude that the three European isolates are from the same lineage and that they share a more recent common ancestor with each other than with Schu S4. This was also confirmed by an examination of all the SNPs in a subspecies *holarctica* strain LVS. As shown in Table 5-2, LVS is identical to Schu S4 at all eight positions that represent SNPs in the European isolates. Since the two subspecies *tularensis* and *holarctica* are known to be genetically distant from each other (Johansson *et al.*, 2004), the LVS/SchuS4 sequences are likely to represent the ancestral state, while the SNPs in the three European isolates are due to substitutions that have occurred in the European lineage. However, when these changes occurred is still a mystery.

In order to uncover the origin of the subspecies *tularensis* in Europe, the immediate questions are how did the common ancestor of the European isolates arise and how did it come to be in Europe, particularly as, according to the published VNTR results, FSC 198 and other European isolates are the closest relatives of Schu S4, far closer than any other clade A.I isolates from central USA or elsewhere in North America. One possibility is that past human activities led to the establishment of Schu S4-like populations in Europe.

The most plausible potential sources of FSC198 and the other European subspecies *tularensis* are laboratory stocks of Schu S4 in North America or in Europe. Since Schu S4 is a recognised laboratory strain propagated in laboratories around the world, sub-culture in different laboratories might explain the differences between the European isolates. The most worrying scenario is that these isolates were released deliberately into the European environment. Alternatively, they could have entered the environment through inappropriate disposal of laboratory waste, or even escape of mammals or anthropods that have been infected in the laboratory.

Another possibility is that FSC198 and other European isolates are derived not from the laboratory strain Schu S4, but from a related wild strain from the USA. It has been suggested that the *Francisella tularensis* clade A.I subgroup could be transferred via dogs within continental USA, and via the cotton tail rabbit from USA to Europe (Farlow *et al.*, 2005). Movements of animal or animal products might also carry insect vectors containing the progenitor of the European strains. However, it seems very unlikely, bearing in mind the published VNTR results (Johansson *et al*., 2004), that a randomly selected wild strain would be so similar to Schu S4. In this study (see Table 5-2), none of the eight SNPs

identified are specific to Schu S4. Therefore, it is most likely that Schu S4, rather than any other wild strain, represents an immediate precursor of the European strains.

Another possibility is that a laboratory error occurred in Slovakia, so that the alleged environmental isolates represent contaminants derived from an existing laboratory stock of Schu S4. However, the three European isolates were distinct from each other at some SNPs and obtained at different times, so this explanation would require repeated contamination and sub-culture from the same laboratory stock, which seems very unlikely.

In summary, the whole genome sequence of a European subspecies *tularensis* strain, FSC198, has been determined in this study. The most likely source of the European isolates of *F. tularensis* subspecies *tularensis* is the laboratory strain Schu S4. Given the pathogenic potential of this subspecies and its status as a public health threat, further environmental sampling to assess the distribution and prevalence of the subspecies *tularensis* in Europe is probably warranted.

The FSC198 sequence is the first publicly available bacterial genome sequence to be determined in the United Kingdom outside of the Wellcome Trust Sanger Institute. Additional *Francisella* genome sequencing projects are now underway, with several new genome sequences appearing in the public databases (Table 5-4) and/or in the process of assembly and finishing. These new genome sequences will help prime the development of comparative genomics and proteomic studies and the search for new vaccine candidates (Titball and Petrosino, 2007).

Table 5-4 Current *Francisella* genome sequencing projects.

F. tularensis strains	Sequence status	Length (bps)	Sequencing Center	GenBank Accession number
F. tularensis subsp. *tularensis* SCHU S4	Complete	1 892 819	Swedish Defence Research Agency	AJ749949
F. tularensis subsp. *holarctica* OSU18	Complete	1 895 727	Baylor College of Medicine	CP000437
F. tularensis subsp. *tularensis* FSC198	Complete	1 892 616	University of Birmingham	AM286280
F. tularensis subsp. *holarctica* LVS	Complete	1 895 994	Lawrence Livermore National Laboratory	AM233362
F. novicida U112	Complete	1 910 031	University of Washington	CP000439
F. tularensis subsp. *holarctica* FTA	Complete	1 890 909	US Department of Energy Joint Genome Institute	CP000803
F. tularensis subsp. *tularensis* WY96-3418	Complete	1 898 476	Translational Genomics Research Institute	CP000608
F. tularensis subsp. *mediasiatica* FSC147	Complete	1 893 886	The U.S. Department of Energy (DOE) Joint Genome Institute	CP000915
F. tularensis subsp. *holarctica* FSC200	Assembly	1 790 358	University of Washington	AASP00000000
F. tularensis subsp. *tularensis* FSC033	Assembly	1 844 205	Broad Institute Genome Sequencing Platform	AAYE00000000
F. novicida GA99-3548	Assembly	1 845 491	Broad Institute Genome Sequencing Platform	ABAH00000000
F. novicida GA99-3549	Assembly	1 897 440	Broad Institute Genome Sequencing Platform	AAYF00000000

CHAPTER SIX

INVESTIGATION OF TWO POTENTIAL VIRULENCE FACTORS IGLA AND IGLB FROM *FRANCISELLA TULARENSIS* ISOLATE FSC198

6.1 Introduction

6.1.1 Identification of *Francisella* virulence factors

Very little is known about the mechanisms of *Francisella* virulence. Replication of the bacterium has been demonstrated in a range of cell types, but mainly in macrophages (Titball *et al.*, 2003). Several factors needed for *Francisella* intra-macrophage growth are responsible for *Francisella* virulence. A biochemical study of the LVS of *F. tularensis* showed that four proteins are induced after *F. tularensis* entry into macrophages (Golovliov *et al.*, 1997). The most prominently induced of these was identified as a 23 kDa protein. It has been cloned and sequenced in this study. However, this protein does not have any homologues within the databases at the present time.

Genetic approaches have also been used to discover the factors that are needed for *Francisella* intracellular growth. A random transposon (Tn) mutagenesis approach identified five genetic loci that reduced the ability of *F. tularensis* to grow and replicate in mouse macrophages (Gray *et al.*, 2002). One of these transposon insertions was located within the gene encoding the 23 kDa protein. Not far from it (approximately 2 kb upstream), there was a second Tn mutant associated with growth in macrophages. The gene encoding the 23 kDa protein was found to be part of a putative operon consisting of four genes, namely *iglABCD* for intracellular growth locus (Gray *et al.*, 2002). The 23 kDa induced protein was identified as IglC, while the upstream gene *iglA* was identified as the other locus revealed by Tn mutagenesis. Three other Tn mutants were found at the genetic loci that encode the ClpB heat-shock protease, and the homologues of glutamine

phosphoribosylpyrophosphate amidotransferase (purine biosynthesis), and alanine racemase (peptidoglycan biosynthesis).

The role of IglC in *Francisella* virulence has been studied for several years. IglC was demonstrated to be necessary for *F. tularensis* to survive in phagocytes (Golovliov *et al.*, 2003a). It also affects the escape of *F. tularensis* from the phagosome via a mechanism that may involve degradation of the phagosomal membrane (Golovliov *et al.*, 2003a; Lindgren *et al.*, 2004), and it is essential for the induction of apoptosis in host cells (Lai *et al.*, 2004).

6.1.2 The Francisella Pathogenicity Island (FPI) is essential for *Francisella* virulence

With the completion of the *F. tularensis* Schu S4 genome (Larsson *et al.*, 2005), two identical copies of a large (30 kb) gene cluster were immediately apparent (see Figure 6-1). Each gene cluster contained a low G+C content of 26.6%, in contrast with the genomic average of 32.2%. The cluster was named the *Francisella* pathogenicity island (FPI), because *Francisella* strains with experimentally induced mutations in FPI genes are highly attenuated in virulence and show defects in *Francisella* intramacrophage growth (Nano *et al.*, 2004). FPI was the first pathogenicity island discovered in the *Francisella* genome. Interestingly, two clinically important *Francisella* subspecies, *tularensis* and *holarctica*, have two copies of the FPI, but subspecies *novicida,* which is avirulent in humans, contains only one copy in its genome (Nano *et al.*, 2004). This may be one reason for the lower virulence of *F. novicida*. Currently, the functions of the proteins encoded by the FPI are the focus of much research (Oyston, 2008).

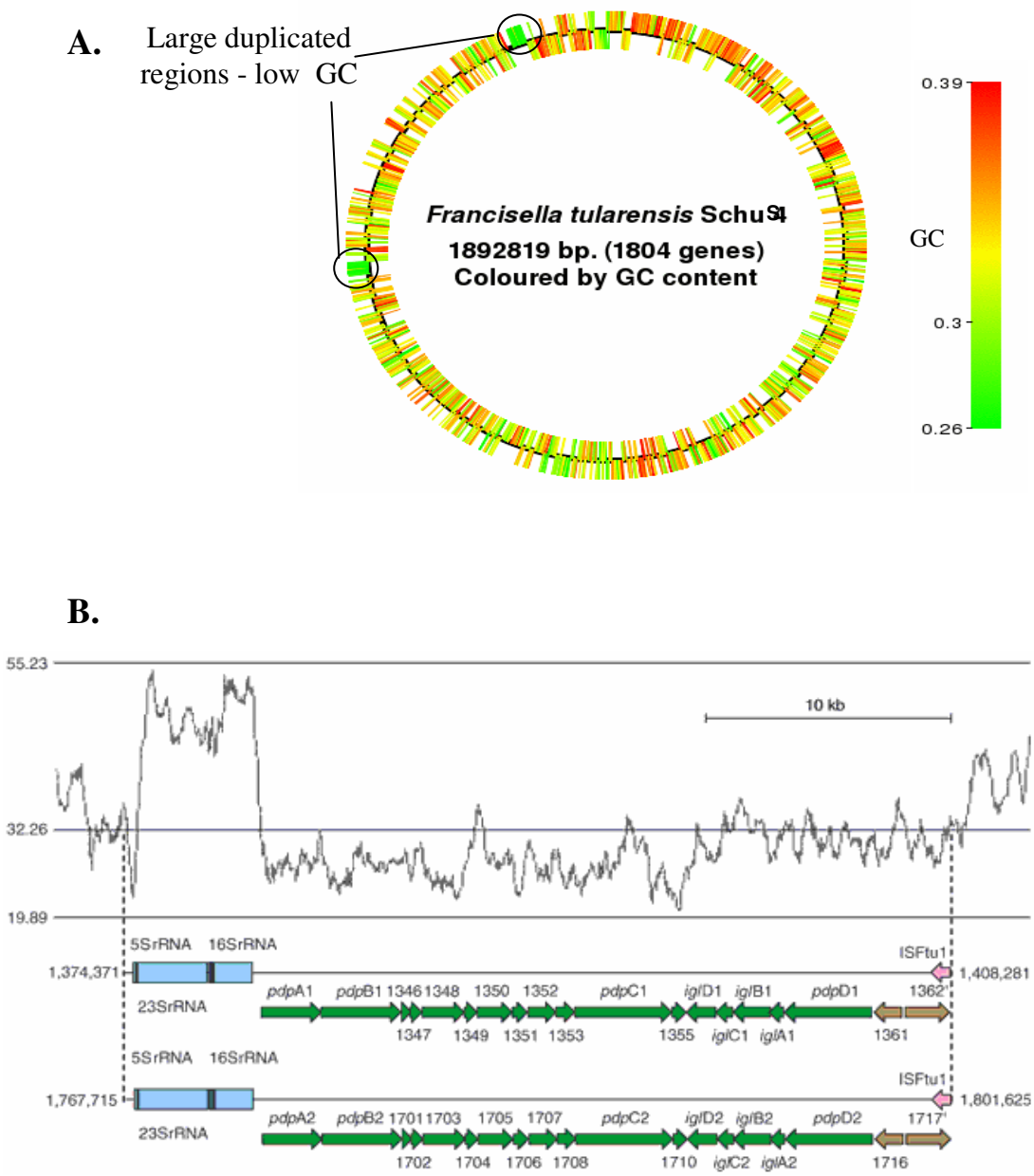

Figure 6-1 The *Francisella* pathogenicity island (FPI). A. Two large duplicated regions were identified by low GC content in *F. tularensis* subspecies *tularensis* strain Schu S4 (image from http://xbase.ac.uk). B. The organisation of two identical copies of *Francisella* pathogenicity islands were revealed from subspecies *tularensis* strain Schu S4 genome (Larsson *et al*., 2005). The leftmost scale shows GC content. Blue represents RNA encoding regions; green represents open reading frames encoding hypothetical proteins; brown represents pseudogenes; and pink represents IS elements. Open reading frame labels refer to the corresponding annotated gene or FTT (for *Francisella tularensis* subsp. *tularensis*) number in the genome sequence of Schu S4 (Larsson *et al.*, 2005).

The *Francisella* pathogenicity island consists of a cluster of 16-19 predicted genes. Tellingly, the *iglABCD* genes are part of the *Francisella* pathogenicity island. Besides the *igl* genes, two other virulence genes, *pdpD* and *pdpA* (for pathogenicity determinant protein), were also found in the FPI. Interestingly, *pdpD* was found in the highly virulent subspecies *tularensis*, but not in the moderately virulent subspecies *holarctica* (Nano *et al.*, 2004). Over-expression of PdpD affects the cellular distribution of the FPI-encoded proteins IglA, IglB, and IglC (Ludu *et al.*, 2008). Disruption of the *pdpA* gene rendered *F. tularensis* unable to grow in macrophages and avirulent in mice (Nano *et al.*, 2004).

Transcriptional regulation of the FPI genes is thought to be affected by two factors: MglA and MglB (for macrophages growth locus) (Baron and Nano, 1998). MglA and MglB are homologues of SspA and SspB from *Escherichia coli*, which are global regulators of the stringent starvation response, and SspA and/or SspB influence the levels of multiple proteins under conditions of nutritional stress in *E. coli* (Williams *et al.*, 1994). There is experimental evidence showing that an intact *mglA* gene was required for transcripts containing *pdpD*, *iglA*, *iglC*, *iglD,* and *pdpA* to be expressed (Lauriano *et al.*, 2004). A microarray analysis of gene expression found a total of 102 MglA-regulated genes, including all of the FPI genes (Brotcke *et al.*, 2006). In the strain lacking an intact *mglA* gene, the FPI genes were repressed by about three to six- fold. IglC, together with its regulator MglA, was reported to be essential for interfering with phagosome biogenesis and subsequent bacterial escape into the cytoplasm (Santic *et al.*, 2005).

6.1.3 *iglA* and *iglB* are conserved and have the same organisation in many other bacteria

IglA and IglB orthologues have been reported in a variety of animal, plant pathogens or symbionts, most of which are living intimately with eukaryote cells (Nano *et al.*, 2004). These two sets of orthologues are each characterised by a conserved domain with unknown function (DUF), i.e. *Francisella* IglA possesses DUF770 and IglB possesses DUF877, respectively. Using the COG (Clusters of Orthologous Groups of proteins) (http://www.ncbi.nlm.nih.gov/COG/) classification, IglA has strong identity to members of COG3516 and IglB has strong identity with COG3517.

IglA and IglB homologues are most often encoded within a conserved gene cluster, encoding what are known as IcmF-associated homologous proteins (IAHPs) (Das and Chaudhuri, 2003). Within IAHPs, the homologues of IglA and IglB are encoded adjacent to each other and organised in the same gene order. Some large IAHP clusters also include ClpB, an ATP-dependent protease. IAHP gene clusters contain about 15-20 genes surrounding an *icmF*-like gene (Das and Chaudhuri, 2003). The deviant G+C% content of the IAHP cluster suggests that these genes were possibly transferred in a phage mediated manner (Das and Chaudhuri, 2003). The *icmF* gene was previously reported to encode a structural component of the type IV secretion system (T4SS) in the study of *Leginella pneumophila*. The *L. pneumophila* T4SS was encoded by 26 *dot* and *icm* genes (refer to Chapter 1) (Sexton *et al.*, 2004).

Recently, Pukatzki *et al*. demonstrated that a set of genes named *vas* genes (for Virulence-Associated Secretion) were essential for the cytotoxicity of *V. cholerae* cells toward host

macrophages. The IAHPs from *Vibrio cholerae* were identified as the components of a novel protein secretion system. It was named a Type VI secretion system (T6SS), in addition to other Type I-V secretion systems (Pukatzki *et al.*, 2006). The *vas* gene cluster-encoded T6SS mediates the extracellular secretion of four distinct proteins [Hcp (hemolysin-coregulated protein) and VgrG (valine-glycine-repeats G) -1, -2 and -3], all of which are lacking recognised N-terminal leader sequences and are secreted in a *sec*-independent manner. However, it is not entirely clear whether Hcp and VgrG are truly secreted effector proteins or are actually components of the T6SS (Filloux *et al.*, 2008). The full picture of secretion by T6SS has yet to be elucidated. However, the lack of T6SS in non-pathogenic organisms strongly suggests a crucial role of T6SS in imparting pathogenicity (Shrivastava and Mande, 2008).

The T6SSs have been reported in many other bacteria, representing the SCI (*Salmonella enterica* centisome 7 genomic island) locus from *Salmonella enterica* (Folkesson *et al.*, 2002),the Imp (impaired in nitrogen fixation) locus from *Rhizobium leguminosarum* (Bladergroen *et al.*, 2003), the HSI-1 (Hcp1 secretion island -1) locus from *Pseudomonas aeruginosa* (Mougous *et al.*, 2006) and the Evp (*Ed. tarda* virulence protein) locus from *Edwardsiella tarda* (Rao *et al.*, 2004). In the *E. tarda* T6SS, the secretion of EvpC, an Hcp homologue, and EvpP, a secreted protein with no homology to either VgrG or Hcp proteins, were mediated by 13 other *evp* genes (Zheng and Leung, 2007). The insertion mutants of either *evpA* or *evpB* led to the lack of secreted protein EvpC, suggesting a role for *evpAB* in extracelluar protein secretion (Rao *et al*., 2004). EvpA and EvpB have 25% and 30% identity to *Francisella* IglA and IglB, respectively (Rao *et al*., 2004). Furthermore, *iglA* and *iglB* have similarities to *impB* and *impC* from *R. leguminosarum* respectively, which

are thought to be involved in temperature-dependent protein secretion. The *impB*-encoding protein was implicated in pathogenicity and protein secretion (Bladergroen *et al.*, 2003). The study on T6SS gene components of *V. cholerae* demonstrated that VCA0107 (IglA-like) and VCA0108 (IglB-like) are inner membrane proteins (Shrivastava and Mande, 2008). Though the mechanism of their assembly was not known, VCA0107 and VCA0108 were predicted to assist in stabilising the T6SS gene complex and in bringing about other intracellular activities essential for the functioning of the T6SS. In addition, IglA and IglB homologues have been identified in many other gram-negative bacteria (see Table 6-1).

In order to define whether the *Francisella* pathogenicity island encodes a T6SS, both IcmF and DotU-like proteins were searched (Nano and Schmerk, 2007). Two FPI-encoded protein, PdpB and PigF, exhibited low homology with IcmF and DotU, respectively (Nano and Schmerk, 2007). The existence of an *icmF*-like gene (*pdpB*) and a *dotU*-like gene (*pigF*) and the conserved gene pair *iglA* and *iglB* suggests the existence of a FPI-encoded T6SS. Importantly, the high conservation of IglAB homologues suggests that these proteins could be the major requirement for a functional T6SS.

Table 6-1 IglA and IglB homologues identified in other bacterial species.

Bacterial species	Protein name (possess conserved domain DUF770)	Protein name (possess conserved domain DUF877)
Francisella tularensis	IglA	IglB
Salmonella enterica	SciH	SciI
Edwardsiella tarda	EvpA	EvpB
Pseudomonas aeruginosa	PA0083	PA0084
Vibrio cholerae	VCA0107	VCA0108
Rhizobium leguminosarum	ImpB	ImpC
EAEC	AaiA	AaiB
Burkholderia mallei	TssA	TssB

6.1.4 Aims

Previous studies have shown the important role of the *Francisella* pathogenicity island, and the *igl* locus within it, in virulence. The most convincing candidate virulence gene, *iglC*, has been widely investigated. In this project, interest was focused on another two potential virulence factors, *iglA* and *iglB*. These genes have homologues in many bacterial species, which occur as conserved gene pairs with the same organisation. In many cases, the conserved gene cluster has been associated with a type VI secretion system.

It has been suggested that proteins encoded by conserved gene pairs are likely to interact physically (Dandekar *et al.*, 1998). This led us to the hypothesis that IglA and IglB physically interact with each other. Additionally, the project aimed to express the two proteins in *E. coli*, to facilitate further studies and to gain insight into the functions of IglA and IglB.

The development of a licensable vaccine is a priority for *F. tularensis* research. Located in the FPI, both *iglA* and *iglB* have potential as vaccine candidates. In addition, deletion mutants of these genes are worth investigating as potential live rationally attenuated vaccines. Since a method for the construction of defined *Francisella* mutants through allelic replacement has been developed (Golovliov *et al.*, 2003b), this study aimed to make constructs of the two genes for mutagenesis, using a *F. novicida* strain, which is closely related to *F. tularensis*, but carries only one copy of the FPI.

6.2 Materials and methods

6.2.1 Bacterial and yeast strains, genomic DNA and plasmids

The genomic DNAs of *F. tularensis* subsp. *tularensis* strain FSC198 and subsp. *novicida* strain U112 were used in this project. The yeast strain *Saccharomyces cerevisiae* PJ69-4A was kindly provided by Dr. Gad Frankel (Imperial College, UK). The plasmids used in this study are listed in Table 6-2.

6.2.2 Yeast two-hybrid screen

6.2.2.1 Principle

The yeast two-hybrid assay was developed by Fields and Song (1989), and has been widely used for studying protein-protein interactions. UAS_{GAL} is the upstream activation sequence for the yeast GAL genes, which binds Gal4 (Figure 6-2). Gal4 functions as a transcriptional activator, which is required for the expression of genes encoding enzymes of galactose utilisation. The process of transcription requires two separate but functionally essential domains for binding to UAS_{GAL} DNA (BD) and for transcriptional activation (AD) (Figure 6-2).

Unless the BD domain physically binds with the AD, Gal4 cannot function as a transcriptional activator. The two-hybrid assay was prompted by expressing two fusion proteins: the Gal4 DNA-binding domain fused to a protein "X" (bait) and a Gal4 activating region fused to a protein "Y" (prey) in yeast. The physical interaction of the two constructed proteins, "bait" and "prey", can bring about the fusion of BD and AD, which leads to transcriptional activation of a reporter gene (Figure 6-2).

Table 6-2 Plasmid list.

Plasmid	Description	Reference
pGAD424	*ori*ColE1 *ori*2µ *LEU1* P_{ADH} : : *GAL4*' activator domain : : MCS Amp^R	(Bartel *et al.*, 1993)
pGBT9	*ori*ColE1 *ori*2µ *TRP1* P_{ADH} : : *GAL4*' binding domain : : MCS Amp^R	(Bartel *et al.*, 1993)
pMal-c2	maltose-binding protein (MBP) fusions, *malE* gene deletion	New England Biolabs
pPV2J	Amp^R, Cm^R, *sac*B, mob	(Golovliov *et al.*, 2003a)
pDONR201	*ori*PUC, $Km^{R,}$ Gateway entry cloning by BP recombination	Invitrogen Gateway system
pDEST17	bacteriophage T7 promoter, N-terminal polyhistidine (6xHis) tag, Cm^R, $Amp^{R,}$ Gateway destination cloning by LR recombination	Invitrogen Gateway system
pCR 2.1	bacteriophage T7 promoter, *Lac*Za gene, Amp^R, Km^R	Invitrogen TA Cloning system

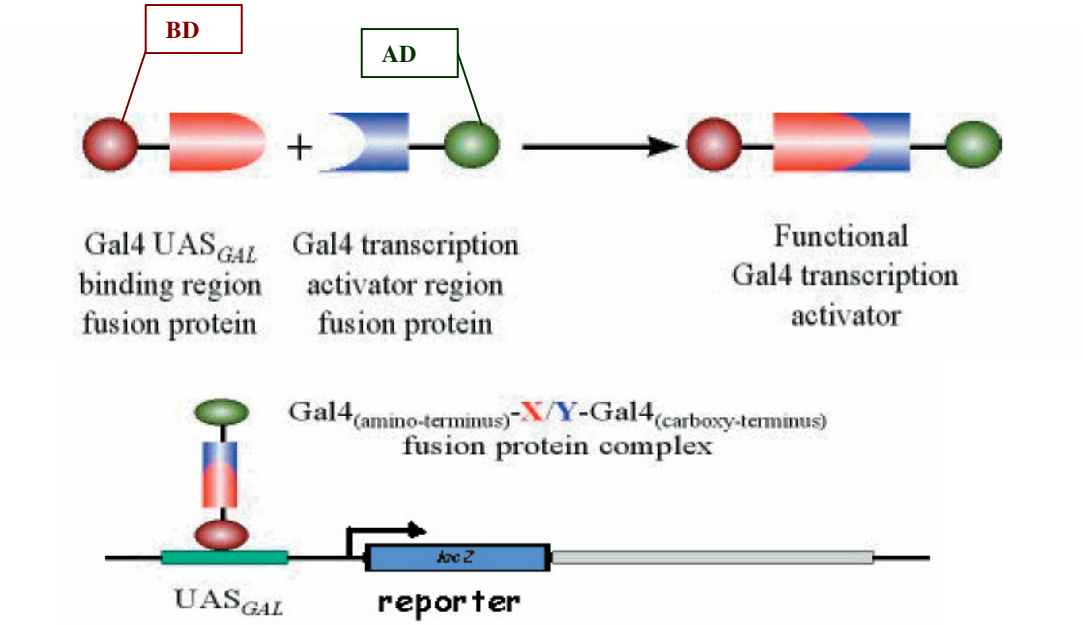

Figure 6-2 The principle of the yeast two-hybrid assay (Sobhanifar, 2003). BD, binding domain; AD, activation domain.

Following the yeast two-hybrid screen, measurement of β–galactosidase (β-gal) activity is an important step for indicating transcription of the *lacZ* reporter gene. β-galactosidase is encoded by the *lacZ* gene and able to hydrolyze (cleave) β-D-galactosides, which facilitates growth on carbon sources like lactose. In the assay, the substrate o-nitrophenyl-β-D-galactopyranoside (ONPG) was used in place of lactose. β-galactosidase cleaves ONPG to yield galactose and o-nitrophenol, which has a yellow colour and absorbs light at 420 nm.

6.2.2.2 Methods and protocols

In this study, two plasmids pGBT9 and pGAD424 are used as the binding and activation domains (BD and AD), respectively, of the yeast transcriptional activator GAL4 (Bartel *et al.*, 1993). Both vectors carry ampicillin resistance genes. The constructs were introduced into yeast strain PJ49-4A using the high-efficiency lithium acetate transformation procedure (Geitz and Schiestl, 1995). The pGBT9 "bait" plasmid carries the yeast TRP1 marker gene, so the "bait" constructs were selected on medium lacking tryptophan, whereas the pGAD424 "prey" vector carries the yeast LEU2 marker gene and the "prey" constructs were selected on medium lacking leucine. The yeast strain, PJ49-4A contains three separate reporter genes, *HIS3*, *ADE2* and *lacZ*. Therefore, co-transformants were replica-plated onto medium with the omission of tryptophan, leucine, adenine and histidine to select for activation of both *ADE2* and *HIS3* reporter genes (Creasey *et al.*, 2003).

The *iglA* and *iglB* genes were amplified from FSC198 genomic DNA using primers that were designed to create in-frame fusion proteins and incorporate *Bam*H1/*Pst*I sites at both ends (Table 6-3). The genes were then cloned into the *Bam*H1/*Pst*I sites of the yeast two-hybrid vectors pGBT9 and pGAD424, respectively. Four plasmids with DNA inserts were

Table 6-3 Primers used in IglA and IglB study.

Primer name	Sequence (5'-3')*
Y2H-IglA-F	AACGGATCCATGGCAAAAAATAAAATCCCAAATTC
Y2H-IglA-R	AAACTGCAGCTACTTATCATCTACTTGTTGATTACTTAAGTCT
Y2H-IglB-F	TCAGGATCCAGATGATAAGTAGAGAGGATTTTGTTATGA
Y2H-IglB-R	CTCCTGCAGTTAGTTATTATTTGTACCGAATAATTCTGGT
*attB*1-IglB	GGGGACAAGTTTGTACAAAAAAGCAGGCTTGGTAAGTAGGGAGGATTTTA
*attB*2-IglB	GGGGACCACTTTGTACAAGAAAGCTGGGTTTAGTTATTATTTGTACCGAATAA
Mt-*iglA*-up	TTGATCTAGAGCGGCTATGTGCTTCGATTTGT
Mt-*iglA*-down	TTTAGTCGACGGATGCTCAAGCAAAGCTTCAA
Mt-*iglB*-up	TTATTCTAGAAGGAAGATCTGTGGATGCAAAA
Mt-*iglB*-down	TCTTGTCGACCACCCATAAGTTCTGTTGGCTCT
Inv-*iglA*-F	TTTTACGTACTTATTGTCCTTTTTTTCACAACACC
Inv-*iglA*-R	AGATACGTAGTAGGGAGGATTTTATTATGAC
Inv-*iglB*-F	CGCTACGTAGTAAGAGGTTCCAACTTTCAC
Inv-*iglB*-R	CGCTACGTATTACGCCCCGCCCTGCCACTCATC
Inv-*cat*-F	CGCTACGTAGTAAGAGGTTCCAACTTTCAC
Inv-*cat*-R	CGCTACGTATTACGCCCCGCCCTGCCACTCATC

* The restriction site in each primer is underlined.

constructed at this stage: pGAD424-IglA, pGAD424-IglB, pGBT9-IglA and pGBT9-IglB. The in-frame constructs were confirmed by sequencing the fusion plasmids with the vector primers (sequence not shown; primers were kindly provided by Helen Betts).

A positive control was set up by using co-transformant yeast with plasmids pGAD424-EspD and pGBT9-CesD. A negative control was established by the combination of plasmids pGAD424-CesD and pGBT9-EspD. The plasmid DNAs for both controls were kindly provided by Dr. Helen Betts from our laboratory. Both the positive and negative controls had previously been shown to work successfully (Betts, personal communication).

To measure β-galactosidase activity, the accumulation of yellow colour (increase in 420 nm absorbance)/minute is monitored (Miller, 1972). The following equation was used to calculate units of enzyme activity: 1 Miller Units = 1000 x OD_{420} / (T x V x OD_{600}). T represents the reaction time in minutes, V represents the volume of culture used in the assays, and OD_{600} reflects the cell density. OD_{420} is the absorbance of the yellow o-nitrophenol.

6.2.3 Protein over-expression and purification

Proteins IglA and IglB were over-expressed in *E. coli*. An N-terminal MBP-IglA (Maltose Binding Protein) fusion protein was created. The gene *iglA* was cloned into the *Bam*H1/*Pst*I sites of the MBP vector, pMal-c2 (Table 6-2).

The His-tag-IglB fusion protein was made following the Gateway cloning procedure (Invitrogen, UK). The *iglB* gene was amplified by PCR with primers *attB*1-IglB and *attB*2-

IglB (Table 6-3). The gene was cloned into the Gateway entry vector (pDOR201), and then sub-cloned into the Gateway destination vector (pDEST17). The fidelity of all the constructs was confirmed by DNA sequencing. Next each plasmid DNA was transformed into *E. coli* BL21 competent cells and the cells containing the plasmid DNA were subsequently induced by 1 mM Isopropyl-β-D thiogalactopyranoside (IPTG) at mid log stage. The over-expressed proteins were examined using sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), with the gels being stained with Coomassie blue. Once expression of protein was confirmed by SDS-PAGE, and then 1 litre of the culture was prepared for protein purification. The purified protein was subsequently used to raise antibodies in rabbits.

Protein purification was performed using column chromatography methods, using amylose resin for the MBP fusion protein (New England Biolabs, UK) and Ni-NTA resin for the His-tagged fusion protein (The QIAexpressionist kit, Qiagen, UK). The purified proteins were analysed by SDS-PAGE and Western blots.

6.2.4 Creation of constructs for mutagenesis

A PCR strategy known as inverse PCR was used for making chromosomal mutants of *iglA* and *iglB* (Figure 6-3 i). Regions containing ~500 bp upstream and downstream of the *iglA/B* gene were PCR amplified from *F. novicida* U112 genomic DNA using *Taq* DNA polymerase and primers Mt-*iglA/B*-up/down containing *SalI* and *XbaI* restriction sites, respectively (Table 6-3). The PCR fragments obtained were cloned into pCR2.1 vector by the TA cloning technique (Invitrogen, UK). The insertion was confirmed by sequencing using the vector primers (M13 pairs) and extracted plasmid DNA.

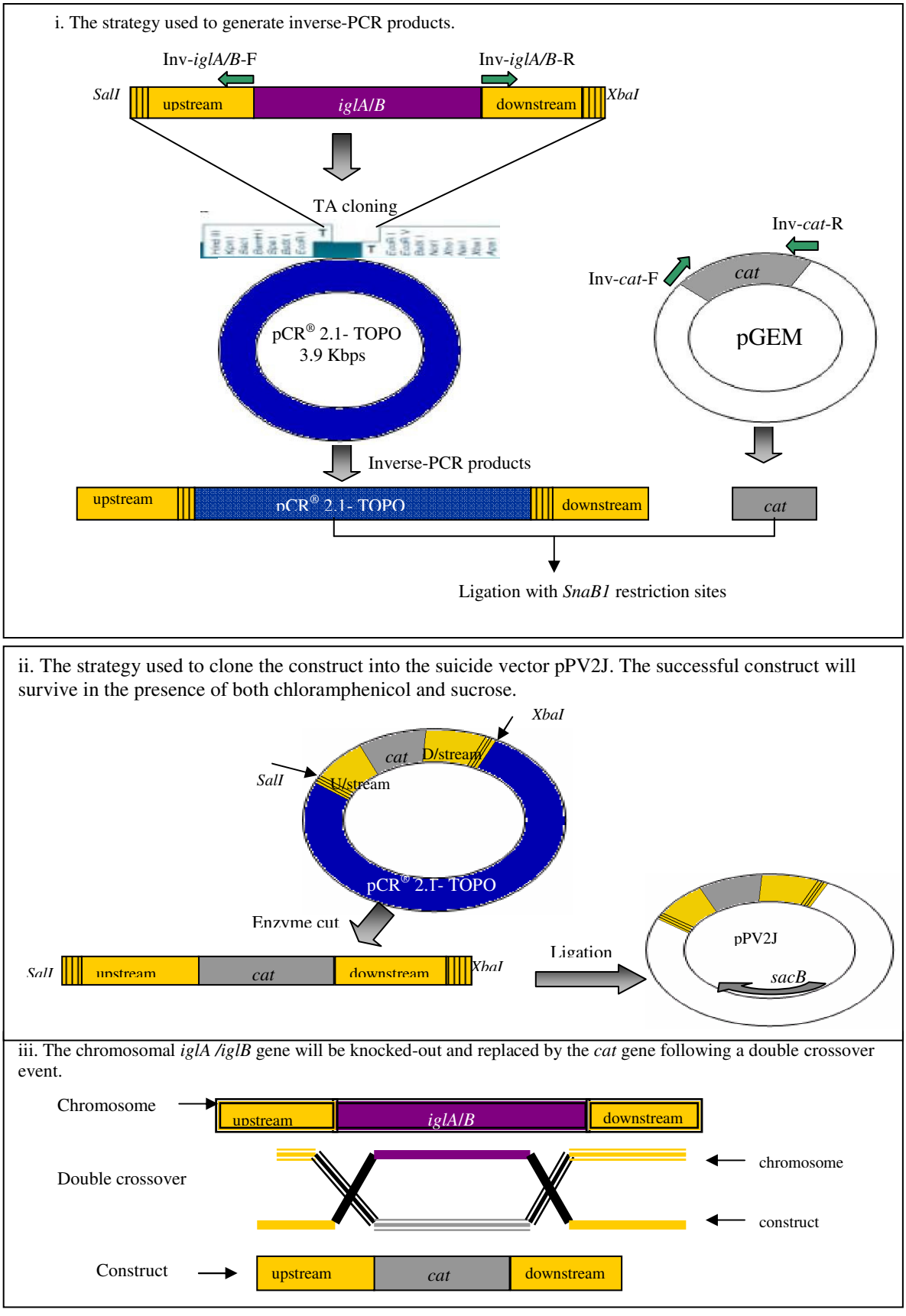

Figure 6-3 The strategy used to make the *iglA/B* constructs.

Inverse PCR primers were designed to incorporate a *SnaB1* restriction site at both ends of the PCR product (Inv-*iglA*-F/R and Inv-*iglB*-F/R, Table 6-3). Amplification of the *cat* gene from the pGEM vector was performed using the primer pair Inv-*cat*-F/R (Table 6-3), which also incorporated *SnaB1* restriction sites at both ends of the PCR product. Once the two products are ligated, a new circular plasmid was constructed, containing pCR2.1, the upstream and downstream regions of *iglA/B*, and the *cat* gene (Figure 6-3 ii). After enzyme treatment to remove the pCR2.1 vector, the constructs were cloned into the suicide vector (pPV2J) at the appropriate restriction sites (*SalI*/*XbaI*).

The vector pPV2J carries a *sacB* selectable marker, which is toxic for Gram-negative bacteria in the presence of sucrose (Golovliov *et al.*, 2003b). The desired construct will survive in the presence of both chloramphenicol (selects for the *cat* gene) and sucrose (selects for the *sacB* gene). Following a double crossover event, the chromosomal *iglA/iglB* gene will be replaced by the *cat* gene (Figure 6-3 iii).

6.3 Results

6.3.1 IglA interacts with IglB but only in one direction

The interaction between the proteins IglA and IglB was examined using the yeast two-hybrid system. The genes *iglA* and *iglB* were cloned into pGBT9 (bait) and pGAD424 (prey) yeast matchmaker vectors. The following pairs of the fusion plasmids were co-transformed into yeast strain PJ69-4A:

- test 1: pGAD424-IglA and pGBT9-IglB
- test 2: pGAD424-IglB and pGBT9-IglA
- positive control: pGAD424-EspD and pGBT9-CesD
- negative control 1: pGAD424-CesD and pGBT9-EspD
- negative control 2: pGAD424-IglA and pGBT9
- negative control 3: pGAD424 and pGBT9-IglB

The negative controls 2 and 3 aim to exclude the possibility that any detected interaction occurred between vectors rather than the fused proteins.

Co-transformants were firstly selected on MUHA plates, which lack leucine and tryptophan. The "prey" vector pGAD424 carries a gene *LEU2,* encoding a protein that is required for the synthesis of leucine, while the "bait" vector pGBT9 carries a gene *TRP1,* encoding a protein that is required for the synthesis of tryptophan. After three days, the colonies from the MUHA plates were replicated onto MU plates, which lack tryptophan, leucine, adenine and histidine. The omission of two of the amino acids, adenine and histidine, results in the activation of the reporter genes *ADE2* and *HIS3* once the two fusion

proteins interact. Only co-transformants in which protein-protein interaction occurs can grow on MU plates. It took four to seven days for colonies to grow on MU plates.

Both the positive and negative controls gave the expected results. Protein-protein interaction between IglA and IglB was indicated by the growth of the co-transformant of pGAD424-IglA and pGBT9-IglB on MU plates. In this experiment IglA was the activating domain ("prey") while IglB was the binding domain ("bait"). However, the reciprocal experiment, in which pGAD424-IglB and pGBT9-IglA were co-transformed, failed to show growth on MU plates.

6.3.2 IglA and IglB were expressed *in vivo* in *E. coli* cells

A previous study predicted the molecular weights of IglA and IglB to be 20 kDa and 55 kDa, respectively (Golovliov *et al*., 1997; Gray *et al*., 2002). The sequences of the *iglA* and *iglB* genes are 591 bp and 1,545 bp in length, respectively. In this study, the *iglA* gene was cloned into a pMal-c2 vector to express an N-terminal MBP fusion IglA, while *iglB* was cloned into a Gateway expression vector, pDEST17, for expression of an N-terminal His_6-tagged fusion IglB.

The constructs were confirmed by DNA sequencing. The confirmed fusion plasmids were then transformed into *E. coli* BL21 cells. Upon induction with IPTG, the MBP-IglA fusion protein was visualised as a strong band with a molecular weight of ~60 kDa on a Coomassie blue-stained SDS-PAGE gel (Figure 6-4A). Expression was subsequently confirmed by a western blot, using an anti-MBP antibody (Figure 6-4B). The molecular mass of MBP is about 40 kDa, therefore, the predicted molecular weight of IglA is about

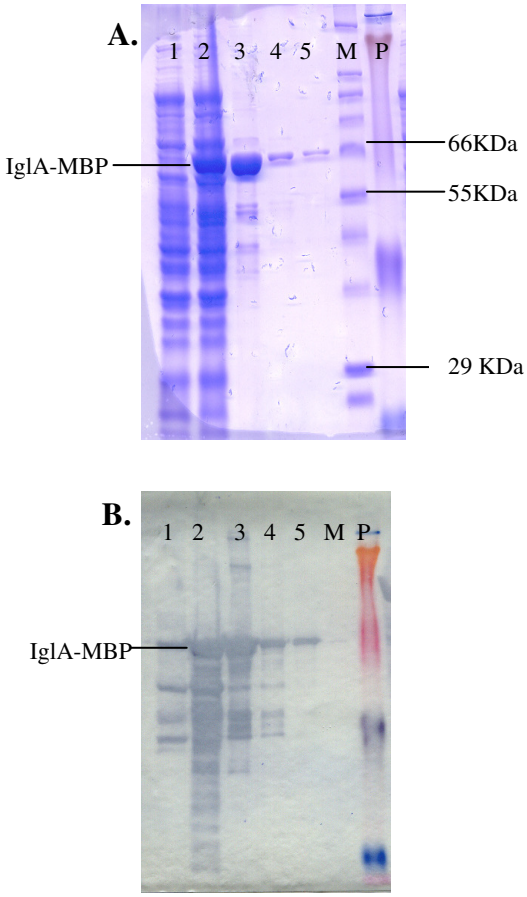

Figure 6-4 A and B. SDS-PAGE gel and Western blot illustrating the over-expression of IglA-MBP fusion protein. Lane 1: uninduced cells. Lane 2: induced cells. Lanes 3-5: three fractions of purified IglA-MBP protein eluted from the amylose column with maltose. Lane M: molecular weight marker (Invitrogen, UK) with 29, 55, 66 kDa marked. Lane P: prestained protein marker (New England Biolabs).

20 KDa in this study. The purified fusion protein was eluted from an amylose column with maltose following the manufacturer's instructions (New England Biolabs). About 3 mg of purified MBP-IglA was recovered after elution, and stored at -20^{o}C.

His_6-tag fused IglB protein was visualised as a large amount of product with a molecular weight of ~55 kDa on an SDS-PAGE gel (Figure 6-5A). The result was confirmed by a western blot using rabbit antiserum directly against the His_6-tag (Figure 6-5B). The His_6-tagged IglB was purified using a Ni-NTA affinity chromatography column following the manufacturer's instructions (Qiagen, UK). However, purification was only possible using 8M urea under denaturing conditions. This suggested that IglB might be insoluble and exist in inclusion bodies. More than 3 mg of purified IglB fusion protein was obtained, and stored at -20^{o}C.

6.3.3 Antibody production using the purified proteins

Polyclonal rabbit anti-IglA and anti-IglB antisera were obtained from Eurogentec (http://www.eurogentec.com). These were produced by immunisation with the purified fusion proteins MBP-IglA and His_6 tag-IglB. To test the activities of each antibody, anti-IglA and anti-IglB were used as the second antibody in western-blot assays. IglA protein can be detected by the antiserum at a maximal dilution of 1:1250 (Figure 6-6). Protein dimers of IglA were detected, suggesting that IglA might be hydrophobic and prone to unfolding. Under this circumstance, the protein could aggregate even in the presence of the detergent SDS.

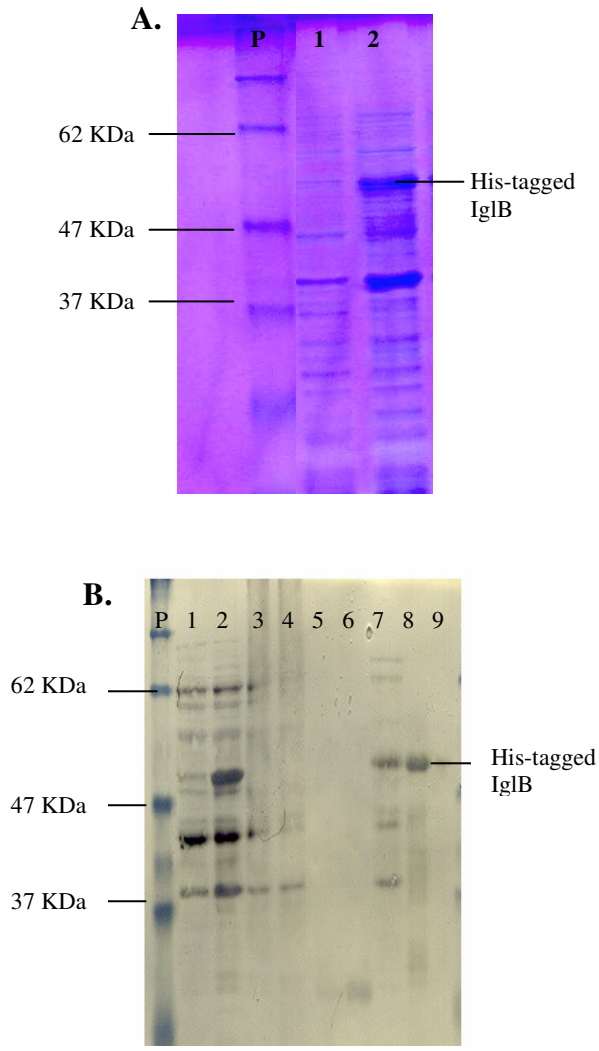

Figure 6-5 A. SDS-PAGE gel showing the over-expression of His6-tagged IglB. Lane P, prestained protein marker (New England Biolabs) with 37, 47, and 62 kDa marked. Lane 1, uninduced cells. Lane 2, induced cells. B. Western-blot illustrating the overexpression of His_6-tagged IglB. Lane P, prestained protein marker (New England Biolabs) with 37, 47, 62 kDa marked. Lane 1, uninduced cells; lane 2, induced cells; lane 3, supernatant from cleared cell lysate; lane 4, flow-through; lanes 5-6, purified protein fractions under native conditions; lane 7 cells resolved in urea; lanes 8-9, purified protein fractions under denaturing conditions.

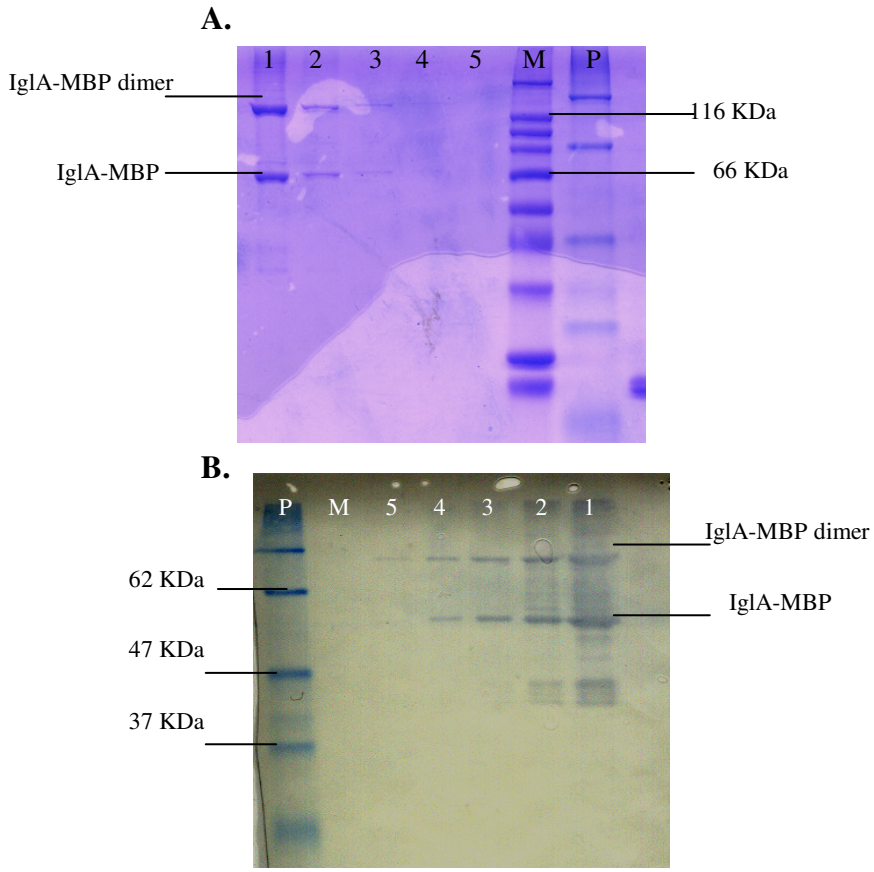

Figure 6-6 A. and B. SDS-PAGE and western-blot of purified MBP fusion protein IglA. The antibody used in the western blot was the polyclonal rabbit antiserum produced against IglA. Lane P: prestained protein marker (New England Biolabs) with 37, 47, 62 kDa marked. Lane M: molecular weight marker (Invitrogen, UK) with 66 and 116 kDa marked. Lanes 1-5 represent dilutions of purified MBP fusion protein IglA as 2x, 10x, 50x, 250x and 1250x, respectively. A dimer of MBP fusion protein IglA was observed in both SDS-PAGE and the western-blot as a band twice the molecular mass of IglA-MBP.

IglB protein was not detected by a western-blot using the anti-IglB antisera. When testing the stock IglB protein by SDS-PAGE, protein degradation was apparent (data not shown). Therefore, a fresh culture of His_6-tagged IglB was prepared from the cell lysate of IPTG-induced *E. coli* BL2l cells. The induced fusion IglB protein was detected by SDS-PAGE, but not identified by a western-blot at this stage, suggesting that the titre for the anti-IglB product is very low. It is likely that the stock IglB protein was degraded during transportation prior to antibody production.

6.3.4 Construction of IglA and IglB mutants

Two large inverse PCR products (about 6 kb in size) were obtained in this study (Figure 6-7), consisting of the sequence of pCR2.1 vector and the sequence of *iglA/iglB* upstream and downstream regions. The obtained inverse PCR product was ligated with the *cat* gene using *SnaB1* restriction sites. The constructed plasmid was transformed into *E. coli* Top 10 cells and grown on LB agar plates in the presence of chloramphenicol to select for the *cat* gene.

One construct was identified in which the *iglB* gene was replaced by the *cat* gene. The derived construct showed 76 bp of the sequence of pCR2.1 vector, ~200 bp of the downstream *iglC* gene, and ~600 bp of the *cat* gene (Table 6-4). However, the construct was not in-frame. *iglB* was not truncated completely from the junction of *iglB* and *iglC*, but partially, with the deletion starting in the middle of the gene, 308 bp from the start codon of *iglC*. Unfortunately, no *iglA* mutants were confirmed by sequencing, and further efforts to obtain this construct are required. However, a method for creating *Francisella* gene fusions was demonstrated here.

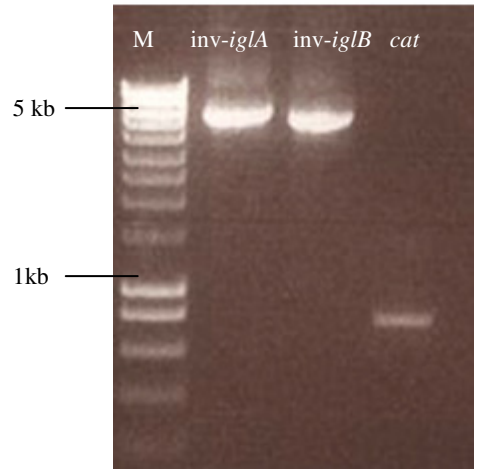

Figure 6-7 Gel image illustrating the inverse-PCR products and the *cat* gene. M represents the 1 Kb DNA marker (Hyperladder I, Bioline) with 1kb and 5kb marked. The inverse-PCR products inv-*iglA* and inv-*iglB* are ~6 kb in size and the *cat* gene is ~800 bp in size.

Table 6-4 Sequence of a construct in which the *iglB* gene was replaced by the *cat* gene. The first 76 bp of sequence (shown in blue) correspond to the pCR2.1 plasmid, the next ~200 bp (shown in orange) is the downstream *iglC* gene, followed by ~ 600 bp of the *cat* gene (shown in grey).

```
1       GATCGCCAGC TTGGTACCGA GCTCGGATCC NNNAGTAACG GCCGCCAGTG TGCTGGAATT
61      CGGCTTTCTT GTCGAGCACC CATAAGTTCT GTTGGCTCTA TACTAATACT AAAAGCCTTA
121     GCACTTATTG GATATAACTC TAAATTAGAT AGATCTATCA TAATACCCCA TGCTTCATCA
181     GTTTTGTACT CTTGTTTTTT ATTAGAATTA CCTAACTTAA TTTTCATATC TGTAGCACTT
241     GCTTGTAATA TGCTCGAAAC TTTCTCTTCA AGTAGTAAGA GGTTCCAACT TTCACCATAA
301     TGAAATAAGA TCACTACCGG GCGTATTTTT TGAGTTATCG AGATTTTCAG GAGCTAAGGA
361     AGCTAAAATG GAGAAAAAAA TCACTGGATA TACCACCGTT GATATATCCC AATGGCATCG
421     TAAAGAACAT TTTGAGGCAT TTCAGTCAGT TGCTCAATGT ACCTATAACC AGACCGTTCA
481     GCTGGATATT ACGGCCTTTT TAAAGACCGT AAAGAAAAAT AAGCACAAGT TTTATCCGGC
541     CTTTATTCAC ATTCTTGCCC GCCTGATGAA TGCTCATCCG GAATTCCGTA TGGCAATGAA
601     AGACGGTGAG CTGGTGATAT GGGATAGTGT TCACCCTTGT TACACCGTTT TCCATGAGCA
661     AACTGAAACG TTTTCATCGC TCTGGAGTGA ATACCACGAC GATTTCCGGC AGTTTCTACA
721     CATATATTCG CAAGATGTGG CGTGTTACGG TGGAAAACCT GGCCTATTTC CCTAAAGGGT
781     TTATTGAGAA TATGTTTTTC GCCTCAGCCA ATCCCTGGGT GAGTTTCACC AGTTTTGATT
841     TAAACGTGGC CAATATGGGA CAACTTCTTC GCCCCCGTTT TCACCATGGG CAAATATTAT
901     ACGCAAGGCG ACAAGGTGCT GATGCCGCTG GCGATCAGGT TCATCATGCC GTTTGTGATG
961     GCTTCCTGTC GGCAGAAGGT T
```

6.4 Discussion

The intracellular lifestyle of *F. tularensis* is well characterised (Oyston *et al.*, 2004). With the completion of the genome sequence of *F. tularensis* strain Schu S4 (Larsson *et al.*, 2005), many potential virulence-associated genes have been identified that are possible vaccine candidates. In particular, there is evidence that the gene locus *iglABCD* within the *Francisella* pathogenicity island (Larsson *et al.*, 2005) is associated with virulence (Gray *et al.*, 2002; Nano *et al.*, 2004). Within this gene cluster, the conserved gene pair, *iglA* and *iglB*, shows homology to components of a T6SS (Nano and Schmerk, 2007). In this project, the proteins encoded by the *iglA* and *iglB* genes were investigated.

In order to understand the structure and function of each protein, it is worthwhile determining which proteins interact with each other. This is important in revealing the relevant biological pathway among the extensive networks of protein partners. In this study, a protein-protein interaction between IglA and IglB was identified using the yeast two-hybrid assay. The interaction was found in only one orientation, when IglA was the "prey" and IglB was the "bait", suggesting that the C-terminus of IglA could physically associate with the N-terminus of IglB, but not the reverse. However, the experiments reported here do not provide evidence about the strength of the interaction, since a subsequent β-galactosidase assay was not successful. The yeast two-hybrid control reactions were set up with two proteins, EspD and CesD, from an EPEC strain. EspD is a translocator protein, whose secretion is dependent on a T3SS, whereas CesD functions as a chaperone to assist EspD secretion in EPEC (Wainwright and Kaper, 1998). Thus, the interaction between CesD and EspD is essential for the secretion of EspD, as well as of another effector protein, EspB. Interestingly, the interaction between CesD and EspD could also occur in one

orientation only, when the N-terminus of CesD interacts with the C-terminus of EspD (Wainwright and Kaper, 1998).

In addition to the yeast two-hybrid system, the interaction could be discovered or confirmed using pull-down affinity chromatography. In this study, IglA and IglB have been expressed as fusion proteins with MBP and His_6-tag, respectively. The His-tagged IglB (the bait) can be used to capture its protein-binding partner, MBP-IglA (the prey) in this case. In other words, the two fusion proteins will be added to the same affinity column sequentially: first, allow His-tagged IglB to bind to the Ni-NTA resin, which generates a "secondary affinity support" for capturing other proteins that interact with it; second, the prey protein, MBP-IglA, is added. If IglA physically interacts with IglB, the proteins will bind to each other and form a large protein complex. Once the complex has been eluted from the affinity column, it can be identified by SDS-PAGE according to its molecular mass. To correctly interpret the results, it is important to set up control experiments. In this case, a negative control (IglB- and IglA+) could be designed to examine a non-treated affinity support, while a positive control (IglB+ and IglA-) could be used to identify whether the affinity support is functional. However, the problem with the pull-down technique lies in the choice of protein candidates. For example, if a protein-protein interaction is transient, any interaction might be temporary and as proteins tend to disassemble over time, it is relatively difficult to capture the protein interaction at the right time.

In a recent study, the interaction between IglA and IglB from *F. novicida* was demonstrated by performing immunoprecipitation (de Bruin *et al.*, 2007). Using anti-IglA

antibody, a co-precipitating protein with a molecular weight of about 60 KDa was detected on an SDS-PAGE gel. This protein was found to be absent in control reactions with non-specific anti-IglA antibody and in immunoprecipitations with an *iglA* null mutant (de Bruin *et al.*, 2007). The co-precipitated protein was excised from the gel and subjected to sequencing analysis, which suggested it was IglB from *F. novicida*. In addition, de Bruin *et al.* discovered that knocking out IglB led to the loss of IglA expression, confirming their interaction. Besides the interaction between IglA and IglB, complex protein interaction or co-secretion among the four T6SS-encoded proteins IglA, IglB, PdpB and DotU might exist, but as yet it is not clear. Two FPI-encoded virulence proteins, PdpD and IglC, have been shown to require those four proteins for their surface localisation (Ludu *et al*., 2008). On the other hand, overexpression of PdpD affects the cellular distribution of several proteins, including IglA, IglB, and IglC (Ludu *et al.*, 2008).

It is possible to suggest roles for IglA and IglB based on the interaction between them that has been identified in this study. It is likely that IglA and IglB act as chaperones involved in protein secretion, and that their interaction helps to secrete *Francisella* virulence proteins, eg, PdpD (Ludu *et al.*, 2008). Alternatively, IglA and IglB possibly function in a complex signaling mechanism, where their interaction might be involved in signal mediation or a two-component regulatory system. An IglB homologue, VCA0108 from *V. cholerae*, was found to belong to the quorum-sensing regulatory protein family (Shrivastava and Mande, 2008). This family also includes staphylococcal accessory gene regulator AgrB, a transmembrane protein. The staphylococcal quorum-sensing regulatory system can upregulate the expression of many secreted virulence factors, while AgrB appears to be involved in processing signals to other protein products (Yarwood and

Schlievert, 2003). However, the exact roles of both the IglA and IglB proteins remain to be elucidated.

In this study, polyclonal antibodies to IglA and IglB were raised in rabbits. Western blots and immunoprecipitation are two main applications of antibodies. In western blots, the antigen-specific antibody can be used as primary antibody to form a primary antibody-antigen complex. This complex is then identified by incubating the blot with a secondary antibody against the primary antibody, which is often conjugated to an enzyme (i.e. HRP) to facilitate detection. In immunoprecipitation, the protein antigen is precipitated out of the solution using an antibody that specifically binds to that antigen. These methods are useful to establish whether a protein antigen exists in a prepared solution, i.e. in cell lysate, or sub-cellular fractions (membrane, periplasmic or cytoplasmic fractions). Therefore, the recognition of a protein antigen by an antibody also helps to identify the location of that particular protein in bacterial cells. Alternatively, the enzyme-linked immunosorbent assay (ELISA) can be used to analyze soluble protein antigens (Hornbeck *et al.*, 2001). In ELISA, as well as the presence of antigen, the concentration of antibody used to detect the antigen, can be identified.

One aim of this project was to construct *iglA* and *iglB* knock-out mutants. If expressed *iglA* or *iglB* gene product contributes to *Francisella* virulence, the hypothesis would be that a null *iglA* or *iglB* mutant would lead to a degree of loss of virulence-related phenotypes under some experimental conditions.

CHAPTER SEVEN

GENERAL DISCUSSION AND CONCLUSIONS

7.1 PCR-based comparative genomics

The scientific value of genome sequences, combined with the increasing capacity and decreasing cost of sequencing, has led to great emphasis being placed on bacterial genomics over the past decade (Binnewies *et al.*, 2006). The availability of multiple genomes from closely related organisms has led to the field of comparative genomics, a powerful approach to evaluate genomic diversity and the processes of genome evolution. The work presented in this thesis has contributed to the field of genomics both directly, by obtaining sequence data to help the progress of two genome sequencing projects, and indirectly, by developing methods to investigate genomic diversity and identify novel genomic regions without the need for complete genome sequencing.

Many bacterial species have shown an unexpected degree of genomic diversity (Ohnishi *et al.*, 2002; Yang *et al.*, 2005; Stavrum *et al.*, 2008). Although the genomes of most model strains have now been sequenced, they have captured only a fraction of the full extent of bacterial genomic diversity. There are increasing needs to create avenues for assessing genomic diversity and for detecting large-scale chromosomal differences. Currently, many bacterial sequencing projects are underway that will not only help to expand the number of available genomes, but also enable many comparative studies that will link genotype and phenotype at the genomic level (Schuster, 2008). In this thesis, PCR techniques for exploring genomic diversity were developed - Tiling-path PCR, Whole-genome PCR scanning and long Single-Primer PCR - that facilitated the discovery of differences between genomes, and the discovery and analysis of pathogenicity islands. These approaches are simple, relatively cheap, provide quick results, and represent a shortcut to obtaining genomic information from non-sequenced strains. For large scale comparisons,

the benefit is greater, since the same primers can be applied to large numbers of test genomes.

The Tiling-path PCR and deletion-scanning PCR used in this study are principally the same as the previously described systematic PCR scanning (Ohnishi *et al.*, 2002). In that analysis, eight *E. coli* strains from the O157 serotype were compared at the whole genome level by PCR scanning. Similar PCR-based approaches were applied in this project but extended to a wide range of organisms, including the phylogenetically diverse *E. coli* strains, *Campylobacter jejuni* strains and *Francisella tularensis* strains. The results demonstrated that the methods are applicable to different organisms. Furthermore, the studies on ETT2 and Flag-2 showed that the tiling-path PCR is a powerful tool in genetic comparison of a single locus. For instance, by combining with deletion-scanning PCR, the 8.7 kb deletion found in an EPEC2 strain B171-8 was also confirmed in another 31 *E. coli* strains, suggesting they are evolutionally related.

Long single-primer PCR was firstly reported in this project. The previous length limitation of single-primer PCR has been improved under long PCR conditions, including the use of long PCR DNA polymerase. The S-P PCR employed three sets of conditions: the first and third 30 cycles used a stringent annealing temperature, while the second 30 cycles used a low annealing temperature (Karlyshev *et al*., 2000). In this study, the second PCR step was reduced from 30 to 2 cycles. Since the DNA polymerase is prone to become inactive during long PCR cycles, the reduction at this stage is necessary to remain DNA polymerase activity for the next step. In addition, the decreased number of PCR cycles in the second step also reduces the non-specific amplification of primer to template,

producing relatively clean PCR products. The PCR sequencing using a nested primer described in the previous study (Karlyshev *et al*., 2000) was modified to incorporate cloning technique in this study. The advantage was seen in that sequencing can be performed from both ends with the plasmid primers. This is particularly useful for characterising large genomic discrepancies. With these improvements, S-P PCR has expanded its applications in genetic studies. For example, to resolve those regions with an uncertain large insertion or rearrangement following whole-genome PCR scanning, long S-P PCR can compete with other methods, such as the use of a bacterial artificial chromosome (BAC) library.

Comparative genomic hybridisation to microarrays has been used extensively in bacterial genomics (Dorrell *et al.*, 2001; Stabler *et al.*, 2003; Butcher, 2004; Witney *et al.*, 2005; Lindsay *et al.*, 2006). However, unlike microarray-based comparative genomics, the PCR-based methods described here allow the discovery and characterisation of sequences present in a reference genome but absent from the test genome. They also allow the user to detect differences in genome organisation, and to rapidly identify the regions of a test genome that are co-linear with the reference genome. Furthermore, the creation and use of the microarrays relies on specialised equipment and reagents and demands a high capital expenditure, whereas PCR is a routine procedure.

In Chapter Three, the remnants of two *E. coli* Type III secretion gene clusters (ETT2 and Flag-2) were found to be present in most strains, both pathogens and commensals. Using tiling-path PCR, the study of the distribution of entire genomic islands was performed by several long inter-locking PCRs, allowing investigation across the full length of the islands.

Application of the PCR approaches to *E. coli* has added value to the current genome-sequencing programme, allowing quick determination of the extent to which novel findings in genome sequences are typical of the entire species.

The genomic study described in Chapter Four obtained sequence information from a novel *C. jejuni* strain, M1, and compared the genome organisation and content to a reference strain, *C. jejuni*11168. The genomes were shown to be broadly colinear and regions of difference were characterised. This work has provided background information that could be used in the assembly of the complete genome sequence of *C. jejuni* M1.

Chapter Five describes the complete genome sequencing of the European *F. tularensis* subspecies *tularensis* FSC198. The sequencing of this strain, together with resequencing of SNPs and VNTR differences in several additional strains, indicated the most likely source of the reported European isolates of *F. tularensis* subspecies *tularensis* to be the laboratory strain Schu S4. This project highlights the utility of bacterial whole-genome sequencing for the purposes of public health epidemiology. Whole-genome PCR scanning allowed confirmation of the conserved genome organisation and facilitated the finishing of the genome sequence.

There are limitations to the PCR-based methods developed. In all case studies, there remained unresolved PCR results, and it is in the nature of PCR that a reaction can fail for many reasons, for example low quality of primers, the use of an inappropriate DNA polymerase, poorly optimised reaction conditions or the presence of SNPs in the primer binding site. Therefore, a failed PCR cannot be interpreted due to an absence of the target

region from the test genome. The supplementary experiments required to resolve these failed PCRs can sometimes be laborious.

7.2 Next-generation sequencing

More than 30 years after its invention, Sanger DNA sequencing still remains the most widely used technology in research today (Sanger *et al.*, 1977 a and b). However, it remains expensive, labour intensive and time consuming when applied to large-scale sequencing projects. For the last five or six years, the incorporation of Sanger sequencing methods into comparative study, which focused on studying genomic differences among closely related genomes, has helped in the production of a *de novo*, high-quality, finished assembly of a given genome (Brown, 2008).

Now, the emergence of next-generation sequencing technologies is starting to make cost-effective and high-throughput, high-capacity whole-genome sequencing a reality (Marguerat *et al.*, 2008). The latest next-generation sequencing instruments can generate as much data in 24 hours as several hundred Sanger-type DNA capillary sequencers, but are operated by a single person (Schuster, 2008). These instruments enable the generation of millions of sequence reads in parallel, rather than 96 or 384 at a time by conventional capillary-based sequencing. Currently, three commercially available next-generation DNA sequencers are the Roche (454) GS FLX sequencer (http://www.454.com/enabling-technology/the-system.asp), the Illumina/Solexa genome analyzer (http://www.illumina.com/pages.ilmn?ID=203), and the Applied Biosystems SOLiD sequencer (http://marketing.appliedbiosystems.com/images/Product/Solid_Knowledge/flash/102207/solid.html) (Mardis, 2008b). Another two systems were announced: the Helicos Heliscope (www.helicosbio.com) and the Pacific Biosciences SMRT (www. pacificbiosciences.com).

Although these instruments are diverse in both sequencing biochemistry and generation of the cyclic-array, the work flows of next-generation DNA sequencing are similar in principle (Shendure and Ji, 2008). The first is to generate a single-stranded DNA library. Libraries can be constructed by any method that gives rise to randomly fragmented genomic DNA, followed by *in vitro* ligation of common adaptor sequences. The library is therefore established without the prerequisites of cloning and colony picking as in Sanger sequencing, which limit the parallel scale of sequencing. The second is to generate multiple clustered DNA copies/amplicons serving as sequencing templates. To do this, the 454 FLX, SOLiD and Polonator systems rely on emulsion PCR, while the Solexa system relies on bridge PCR. Emulsion PCR is performed by adding the library fragments into an excess of DNA capture beads, exemplified by the Roche 454 FLX system. Each bead captures only a single molecule. The PCR reaction is performed in water-in-oil microreactors, where PCR reagents are present and thermal cycling occurs. As a result, approximately one million copies of a single DNA fragment are captured on the surface of each bead (Mardis, 2008b). After breaking the emulsion, beads bearing clonal amplicons can be selectively enriched prior to the sequencing process. Another approach is called bridge PCR, in which both forward and reverse PCR primers densely coat the surface of a flow cell, and two primers are attached at their 5' ends by a flexible linker (Shendure and Ji, 2008). On the Solexa platform, the flow cell is an 8-channel solid device. The bridge amplification of fragments is allowed on the surface, followed by the addition of nucleotides and enzyme. As a consequence, all amplicons originating from any single template molecule remain clustered and immobilised to the point of origin on an array. Each clonal cluster consists of ~1,000 copies of a single molecule (Shendure and Ji, 2008). Within each of eight independent channels on a flow cell, several million clusters can be

generated to the distinguishable locations, which greatly increase the ability of sequencing in parallel.

After clonal cluster generation, the next step is referred to as sequencing by synthesis. This is enzyme-driven sequencing by either a polymerase or a ligase (Shendure and Ji, 2008). Of the currently available sequencing instruments, the 454, Solexa and HeliScope are driven by DNA polymerase, whereas the SOLiD and Polonator are driven by DNA ligase. For instance, the 454 system uses the pyrosequencing method, in which the amplicon-bearing beads are preincubated with *Bacillus stearothermophilus* (*bst*) DNA polymerase. Other enzymes required for pyrosequencing include ATP sulfurylase, luciferase and apyrase, and the substrates adenosine 5´ phosphosulfate (APS) and luciferin. The enzyme-containing beads can catalyse the downstream pyrosequencing reaction step, which relies on the detection of pyrophosphate (PPi) release on nucleotide incorporation rather than chain termination with dideoxynucleotides (Shendure and Ji, 2008). In brief, each sequencing cycle is initiated by the introduction of a single nucleotide, driven by DNA polymerase incorporating correct, complementary dNTPs into the template. The released PPi will be converted to ATP in the presence of ATP sulfurylase and APS. This is followed by the addition of substrates (luciferin and ATP) to drive light production (Shendure and Ji, 2008). The light is detected by a camera and the 454 base-calling software can calibrate the light emitted. The unincorporated dNTPs and ATP are washed with apyrase and the reaction cycle restarts and keeps going. In this case, however, the calibrated base-calling cannot properly interpret long stretches (> 6) of the same nucleotide (homopolymers) (Mardis, 2008b). Therefore, these areas are prone to base insertion and deletion errors. Other methods, using either polymerase or ligase to generate DNA

sequencing, and subsequent *in situ* interrogation after each cycle by fluorescence scanning or chemiluminescence, have been reviewed (Lizardi, 2008; Mardis, 2008b; Morozova and Marra, 2008; Shendure and Ji, 2008). However, the read lengths produced by next-generation sequencing are very short compared to Sanger sequencing. This shortcoming shows up the difficulty in *de novo* sequence assembly. Currently, the Solexa, SOLiD and HeliScope systems can yield only about 35 bp independent reads, and the Polonator produces even shorter reads of about 13 bp. An average read length of 250 bp produced by the 454 system is still much shorter than the ~1,000 bp achieved by Sanger sequencing.

Although there is some limitation in read length and accuracy, the next-generation DNA sequencing technologies have many advantages relative to Sanger methods. In addition to its high degree of parallelism of sequencing, next-generation DNA sequencing has greatly reduced the cost for sequencing. Sanger sequencing costs $0.5 per kilo base now, while with the massive parallel Polonator and HeliScope systems, the sequencing cost is only $1 per mega base (Shendure and Ji, 2008). Moreover, it is believed that, with the gradual improvement of other informatics technologies, such as computational system, data storage and interpretation, next-generation sequencing will have even greater potential and more applications (Mardis, 2008b). Next-generation DNA sequencing is transforming today's biology (Schuster, 2008). The major impacts of next-generation sequencing on the fields of biology and genomics have already been seen.

The applications of next-generation sequencing are rapidly revolutionising bacterial genomic studies. In the near future, bacterial whole-genome sequencing with the next-generation techniques is likely to be routine and simple. For example, the choice of isolates

for current sequencing is usually biased on clinical phenotype or simply availability, rather than universal sampling from reliable phylogenies. But once sequencing a whole genome is straightforward, the selection of sequencing targets would be broadly expanded. For example, one can imagine a time in the near future when all ECOR strains have been completely sequenced. In such a future, one assumes that genomic studies will be totally different from the methods listed in this thesis, although some of these approaches will still be of use in the finishing stage of a genome project. With direct access to rapid whole-genome sequencing, comparative studies will allow us to (i) align the sequence reads with the reference genome(s); (ii) evaluate them for single nucleotide and/or in-del variants, and, to a wide scale, for copy number variants in large sequence blocks (>1000 bases); (iii) detect the presence of antibiotic resistance genes or pathogenicity islands by comparing novel sequences to the publicly available databases; and (iv) evaluate any discovered variation in a functional and a biological context (Mardis, 2008a). With next-generation sequencing, comparative bacterial genomics is about to enter a new era of discovery. It will expand our view on bacterial diversity, of which nearly 99% is still unknown today (Snyder *et al.*, 2009). Furthermore, next-generation sequencing is likely to be used for resequencing projects, such as for those strains and isolates from the large-scale reference genomes, eg. *C. elegans*, *Drosophila*, and human (Mardis, 2008b). Resequencing studies will help to better characterise and catalogue genomic variations within complex genomes. The current bottlenecks in effectively identifying SNPs, copy number variations and chromosomal inversions may be overcome. Furthermore, the field of "meta-genomics" has been launched where entire biological communities are sequenced, *en masse*, to survey the variety of all the organisms of an ecosystem where they live *(*Marguerat *et al.*, 2008). This

helps to explore the influence of environmental and temporal changes over time on the mix of the microbial species within the community (Snyder *et al*., 2009).

The short read length production of next-generation sequencing has developed new applications in which sequencing of a small fraction of a DNA molecule is required, rather than of the entire chromosomal DNA (Morozova and Marra, 2008). For example, next-generation technologies were applied to transcriptome sequencing, such as mRNA expression profiling and non-coding RNAs discovery (Morozova and Marra, 2008). One application of the transcriptome studies is to annotate the genomes through the identification of expressed genes, which remains a challenge in high-throughput genome projects. Currently, the most commonly used method is based on microarrys. However, several limitations of microarrays remain in that the synthesis of DNA probes does not fit well with large genome size, and in the possibilities of cross-hybridization (Wold and Myers, 2008). The sequencing-based methods could bypass these technical problems of microarrays. The shorter read lengths produced by next-generation DNA sequencing are compatible with the length of small RNA and could certainly be utilised with these approaches.

APPENDIX I

Primers used for whole-genome PCR scanning on the *Francisella tularensis* FSC198 genome (primers were designed by Dr. Chaudhuri). Gene names and nucleotide position numbers of each primer are specified relative to the *F. tularensis* Schu S4 genome.

Primer name*	Sequence of primers(5′-3′)	Gene name	Start position
1F	GCTCTGGCTGAATTAGGACATCAAG	FTT0008	8042
1R	TTGGCGTTTTGTTCAAGGACTCGGT	figD	25517
2F	GCCAAAGATGACAAAAGCGTTACCC	FTT0025c	23789
2R	TCGGACCCTCCATTGATCCTTCTAA	nuoL	41339
3F	GCTATGAGAATGGCGAAGAGAGATG	nuoI	39298
3R	GACCTACACCGCCCCACATATATAA	FTT0054	56449
4F	AGGGTTGGGATTATGGTTGCCTTTC	FTT0053	55817
4R	TATGGCGAACTCCCAACAGATGAAC	gltA	73283
5F	ATCCACCCTGATGTTCTTGCTAGAG	gltA	72388
5R	GCTCTCTAAACCTTCGCCATAGTAC	acnA	89650
6F	CGCACGCACCAAGGAAAATGGATAT	hemN	87279
6R	TGCTCCACCTCCTCCTTTTAAGAAC	FTT0101	104836
7F	CCACAACCAACTCCTTGCTTTGAAC	isftu1	102826
7R	TCCGATGAGGATAGCCTACCTGTTA	polA	120243
8F	GTTGCTAAGTACGCTGCTGAAGATG	polA	119708
8R	CGCATACCTCCAGAAAACTCATGTG	oppD	137110
9F	TGGTGGTTACTGGTCTATCCATCTG	oppC	136521
9R	GAGCATCAGCAAGACACTCAAACTC	rplJ	154035
10F	GAGCTTCAGTTCTGCCTAATGGTAC	rplA	153126
10R	TGGCACTATACGCTCTTTTGAACCC	xerD	170606
11F	TGGTAGATACGAAGGGGTTGATGAG	trmD tRNA	169146
11R	GCTTATTGTCACACCCCATCTCATC	isftu1	186436
12F	AGCTTGATGAGGTCATGGAAGGTAG	prfA	184561
12R	GCTGCTGACTTAGGCGATAGACTAT	ddlB	201601
13F	GGTCTGCATTGCTTGATACCAAGAG	rpsA	199105
13R	CGCACCCGCCACATAAGCATATTTA	blc	216684
14F	TGAGGCACAGGATCTGAACCATCAA	glnA	215054
14R	CCTGCTGTGGTTTAACCTCTTGTAG	priA	232301
15F	GGGACCACTACTGATGGGATTTGTA	rbn	230627
15R	CTGGAGCATTTACAGGTGCGACAAT	yidC	247996
16F	TGGGAACGATGCGAACAGGAATGTA	yidC	247329
16R	TCCATCATCCCAGGCATAGAAACTC	ppdK	264374
17F	GGCAAATCTTACTGGCTGCTCTAGT	feoB	263189
17R	CGCCAGAACAGTTTGTTAGAGGCTA	FTT0265	280504
18F	CAGCAAAGGTTTCTCGCACAAAGTG	FTT0265	278779
18R	GCGATACTCAGCATTAGCACCATCA	cyoA	295904
19F	AGCACACCAGCGATACCAAAGTGAT	yajR	294223
19R	TACCCAAGAGCATCACAGAGACTAG	FTT0297	311486
20F	GGCGTGATGTTGCTATAGGGCTTTT	FTT0295	310205
20R	CCAATGTCAGCTCAAGTCTCAGATG	FTF0311c	327685
21F	CAGCGTAGATGGACTTGCTTTTGTG	FTT0311c	326958
21R	GCCTTCTTCAAACCAGGAACATCAC	rpsN	344009
22F	TGGCGAGCTTGTAGACTTATAGGAC	rplX	343123
22R	ACCAGCCTGACCATATCCAGCATTA	FTT0360	360295
23F	CCTAAACCTGATGTCAGTCGTACTG	FTT0359	359734
23R	GGCGACATAATTGGCATGGTAAGAG	FTT0375	377023
24F	TGCCGAGCTTCCAATCTTATCAGGA	FTT0375	376523
24R	GCTCTGCGGTTGTAACTCCTTCTTT	map	394044

25F	GCAAAGGACCCTGAAATATGGACTC	FTT0389	392390
25R	CCCATTCATAACCGACATCACTGTG	lolC	409628
26F	GGGGAAGAACAACGCTTGATTTAGC	Intergenic region	409083
26R	CCCAACCACGTTGTAAATGGCGTAA	glgC	426551
27F	CCACTTGAGGATTTTGGAGGGTTTC	pgm	425363
27R	CGGCACAGCATTATCACCGTGAAAT	thrB	442670
28F	AGCAGTTGGTGAGGGTATGAAGTCT	thrA	440970
28R	CGCCCGTTTGTGAAACTGTATAGCT	Intergenic region	458384
29F	GCATTGCCAAAAGCCCAAGTGTAAG	FTT0442c	456143
29R	AGGCTACGAGGAAGACCTACCAATA	FTT0455c	473301
30F	GCGAAGGCTACTCTAATGTCTATGG	yfdH	471987
30R	GCTGCCCAATGCCCAAAGATTATCT	apaH	489359
31F	GCGAAAGATTGATCCTAGCACACTG	ksgA	488560
31R	ACACCATCCTCAACTTGCGAGAGAA	mutL	506147
32F	AACAAGTTGGTGATGCCCATCTGTG	mutL	505559
32R	ATTGGTCCTAACTGTCCAGGGATCA	sucD	523018
33F	CGATGAGGAGGAGTATGCTAGATTG	FTT0502c	522314
33R	ACCTTTATCGCCAGCAAATCCACAG	FTT0516	539728
34F	TGTGCGTAGAGCTGTATTGCCTTGT	FTT0516	538229
34R	AGGAACAGGACGTAGAAATGCTCAC	nrdA	555298
35F	GGCGAAGGACAAGTTAAGCCATTTG	FTT0531	552842
35R	AGCCCAAGTTCGCCAGCTAATTTAG	FTT0551	570000
36F	GAGCTAAAACAAGCAGCGACTGATG	vanY	567658
36R	GCTGGATGCTTAGCACTGACAGTAA	FTT0567c	584734
37F	CCTGAGATAAACGCCTTGGCTACAA	potI	582260
37R	CACCTGTTGCTTGACCTGCTAGATT	fopA1	599790
38F	GGGACCAGTGCTTATGACAGATGTT	FTT0580	598318
38R	AAAACCAACTCGCTCCCACTCTGTT	FTT0597	615427
39F	AACGCTGTTGTACCTGTACCAGATG	FTT0594c	613214
39R	CCTAAGTGATAGTCCTGCCAGTAAC	FTT0611c	630751
40F	CAACCTCAAGCAGTATCCACCAGTT	FTT0610	629491
40R	ACCTCACCTGTCATTGCCACATCAT	lon	646774
41F	TCGAGTGGCGGTGATGTTGTAGAAA	lon	644814
41R	AGTCCCAGAACCTTTTGGTGCTACT	ilvC	662329
42F	GGTGCAAAGATGGGCGTCAAAGATA	ilvB	661101
42R	ACCTCTGCTACTGTAGTTGCATCAC	ruvA	678177
43F	AGGTCCAGCACATGGTTCAGTCTTT	ruvC	676125
43R	GCTTCTCTTTGCGGCACGATGATTT	FTT0676	693607
44F	GTTGGCAGATACATTGGAGCACAAG	FTT0673c	691693
44R	ACCTCATCCGCCCCTCTTTTGATAA	tyrS	709098
45F	TGTCCTAATGTGGCTTGACCATCAC	hslV	706999
45R	AGCATCACGAGCTAAAGCAGCAGTA	rumA	724344
46F	TACCACCACTACGCACGAGTAATAC	rumA	723470
46R	GGTCACCCTCCTTTTGGTCATAATG	dgt	740793
47F	CCACTTGGTGCTTGGGGATTTAGTT	FTT0719	738744
47R	GCCTGTTTGACAGCCACCATATTAC	FTT0734	756123
48F	GCCAGGTAGTGACGAAGAAATCAAG	FTT0733	755363
48R	ATTCGCTCTGGTAGCACTACAGGTA	FTT0750	772884
49F	CCAAACCAGTCAACACAAGCCTGTA	FTT0749c	771482
49R	ACGGCTTCTTTGAACTTGCTGCTCT	secA	788892
50F	TGGAGACAGCGGCTTTATTTGCTAC	deoD	786508
50R	TTGCCTTCACCTTGAGCTATTGGAC	FTT0784	803592
51F	GGAGCCGAGCTTGCTTTTACTTATG	fabI	801646
51R	CGCATCATCCCAAGCATAGTTGAAC	FTT0800	818870
52F	TCAGCAAAACTCTGGGCAGGCATTA	FTT0798	816847
52R	TGGCAGAATGGGACGTTGATGATAC	FTT0816c	834131
53F	CTGCTCCAGCAAGTTCACTTGAGAT	FTT0814c	833218
53R	TACCCTCAAACTCCTCAACCTCTGT	ispH	850725
54F	CGTGCTGTTGAGACTGTAGAGAAAG	ispH	849888
54R	GGATTACCAAGAGGACTCATCACAC	FTT0853	866965
55F	AGCTGGTGGATATGGTGGCTATGAT	FTT0852	866157

55R	AGAGCAAGTAGGCTGTAAGGCAAGA	FTT0874c	883583
56F	CCGACCTTTACAAGAGGCTGAACTA	FTT0872c	881575
56R	TGCTCTAGGTGGTGCTAAAGCTACA	FTT0890c	898868
57F	GTACGTTGTTGCCATGCGTTTAGAG	FTT0887c	896725
57R	CGCAAAGATGTTGTGCAAGAAGAGC	topA	914267
58F	GGATGGTGAGGTTGTTTCCGAAGAT	lpnB	911817
58R	CCAGATGCTGTTTGACCGAAAGTTG	FTT0918	929197
59F	AGTGGAGCCTCTTCAACAACACAGA	FTT0918	927828
59R	TCTCCCTCCGTTTCTGACCATTCTA	FTT0932	944899
60F	AGACCAGTTGGTTTAAGAGGCTCTG	FTT0930c	942503
60R	CCCTGTAATGATTGCGATGGCAGTA	FTT0947c	959730
61F	GGCGGTATCACGGTATATTCGGATT	FTT0945	957478
61R	TATGGGACTACCGTGTGCAACTGAA	aroG	975023
62F	AAAGCAATAGAGGGTGGCTCGTATG	FTT0960	972681
62R	AAGGCTAATGTGGTCGATGGCAGAA	FTT0979c	989977
63F	GGTCCATGTAAGACACTAGCACCTA	trxA2	987951
63R	CAACACGCCAAGCATAGTCATCACT	FTT0995	1005266
64F	GACCACGCCAAATTACTGCTGATTC	FTT0994c	1004313
64R	CGCTCACCGACAGGATTATGATAAC	FTT1009	1021526
65F	GGCTCAACCTTACCTGCTGTTACTA	FTT1006	1019488
65R	GGAACTGGTGTAAAGGGAAACTCTG	FTT1026c	1036851
66F	CGCTCACCACCAGATAAACTCATAC	yhbG	1035444
66R	TTCCCCAGCGATCATAAGAGTCAAG	FTT1040	1052555
67F	TTGGCTCAACTAAAGCAGGTACGGA	dacB1	1051965
67R	GCTGCTGAAACGACTTACTCTCTAG	dnaB	1069496
68F	CGTGCCTGGATATGGATTAAATGGG	FTT1058c	1067181
68R	TGACGGTATCCTTTCAAGGGCATAG	FTT1072	1084480
69F	GTGGAACTCACAGGCAAAACCTTGA	Intergenic region	1083474
69R	CGGATGAATGTCGATATCAGGAGAG	FTT1091	1100632
70F	CAGCTATTGTTACAGCACCCCTTAG	Intergenic region	1098786
70R	TGCGGTACACCACTATTTAGAGCTG	msrA1	1116275
71F	TGGAGGTTTGACTGGTGGCTTATGT	bcr1	1114992
71R	CTGCGTTATTGGAGGGTTGTGGTAA	FTT1122c	1132514
72F	CCCTCTTCTGTTAGCTTACGCATCT	tgt	1130857
72R	TCGCATCACCCATTTGAAACATCGG	pilC	1148237
73F	AGGCGATGCAATAATAGGTAGTGGC	pilB	1145765
73R	ACCGCAGATGTGATACGTTCAGCTT	putA	1163354
74F	AGCTGTAGTCAGAACACCATCACTC	FTT1149c	1161201
74R	GATGGGTAAATGGACAGAGCTAGAG	era	1178521
75F	CCCCATTACGTGCTTTCCATAGAGT	FTT1162c	1176283
75R	TAGGCACCGTAATCAAGAGCAGATG	FTT1177c	1193637
76F	GCGTGGAGTCTTTGTATCATCTGAG	hsdR1	1191774
76R	TGGCAAAGACAGATGATGGCACCAT	FTT1192c	1209250
77F	AATCGCCCTTTCCTGCTGCTGTTAT	xseA	1206834
77R	ACGAGGAGGCGTACCAGTTTTAAGT	gidA	1224129
78F	AGCTACACAAGTAACCCGCACTAAG	FTT1203c	1221836
78R	CGCTAGTAAAACCAACCATCGGTAC	FTT1220	1239345
79F	GCAACCTTATCCCACAAACCAAGTG	ubiH	1237831
79R	TGTTTGGGTGTTGGCGTGATAAGTG	lpcC	1254976
80F	CCGTTTCTGGTAGAACTATGGAGTG	FTT1234	1253053
80R	CGAGCCAACATTAAGCCCAGAGATA	FTT1249	1270251
81F	GTCGCAAACAACTTCCCAACTAGAG	FTT1247	1268450
81R	TGGTGCTCAAGGATGCGTGGAAATA	FTT1263c	1285485
82F	GCTCTGCTTGACCAACCCATATCAT	FTT1261c	1283223
82R	TAGCAGACACTTACCCTGATGTGAG	ptsN	1300768
83F	TCGACACCTTCAAGAGCCTCTTCTA	yfhQ	1300185
83R	GCTTGGTGTTGGTGGGTTAATCAGA	FTT1292c	1317606
84F	TGGCGTCACTATCTTCTGCTTAGGA	FTT1291	1317045
84R	GCTGGTGTTGATGAAGCCAAAGAAG	hflB	1334418
85F	GTTTTACTGACTTGCCACCACTACC	hflB	1333275
85R	GTTGCTAGAGCTAAAGAGGGGTTAC	FTT1322	1350325

86F	CGTGCCTCTTAGTATTGGTTCAACG	FTT1320	1347901
86R	GCCCATAAGTATTGCTGACCATGCA	cydC	1365491
87F	TCTCGTGGCTAGGACAAAACTCATC	cydD	1364906
87R	CTGCGATTGAGCTTTCTTCTGGCTT	pdpA1	1381969
88F	AATGGCGACACAACCTCTTCTTCTG	pdpA1	1380031
88R	CAGCCATCAGCTATACGTTCGAGAT	FTT1355	1397408
89F	CAGGATTGGTCCAAGATGGAGAGAA	pdpC1	1395354
89R	GCTGGTGGAGCATTCTTAGAGTTTC	pgk	1412849
90F	TAAGCACCATGCGAAGTACCGATAG	fbaB	1410703
90R	TACCAGCAACGATCAAACCAGAGTC	suhB	1427838
91F	CAGGCACTGCTCTTTCTACTACAAC	FTT1380	1425732
91R	CAGCGAAACCCTATCTTAGCCTCAA	FTT1395c	1442962
92F	TGCCCTGTAGCTCACCTTTTCCTAA	recB	1441772
92R	TCGCCCTCAGAATGATGCTGTCAAT	hemD	1458828
93F	TAGCACTTACACTCGCCACTCTACT	FTT1407c	1457593
93R	AGTGGCTTCTTTGGTGCAGGACTTT	FTT1426c	1475096
94F	TCAGCTTTGCTTTCGCCACCGATAA	FTT1424c	1472987
94R	CATCGTGAGGGGAAGGAGTTTTATG	FTT1443c	1490396
95F	CCCGTTTCAACACCTTCTGCTAAAG	rpoA2	1489471
95R	TGACGAGCCCTTTGCTGATTCATCA	wbtH	1506548
96F	CGGAATACCAAACCCCATCTTAGAC	wbtH	1505882
96R	GGCGTGAAGGCACGTCAATTTCTTT	deaD	1523179
97F	GCGTAGTTATCCTTGCCTATGTCAG	gmk	1522280
97R	GCGAGGTTTTGTCACAGGGTATAGT	FTT1486c	1539687
98F	GGAACTGCCAATAAGTCGGTTGTAG	aceE	1538363
98R	CGGCTATAAGATAATCTGCGGAAGC	FTT1501	1555584
99F	CTGGTCCAAATATGGGCGGTAAATC	muts	1554085
99R	CTTTTGTTGTCGTTGTCGCTGCTAC	FTT1517c	1571349
100F	CCACCAACGATCATCAAATGCTCAG	merA	1570378
100R	TAGGGTGGCTATAGGAGCTAATGTG	fadE	1587902
101F	GCCATTATGTTACCCAACCTGCTAC	fadD2	1585742
101R	GCCTCATCGTTGCTGGCAATTTCAT	FTT1539c	1603262
102F	CCATCTTGACGAACCTCATAGACAC	FTT1538c	1601034
102R	TCGGTTTGCCTGGCGATGTAATTTC	lepB	1618302
103F	CCGTTCCATCAAGATGAGGTTTGTC	truB	1616418
103R	CATTTACCGCAGCGTATGCAGATAC	ompH	1633576
104F	GTCCAGCTAGACCAACACCATTTAC	lpxA	1631158
104R	TACGATCAAAGCCTGTGCTGGACTT	Intergenic region	1648303
105F	GCTGTTGTCGGCTTTGGGTTACTAA	FTT1582c	1646976
105R	CCCCACTAAGATATACATACGGTGC	FTT1598	1664422
106F	CTTGGGGTGGTGTATCGTATATGAC	FTT1598	1663535
106R	CCGCACCATCACCTTGCTTGAAATA	cysS	1680731
107F	TGGCGGTTCTGATCTTAGATTCCCA	cysS	1680175
107R	GCGACAAGGGTACTATGGCATTTAG	FTT1633c	1697515
108F	CGATACGACCAGCTTGACCTTGAAT	FTT1632c	1696946
108R	TGCTATGCGGAGCTGGTCTTAATCT	FTT1649	1714253
109F	GCTAGGCAGTTTTGCATACCTGTTC	Intergenic region	1712382
109R	CAGGCATAGTATAGTCAGAGGTAGC	carA	1729871
110F	CGAAAGGCTGAAGGAATGACTGAAC	tmpT	1727435
110R	GTGGTAACCCCAGCTTTTCGTAACT	FTT1676	1744912
111F	GCACCATGATGAAAGCTGCTGAACT	ribA	1742810
111R	CCACAGTCCATATAGCCGCTAATTC	FTT1690	1759918
112F	TGCTACACCAACAGGCATACTAGGT	FTT1688	1757906
112R	CTGCGATTGAGCTTTCTTCTGGCTT	pdpA2	1775313
113F	AATGGCGACACAACCTCTTCTTCTG	pdpA2	1773375
113R	CAGCCATCAGCTATACGTTCGAGAT	FTT1710	1790752
114F	CAGGATTGGTCCAAGATGGAGAGAA	pdpC2	1788698
114R	ATTCGCCAGTGGATTGGAGTAGATG	purF	1806209
115F	GCACACCTTGTGGAGATTTCTCAGT	purL	1805235
115R	AGCCTAATGCTAGAGCTGCTATGAG	fopA2	1822673
116F	GCTATGTAAGCCCTGCTGTGGAATA	FTT1732c	1821283

116R	TTTGGCGTAAGTGCAGCAACAGAGT	recA	1838698
117F	GCCAGAACTGCACACAGACCTAAAA	secB1	1837883
117R	GATCGCCAGGGTATAAGAGCTTTAG	purT	1854960
118F	GCAGAAGCAGATGGCTGTAATGTCT	purT	1854326
118R	TTGCCACCTTGGGATGCTGATGTTA	yjjK	1871777
119F	CCTAGCCTTTCTCCAGGTGGAATAT	yjjK	1871275
119R	AGGTGTTTATGGCGGAGCTATAGGT	trpE	1888788

APPENDIX II

Primers used for finishing stage to sequence the *Francisella tularensis* FSC198 genome (primers were designed by Dr. Chaudhuri). Gene names and nucleotide position numbers of each primer are specified relative to the *F. tularensis* Schu S4 genome.

Primer name	Sequences (5′–3′)	Gene name	C strand*	Start position	End poistion
Primer1	ATCAATGAGCAAGCAAGTTTTG	dnaN		1921	1942
Primer2	GCAGGGCTATTTCCTGTTAGAA	FTT0003c	c	2874	2895
Primer3	TTCTAACAGGAAATAGCCCTGC	FTT0003c		2874	2895
Primer4	TGTTTAAGTTTTACAAGCGAGGC	gabD1	c	4101	4123
Primer5	CTCTGGCTGAATTAGGACATC	FTT0008		8044	8064
Primer6	CCCGTCTCCGTTAGTATTAT	FTT0010	c	9024	9043
Primer7	GGTATTAAAGAGGTTCCGGATG	FTT0011		9934	9955
Primer8	TTCAAGAGTGTATGTAACTTCAAA	FTT0012	c	10875	10898
Primer9	ATCATTTGTATACTGGCCGAGG	FTT0024c		22738	22759
Primer10	CATGGACAGGTGGAATAACAAA	FTT0025c	c	23692	23713
Primer11	GTTTGCATGCTTGATATTTGGA	FTT0028c		27206	27227
Primer12	CAGCAATGCTTTTCTCTCCTTT	FTT0028c	c	28154	28175
Primer13	TTCGGGTTGCTTAGGATTAAAA	FTT0056c		58942	58963
Primer14	ATGCTTAGTTTCAAAAGCTCGC	FTT0057	c	59910	59931
Primer15	GTGCCTATCTTGTTCAAAAGGG	FTT0057		60029	60050
Primer16	ACGACCAAGTAACATTCCACCT	atpE	c	61105	61126
Primer17	ATGGTTAAGCCTTTTTGGTTGA	qseC		97724	97745
Primer18	AACGAAGAGGATGGTCAGAAAA	qseC	c	98685	98706
Primer19	CCTTTTATGGCAAGAGAAAAGA	FTT0097		101867	101888
Primer20	GTTCAAAGCAAGGAGTTGGTTG	isftu1	c	102830	102851
Primer21	TTCTTGGAATATCAGGGGCTAA	isftu1		102717	102738
Primer22	TTTGATTTACAAGGGATAACAGC	isftu2	c	103650	103672
Primer23	TTGTTGTCAGAGGATAAAACTCA	isftu2		103512	103534
Primer24	AATCTGTAGCAGCACAAGCAAA	FTT0101	c	104583	104604
Primer25	TCAATTTTAGTTGGTGAGGCTG	Intergenic		105878	105899
Primer26	AAGCTTTAAGCCTTGTTGCTCA	FTT0102	c	106778	106799
Primer27	TTCAACACAGCTTTCTCAGCAT	nupC1		124764	124785
Primer28	CCTGTTTCATAATCAGCATGGA	nupC1	c	125678	125699
Primer29	TTTTACAATCTGAAATCTCAAAAC	glpK		143242	143265
Primer30	ATTATTCAAAGGATAGTAGCAA	Intergenic	c	144420	144441
Primer31	GGTTGTCGATGTGTCTGAAAGA	FTT0136		149112	149133
Primer32	TTGTACCTACGTTTACATGCGG	tufA	c	150075	150096
Primer33	ATGGATGAAGGGTTACGTTTTG	tufA		151149	151170
Primer34	CCGATTTAACAAGATTCCAAGC	nusG	c	152123	152144
Primer35	TTCCAATTTTGCATTTTATCGG	FTT0170c		185687	185708
Primer36	AACACTAGCCTCAAAATTCTCCA	FTT0172	c	186955	186977
Primer37	TTGCTGAAAAATATCAACGCTT	FTT0179		195130	195151
Primer38	CTTGCGGATAATTGTACTGCTG	FTT0180	c	196055	196076
Primer39	TAACCCTAAGGCAGGTGGAATA	FTT0180		196506	196527
Primer40	GTTCTAGTCGGAGTTGTTGCCT	FTT0181c	c	197526	197547
Primer41	AGATCAATAAACGCACCGAAAT	rpsA		199526	199547
Primer42	GTTGCCTAGCTTTAATGCTGCT	FTT0184	c	200499	200520
Primer43	AAAAACTTTTCTTTTAAATTTCCC	Intergenic		218631	218654
Primer44	ATGGTCTTGACTCTTTTGCAGC	FTT0201	c	219804	219825
Primer45	GCTGCAAAAGAGTCAAGACCAT	FTT0201		219804	219825
Primer46	CCAAATACAAGACCAGATAAAGGC	FTT0201	c	220711	220734
Primer47	GCACAAAGAGAACTCAAGAATGG	rbn		231088	231110
Primer48	GCTTTCCGACCTGAACTAAGAC	priA	c	232062	232083
Primer49	CATCTTAAGCGTAATAATGTGT	priA		233967	233988

Primer50	CAGATTGGCCTTTTGTATTTCA	FTT0217	c	235145	235166
Primer51	TTGAGCTCTTTATTATCGCCGT	FTT0220c		237199	237220
Primer52	TACAAATTTTGGCTCACCCTCT	acpA	c	238129	238150
Primer53	TAGTGAAGATTCTGCTTCTACC	FTT0225c		242416	242437
Primer54	CAAAGAGCCCAAGAGTTAGCTT	FTT0227c	c	243591	243612
Primer55	CCGATGATCTCGATACAATGAA	htrB		245855	245876
Primer56	TTCCATACAAAATGCTAATGCG	ddg	c	247127	247148
Primer57	ATCACCAGCAGGATCAAAATCT	yidC		248242	248263
Primer58	AGTAAGAAAACAGCAGCCCAAG	rnpA	c	249187	249208
Primer59	CTAAAAACTGTATTTTTGAGCA	Intergenic		259257	259278
Primer60	ATTCACCATATCTTGGATTGGC	Intergenic	c	260490	260511
Primer61	AGGCTGCTATAGCTCAGCAAAT	leuA		269113	269134
Primer62	CCTACTAAATGCTGACATCCTCT	Intergenic	c	270342	270364
Primer63	TCGGACCTCTATCAGAAGCAAT	FTT0268		283374	283395
Primer64	CACGAGGATAAATAATCCCCAA	FTT0268	c	284288	284309
Primer65	TTTGCAATCACTTTTACTTTTGA	FTT0272		287246	287268
Primer66	GCTAGTGCGGTACAAAGTGCTA	FTT0274	c	288511	288532
Primer67	ATTAATGGCGCGATAATACCAC	FTT0274		288576	288597
Primer68	CGATCGATTTTATTAGCCCATC	FTT0275c	c	289459	289480
Primer69	GCCAATTGCTAAACAAAAAGTTG	cydA		293084	293106
Primer70	AGTTTGATAATGGCAGAGCCTA	yajR	c	294162	294183
Primer71	TCTAAGTTTGTTCTGGCACTTCTT	cyoC		299173	299196
Primer72	AGCATACGAATGAGGAATTTGC	cyoE	c	300137	300158
Primer73	TGGGACCATATATCTTGCATCA	FTT0308		323610	323631
Primer74	AATGACTATATTTTGCTATCAACC	Intergenic	c	324773	324796
Primer75	GGATTTGGGAGCATTTACTCAG	pyrH		330731	330752
Primer76	ATCAACACTCGAGACCGAATTT	uppS	c	331634	331655
Primer77	GAAGATCAATGATTGTGCCAGA	rpsS		340259	340280
Primer78	TTGAGCAACACTTTCAGCAACT	rpsC	c	341141	341162
Primer79	GAGCAATCGTTGAGCCTCTAGT	rplQ		350748	350769
Primer80	AATCAGGCAGCATATTCAAGGT	isftu2	c	352283	352304
Primer81	AGAAAAGAAATTCTCTATCAAATG	isftu2		352178	352201
Primer82	TGACACCAATTCTCTAATTCAAAA	FTT0354	c	353134	353157
Primer83	TGGACAGATAGACCAGAAGAAATG GACAGATAGACCAGAAGAAA	FTT0354		353027	353048
Primer84	TGGACAGATAGACCAGAAGAAATG GACAGATAGACCAGAAGAAA	FTT0354	c	353982	354003
Primer85	TGTTTTACATTTAAATTTCGGT	Intergenic		353741	353762
Primer86	GGAGGTGTTAACCAATCCATGT	htpG	c	355307	355328
Primer87	ATCATCTTCGTTTAAAGCTGCG	htpG		356994	357015
Primer88	GATCCAATGGCTTTTGTTAAGC	FTT0358	c	358228	358249
Primer89	TAGCACACATAAGTCCAGCAGC	FTT0362c		363214	363234
Primer90	GAAATATACCTATTCGGCGGC	FTT0364c	c	364480	364501
Primer91	TGCCTCAGGTCATGACTTACTT	poxF		365715	365736
Primer92	TGGTATTGTTCCATACAGAGCG	gshA	c	366619	366640
Primer93	TGTTGTTAGGTGGGGTGTAATG	gshA		366861	366882
Primer94	TTAGCTGAAGGTGTTTGCGATA	gshA	c	367816	367837
Primer95	ATACAGGAAGCAGATTTTGGGA	FTT0369c		369798	369819
Primer96	AACTTTGCTGCTCTTTGGTAGC	FTT0369c	c	370749	370770
Primer97	TCATCATATGCAACTTTGGAGC	Intergenic		378631	378653
Primer99	TCCAGATATATCAAAAATCTCCC	isftu1		379487	379509
Primer98	GGCTATTTCAATCAGCTTAGAGGA	isftu1	c	379596	379619
Primer100	CAGCGGATATACCAATACTGAAA	FTT0378c	c	380419	380442
Primer101	TTCTTTTCATAATATAGCTGGCAA	FTT0378c		380317	380339
Primer102	TCTATTAGCTCTCGACCAGTTTG	Intergenic	c	381283	381305
Primer103	CACCAGTGTTTAAAATCTTTATG	Intergenic		381172	381195
Primer104	AGCAATGATGAAAATAAGTTGAGA	gdh	c	382131	382152
Primer105	CAAATCATGGCTAACATTTTCC	gdh		383337	383359
Primer106	CGCTATCAATGTACTTCTGAGCA	Intergenic	c	384788	384809
Primer107	CTAAATACTCGTTTCCGTTGGC	Intergenic		385204	385226
Primer108	TTAAATCTTCAATGTGTTTTGTT	psd	c	386679	386701
Primer109	TTAGTTGAAGCTAAAGAAACCGA	glmU		389997	390018

Primer110	AAGAGGGTGCTGTTATTGGAAA	glmS	c	390879	390900
Primer111	CACCGCAACAATTTTATCAGAG	pulB		420198	420219
Primer112	TGGATAAACTTGCTGACCATTG	pulB	c	421168	421189
Primer113	GCAGCAAAGGTATTAATGGGAG	pulB		422259	422280
Primer114	GCTTGGATGCGAAACCTAATAC	glgB	c	423194	423215
Primer115	CTCTTGGCCAGATATTTCAACC	glgB		423690	423711
Primer116	GTAGGTTGATAGCCCCATGAAG	pgm	c	424639	424660
Primer117	GTTGATACAGTTTGAATCGCCA	malP		430724	430745
Primer118	GCTATATCCAAAACCTGAACGC	malQ	c	431649	431670
Primer119	ACCAGCAAAAGAAGAGCTTGAC	FTT0424		438026	438047
Primer120	ATCCAAACAATCAGGTTATGCC	asd	c	438993	439014
Primer121	TGCCTAGTGGACAAACAAAAGA	speA		447323	447344
Primer122	TAGAGAAGTGTCGCTTGGTGAA	FTT0434	c	448296	448317
Primer123	ACTGCTCCACCCTCTAAAATCA	yjfH		453023	453044
Primer124	CAGTCAATGCAACGGTTAAAAA	Intergenic	c	454200	454222
Primer125	AATTTAGCAAGAAAATCACAACT	FTT0443		457332	457355
Primer126	CTAAAAGAATAAGTTACAAAGAGC	tet	c	458400	458421
Primer127	CAAGCGCTGTAGATAATCCCAT	Intergenic		472838	472859
Primer128	TTCAGAGTGGTGTTTCGATAGG	FTT0455c	c	473773	473794
Primer129	TCCTAGCTCAAAAATCGTCAGC	FTT0474		492504	492525
Primer130	TACTTCAGCAGTTGCCTCAAAA	msc	c	493394	493415
Primer131	GAGTTCGACGTGATGTTGTTGT	poxA		494336	494357
Primer132	TTTGCATTTCCTGCATTAGTTG	birA	c	495271	495292
Primer133	GGCGATAAATGGATTTCTGAAG	perM		497552	497573
Primer134	CAGCTAAAAGGACCGCATAGTT	xasA	c	498526	498547
Primer135	TGTTTTGGCTGTTTAGTAGCGA	FTT0496		516548	516569
Primer136	ACCAGATATAGCGCCAATGACT	FTT0497c	c	517495	517516
Primer137	TAATTGCAAATGATGAAGGTCG	Intergenic		519981	520002
Primer138	AGTGAGCTTTTTCTTACGGCAG	FTT0501c	c	520910	520931
Primer139	TGCTTGTCGGAACTATGATGAC	FTT0505		525087	525108
Primer140	CAGGTGCATTTGTTGTTGCTAT	FTT0505	c	526065	526086
Primer141	GTCGGCAATCCTAATAAAGCAG	FTT0512		533876	533897
Primer142	TCAAGTCTTTGGGAGTTGAGGT	lldD2	c	535453	535474
Primer143	TCTAAACGCTTGCTCATTCTCA	FTT0516		538234	538255
Primer144	CGTAGAGCTGTATTGCCTTGTG	FTT0516	c	539116	539137
Primer145	ACCGGTGTTTTCCCTAATACCT	FTT0517		540665	540686
Primer146	AAATGGCTTAATTGGTTCAGGA	prmA	c	541554	541575
Primer147	TATAACCAAGCTGCTTCCCACT	FTT0525		547805	547826
Primer148	GAAAGAGCTTTTGTTGATGCAG	FTT0527	c	548980	549001
Primer149	AGGGGCAAAAGATTTTCTACTG	vanY		567575	567596
Primer150	AAAGAGCATGAGCTTATTTCGG	FTT0550	c	568534	568555
Primer151	TGCTTCATTCCTTCAGCAAGTA	FTT0550		568846	568867
Primer152	CAATGGGTATGGGTAGAGTGGT	FTT0551	c	569821	569842
Primer153	AAAACTGGCTTATCATCAGGGA	FTT0553		573064	573085
Primer154	GCTCTACCAAGGTTTTGTTGCT	FTT0555	c	573981	574002
Primer155	AAAGCCTCTGCAAATGGATAAA	Intergenic		578095	578118
Primer156	TTTTATCAGACTTAATCCATTTGA	Intergenic	c	579337	579358
Primer157	GACTCGACAACTTCGATAAAGC	potI		582250	582271
Primer158	ATGGTCCTACGCCTGAGATAAA	FTT0566	c	583425	583446
Primer159	CCCCATCCTTTAACATCTTTTC	FTT0572		589757	589778
Primer160	TTCACCATGTGGATTTGAATGT	alr	c	590675	590696
Primer161	ATCAAGCTCTTTGGCTATTTGC	fdx		599096	599117
Primer162	GCATTTCGAGGAGTCTCAATGT	fopA1	c	600161	600183
Primer163	AGCTTTATCATAAGCAAACTTAG	FTT0584		604079	604100
Primer164	TTAACTGATGGCGATTTTGATG	FTT0585	c	604974	604995
Primer165	CCGCATAACCCTCGTAACTAAG	FTT0585		605030	605051
Primer166	ATTTTTATCAGCGAGTTTTGCC	FTT0586	c	605953	605974
Primer167	CGGACAAAAACACCATTGTCTA	Intergenic		606650	606671
Primer168	TTCAGACGTGTCAAACAGAGGT	aroA	c	607824	607845
Primer169	ATCTGTGATTTTATTTGAAGGA	FTT0594c		612898	612919
Primer170	TCGTTTGATTACCACTGACACC	rubA	c	613786	613807
Primer171	AGATTGGGAATGTCCTGACTGT	FTT0597		615037	615058

Primer172	ATTTAAAAGGAGCGGGCTTAAC	FTT0597	c	615969	615990
Primer173	CGGCAGACTTAAGTGCTTTTTC	FTT0602c		621081	621102
Primer174	TTTGTCTTTTAATGCCCGATTT	FTT0602c	c	622012	622033
Primer175	CAGGTAACAATAGCATCCCCAT	FTT0604		624131	624154
Primer176	AAAGAGTTTAATAGTTTTTGGTCT	FTT0606c	c	625191	625213
Primer177	TCAAAGATTACTGCCCTAATGGA	FTT0612		631591	631612
Primer178	TGTTGCTGCTAGAGGAATGTGT	Intergenic	c	632504	632525
Primer179	AAGCAAAAATCCTGGTAATCCA	FTT0619		638772	638793
Primer180	TTGGCAGTACTGTCTTGAGCAT	FTT0620	c	639690	639711
Primer181	CTTTTTGCTCAATTATCCCCTG	lon		645861	645882
Primer182	GCCTGGTCAAATTATCCAAAAG	lon	c	646747	646768
Primer183	CTGTCATTGCCACATCATTTCT	hupB		647130	647151
Primer184	CTAGATGCTACTATTGCGGCAG	FTT0628	c	648009	648030
Primer185	TTTGTAAATCTTGTTGGCTTGG	FTT0651		671470	671491
Primer186	GCGATAATCCAGTACATGCAAA	FTT0651	c	672365	672386
Primer187	TCCATCTTTAGGAAGAGCAACC	FTT0668		685435	685456
Primer188	TCTGATAGCAAATGTCATCGGT	FTT0669	c	686373	686394
Primer189	TTTGCCCTAGACGAGTCTTCAT	prsA		692239	692260
Primer190	AAAAGATCATCAGCAGAGAGGG	rplY	c	693176	693197
Primer191	GATACCGCTTCTTTTTCACCAC	Intergenic		694757	694779
Primer192	AATGTGACACTGCTCTTGTATGA	FTT0677c	c	695647	695668
Primer193	AGTTTGGTAATTCTGCAAACGA	sthA		701447	701468
Primer194	TGTGGGTCATACTCAACTCCAA	Intergenic	c	702488	702510
Primer195	TTCAAATATTAACCAACCAAAAA	Intergenic		710349	710370
Primer196	TTTCAGCGATTATTAGTGCGAT	mutM	c	711595	711617
Primer197	TTTTTGTAAAAAGTGTCAGAAAT	udk		721341	721362
Primer198	GGTGGTTCTGGTTCTGGTAAAA	tRNA-Val (TAC)	c	722310	722331
Primer199	GCTTTTTAAATCATGGTGGGTG	FTT0719		738726	738747
Primer200	TGTTGGTTATATTGTACGCCCA	FTT0719	c	739653	739674
Primer201	TATTGTCAATGCTGGAACACCT	Intergenic		743956	743977
Primer202	AAAGCCATAACTAAGGATGTTA	Intergenic	c	745223	745244
Primer203	GATAATGTCAAAGCATGGCAGA	ybhR		752941	752962
Primer204	TGCCATCGATATTACTGTCAGG	Intergenic	c	754109	754130
Primer205	ATCTATGCAATGCTGATTGGTG	FTT0744c		767705	767726
Primer206	TAAATGCTAGCCAAAGTCCGAT	FTT0746c	c	768628	768649
Primer207	AACTTTGAAAAACCATGCAACC	glyS		784463	784484
Primer208	ATTTTGCTGCGAGAGTAGAAGC	deoD	c	786009	786030
Primer209	ATCGCCTGGCATAATTACTGTT	FTT0778		798945	798968
Primer210	AGACTTGGTATTTGTATTTTACAT	Intergenic	c	800122	800144
Primer211	TCGTTCTCAGCTAAAGTAACACG	FTT0793		812649	812670
Primer212	TGAGGCTAAGGTTTGGAAATGT	FTT0794	c	813566	813587
Primer213	CCTCAAAAATAGCATAGTCCGC	FTT0796		814908	814931
Primer214	CGTAAAATAAAATTATCTCTTCGT	FTT0797	c	815886	815907
Primer215	TCTAAACCCACATTTCTTGCAC	FTT0797		816282	816303
Primer216	ATTTTCAAAGACCGACAAGCAT	FTT0798	c	817256	817279
Primer217	CGCTGATAATAAAAATAACGCACA	FTT0799		818484	818505
Primer218	CGGGGTTGTATTTTGATCCTTA	Intergenic	c	819430	819451
Primer219	CACGGGGCTCTCTTAAATACTG	capB		825630	825651
Primer220	TTTGCTAATGCTTTTGCTGCTA	FTT0807	c	826565	826586
Primer221	TAGCGCGATAAAAATTTACGGT	FTT0807		826694	826715
Primer222	TAACTAACGACGCATTTGCATT	FTT0807	c	827607	827628
Primer223	ATGACCTGGTGGTGTGACAGTA	spoT		829265	829286
Primer224	TCGTGGCAAAGATATAGTCGAA	recR	c	830186	830207
Primer225	CTCCACTAGATGGCATAGGTCC	infC		837187	837206
Primer226	ATGCTAATCCGCCAGTGTGT	FTT0821	c	838464	838485
Primer227	TCCAAATAAAATTCCAGCTAAA	FTT0821		838866	838887
Primer228	AGGATTATTAACTTTGTTGGCA	Intergenic	c	840116	840137
Primer229	TTGCAATCATAAAAAGTTCTCA	FTT0830c		847315	847336
Primer230	AGCGTAATGGTCCTGTTCCTAA	FTT0831c	c	848250	848271
Primer231	GTTTCACAACAACTGCCAGAAC	tolB		854874	854895
Primer232	TTATCAAAAGCTTACGGGCAAT	FTT0841	c	855846	855867

Primer233	ACTACCGTTTGCATTAGGCATT	FTT0846		860262	860283
Primer234	CAAGTACAAATGCAATCTGGGA	FTT0847	c	861151	861172
Primer235	ATCATAATTCCACTGTCCACCC	FTT0855c		868347	868368
Primer236	CATGGAGTGCTTTTAAACATGA	ubiA	c	869291	869312
Primer237	GTGGCCAACACTTACAGCATTA	FTT0866c		876412	876433
Primer238	ATAGCCTTCGCAGTTGCAGTAT	FTT0866c	c	877548	877571
Primer239	CATTGGTAAAGGTGTAACTGGAGA	FTT0874c		884057	884078
Primer240	CATCAATGCGTTGTTGAGATTT	aroC	c	885026	885047
Primer241	AGCAAGGCTTTTTATCAAACCA	FTT0890c		898802	898823
Primer242	CTTGAGATGCATCACCATTGTT	FTT0891	c	899737	899758
Primer243	AGTTGCTAAGATAAAACCGGCA	FTT0900		909786	909807
Primer244	GTGTTGATTTTATGGCGAGTGA	FTT0902	c	910742	910763
Primer245	TTGTAGTTGGTCTGCTTCGAGA	FTT0902		910742	910763
Primer246	TCTCGAAGCAGACCAACTACAA	lpnB	c	911710	911731
Primer247	ACTACAGCCTGCTAAAGTTGCC	lpnB		912113	912134
Primer248	GAGCTTGCTAAAGCCTCTCTTG	FTT0905	c	913065	913086
Primer249	AAGCCAGCATAAACTAAAGCCA	maeA		927093	927114
Primer250	AATTGCTGCAGGTAGTCCATTT	FTT0918	c	928045	928066
Primer251	CTTTACTCTCACCATTGCCTCC	FTT0919		930643	930664
Primer252	TTTTCTAGCAGCTTCAGGGATT	Intergenic	c	931916	931937
Primer253	CGTAATTTCTCTCGATTACGCC	Intergenic		931970	931991
Primer254	CACCTTTAGAGGCGAGAGAAAA	FTT0922	c	932917	932938
Primer255	TCACTCCAATAGCCAGTGCTAA	FTT0933		945325	945346
Primer256	AGCGTAGATGGATTTCACCATT	bioD	c	946207	946228
Primer257	CTCTTTATTGAGGGTGCTGGAG	bioC		947216	947237
Primer258	GCGAACTTGATTTTGAATAGCC	bioF	c	948160	948182
Primer259	CAACAGTTTGAACATGAGTTTGC	bioA		950071	950092
Primer260	TGCTGGGGGTATGTTGATTTAT	add1	c	950996	951017
Primer261	AGCAATTTATCTAGCCGTCCTG	add1		951251	951272
Primer262	GATTTCCGAACTCATAGCCATC	FTT0940c	c	952312	952333
Primer263	TTGCAACTGGTGGTGAGATAGT	folK		953957	953978
Primer264	CTTCAATCGATATAGGCTTGGC	folB	c	954924	954945
Primer265	AGTCTTGGTTGCTCAGGAGAAG	FTT0945		957369	957390
Primer266	GAGTTTATACCGGAAGCATTGG	trpG1	c	958259	958280
Primer267	CATCGACACTTGATAGCCAAGA	rhlE		964325	964346
Primer268	CCAAAGACTTCTTGATCAACCC	rhlE	c	965253	965274
Primer269	TTGCCTCAAAGCCTTCTAGTTC	FTT0953c		965589	965610
Primer270	ATAGCAGCAATTACAGCGACAA	FTT0953c	c	966544	966565
Primer271	GCTTTAGATGGGGCTATGACAC	FTT0967c		978884	978906
Primer272	TTTTCTAATTTGAGTAAAGCGCC	FTT0968c	c	979853	979874
Primer273	ATGATTTCACCTGCAAGAGGAT	Intergenic		988455	988476
Primer274	TGCTTGTGAACTTGGTCAAGAA	FTT0978c	c	989352	989373
Primer275	CGTGATGTTTTTGCTTATTCCC	FTT0979c		989681	989702
Primer276	AGGTATCAAAGGAAATCCCCAT	FTT0979c	c	990613	990634
Primer277	TCAGGATTCCTAACAGGTTGGT	FTT0980		991896	991917
Primer278	CATTAACTAAAGCCTTCGGTGG	FTT0981	c	992826	992847
Primer279	ATGCATTATCGTGAGCTCCTTT	FTT0989		997564	997585
Primer280	TCACTGTTAACGAGCCAGAAAA	FTT0989	c	998629	998650
Primer281	ATCCATGGTGTTGATAGTGTCG	FTT0991		1001797	1001818
Primer282	ATTCTGGTACACAGCTTGCTCA	FTT0992	c	1002769	1002789
Primer283	CACGCCAATTTTGGACACTAT	FTT0996		1006138	1006159
Primer284	CATGGATTTTGGCGATATTGTT	FTT0996	c	1007016	1007037
Primer285	TGAAGTAAATCAAGCGATGGAA	FTT0996		1007006	1007027
Primer286	CCCTATTTAATTCCATCGCTTG	ybhO	c	1007923	1007944
Primer287	GCAGATTCAATTGCTTGCATAA	ybhO		1007918	1007939
Primer288	AGCTTTTATGCAAGCAATTGAA	ybhO	c	1008859	1008880
Primer289	TTTTGCGTTGAGTCAGCTAAGA	ybhO		1008363	1008384
Primer290	CACTTTGCGACAAATCAAGAAA	FTT0998	c	1009263	1009284
Primer291	CATATTGCTCATCAAAAGCAGC	FTT1008c		1020911	1020932
Primer292	AAAATCTCTGGCAATTGCTTTG	FTT1010	c	1021830	1021851
Primer293	ATCTCCAGGTTGACATTCAAGA	FTT1015		1025217	1025238
Primer294	CCAGCTTTAGATGTGTTTCGTG	FTT1016c	c	1026190	1026211

Primer295	TTGGTGATAGTCTCACAGCAGG	FTT1025c		1036098	1036119
Primer296	CCTTTTCGCCACGATATACATT	FTT1026c	c	1037065	1037086
Primer297	AAACATGCTCAGCAATTAGCAA	lipB		1040660	1040681
Primer298	AAGCTTGGTGCCAAAAAGTTAG	FTT1032	c	1041585	1041606
Primer299	GCCAATATGGGAGATAAACCAA	Intergenic		1057601	1057622
Primer300	GTTGGCGATGGAAGTGTAGATT	FTT1048c	c	1058579	1058600
Primer301	TGAAATTGGTGATCAAAAGCAG	Intergenic		1062932	1062953
Primer302	ACCATCTGCACCAAAAGAAGAT	rimI	c	1064461	1064482
Primer303	GATGGTAGTCAAAGTGATGCGA	rpsF		1070558	1070579
Primer304	CCTTTTTCTTCGATTATGCCAC	hemF	c	1071503	1071524
Primer305	TTCAAATATTTCGCCACCTTCT	Intergenic		1073046	1073067
Primer306	CCAGCACAATCAGATGGATTTA	Intergenic	c	1074290	1074311
Primer307	TGTTAGTAATGCGTATGCTCGC	Intergenic		1074285	1074306
Primer308	TACTTGCGAGCATACGCATTAC	res	c	1075256	1075277
Primer309	TCTTACCATAGCACTGGCGATA	FTT1073c		1084855	1084876
Primer310	GAGCTTTAGCTTGATCTGGCAT	Intergenic	c	1086329	1086351
Primer311	GCTTGCTTTAGACTTCTTAGCTG	Intergenic		1086875	1086896
Primer312	CCCTTTGGTCTTTACTTGTTGC	hipA	c	1087781	1087802
Primer313	CATCATTTTATCTTGAGCACCG	rdgC		1093050	1093071
Primer314	TACCTTGGCATGATATGTTTGC	FTT1085	c	1093975	1093996
Primer315	TGCAACAGTACCATTGATTCCT	rep		1097372	1097393
Primer316	GCTGCCTGCTCTTCTAATTCAT	FTT1088c	c	1098341	1098362
Primer317	AAAAAGTACGGGATGTTTTCCA	FTT1091		1100780	1100801
Primer318	TTCAGCATTTGATGGTTTTGAC	talA	c	1101748	1101769
Primer319	ATCTCCAGTTTTGCTTGAGGAG	FTT1110		1121173	1121194
Primer320	TCACCTTCATCATGAGCACAAT	rpoH	c	1122345	1122367
Primer321	GCCAAATTGAAGAAGATGCTCTA	yajC		1127020	1127041
Primer322	CTCCTTGTGCAAATGAAGTTGA	FTT1117c	c	1127997	1128018
Primer323	AGAGTGCTCCAAAGTATCTCCG	metN		1134275	1134296
Primer324	GCACTTGCACTAAATCCACTTG	metIQ	c	1135236	1135257
Primer325	GTCTTTTAATTGCTCCCATTGC	phrB		1151248	1151269
Primer326	TCGTCAAGATTTACGCTTAGCA	phrB	c	1152132	1152153
Primer327	TCTGTCCTTTTTGCCATTTCTT	FTT1140		1153576	1153599
Primer328	GAAAGCATACTATAAAGCCAAAGA	FTT1141	c	1154752	1154773
Primer329	ATGCTACCTCTCCTGTATTGGT	hsdS		1187690	1187711
Primer330	AGTTGTTGCCACATCATCAAAC	Intergenic	c	1188641	1188662
Primer331	AAAATAACGGATACAAGGGCAA	Intergenic		1189157	1189178
Primer332	CTCGGCTTTTCAGGTTTCTCTA	hsdR1	c	1190050	1190071
Primer333	GGCGAGGGAGTTTCTACTTTTT	FTT1201c		1220088	1220109
Primer334	CAATCGGATCTTGGGTATCAAT	FTT1202	c	1220966	1220987
Primer335	TAGCAACAGTACGTAAAGGCGA	FTT1209c		1228498	1228519
Primer336	CAATCTCATTCATTTCCGGATT	FTT1209c	c	1229405	1229426
Primer337	TAGCCAGGGTTCTCTTGGTTTA	FTT1209c		1229569	1229591
Primer338	AATTTGCTAGTGGTTTTAAGCCT	FTT1211c	c	1230826	1230847
Primer339	TTGGTGCATACTTGCTTTTGTT	FTT1213		1231794	1231815
Primer340	GCAGCAAACATCCAAGTTAATA	FTT1214c	c	1233172	1233193
Primer341	GAAACTTTTCCCACAAAGTGCT	FTT1220		1239199	1239220
Primer342	ATAAAACATGGGCTTTGGTTTG	dedA2	c	1240080	1240101
Primer343	TGAGAATTCGTGCGTGTGTACT	rne		1243672	1243693
Primer344	AAACAAGAATTGCCACCCTAGA	rne	c	1244606	1244627
Primer345	AAGTTGACGAGCAACTTCTTCC	mesJ		1250227	1250248
Primer346	ATCGTGGCAGAAATAAGCCTAA	yjdL	c	1251399	1251420
Primer347	GTTGTTCCAAGGGCTATTTACC	yjdL		1251418	1251439
Primer348	AACTGCGCCTGCTTAATTAGAT	yjdL	c	1252310	1252331
Primer349	TTATACTTGGGTGGGAATTTCG	yjdL		1252310	1252331
Primer350	CGAAATTCCCACCCAAGTATAA	FTT1234	c	1253283	1253304
Primer351	TTTTTGATATTCGGCAAATCCT	FTT1236		1256083	1256104
Primer352	TGATGAAATTGAAAATGCCAAC	FTT1237	c	1257024	1257045
Primer353	ATGGTAGCGATTTAATTTTGCG	FTT1238c		1257898	1257920
Primer354	AAACAATATAGCACTTAGACTGA	FTT1239	c	1259041	1259062
Primer355	CAGAAATTCTTCCCGTCAGAGA	FTT1239		1260071	1260092
Primer356	ATGAAGATCCGAATGAATACGC	glyA	c	1260986	1261007

Primer357	CTTTTACCATGATAGCCTTCGG	glyA		1261923	1261944
Primer358	AACGACGAGAGGTTTCAAAGAG	FTT1242	c	1262838	1262859
Primer359	ACTGAGCGATTTTCTCTTCAGC	FTT1252		1272553	1272574
Primer360	TCAAAAAGCTACCGAACTAGGG	FTT1253	c	1273437	1273458
Primer361	GTTTAGCGCCAAGAACTTTGTC	FTT1253		1273213	1273234
Primer362	AGGAGAAAATGCCTATGCAAAA	FTT1253	c	1274114	1274135
Primer363	GACATGCCAACCAAATACAGAA	FTT1261c		1283453	1283474
Primer364	AATCATCAAGAACGTCAACACT	Intergenic	c	1285027	1285050
Primer365	GGCGGTATTATTCAACTATTTCCT	Intergenic		1284867	1284889
Primer366	CTCGGATTAACTATAAACCAAGA	FTT1263c		1285821	1285843
Primer83	GATTGATTTGACACCAATTCTCT	FTT1263c	c	1285938	1285959
Primer367	GAAATTCTCTATCAAATGTCTTTC	isftu2	c	1286779	1286802
Primer368	TCCATGCTATGACTGATGCTTT	isftu2		1286537	1286558
Primer369	AGGTGATTGATGAAAGCGAGAT	yhhW	c	1287877	1287898
Primer370	TCAAATCCATAATGAGGCATGA	yhhW		1288354	1288375
Primer371	CTTTGCTATGCCTAAGCCATCT	FTT1267	c	1289309	1289330
Primer372	AGTTGCTGAACCCTTGATGATT	FTT1284c		1302959	1302980
Primer373	TCGAAGTAACCACAGAAGGACA	FTT1285c	c	1303857	1303878
Primer374	GCTTTGGAGCAACAAGGTTATC	FTT1285c		1303714	1303735
Primer375	TTCAACCATGTAGTGAACTGGG	FTT1286	c	1304670	1304691
Primer376	TCCTGGTAACCCTGTCTTCTGT	cbs		1306132	1306153
Primer377	TGTCTGCGTAATTCATATTGCC	FTT1288	c	1307017	1307038
Primer378	AATAGATATCTCATAAATTACCCC	Intergenic		1307330	1307353
Primer379	ATCGTGCCAAGAGTATAAAGGC	23S rRNA	c	1308306	1308327
Primer380	TACCTTTTATCCGTTGAGCGAT	23S rRNA		1308199	1308220
Primer381	TGGTTGTCCAGGTGAAAGTATG	23S rRNA	c	1309156	1309177
Primer382	TAGAAGCTTTTCCTGGAAGCAT	23S rRNA		1309048	1309069
Primer383	CGTAGCGAAAGCGAGTTTTAAT	23S rRNA	c	1309982	1310003
Primer384	ATCCACAGCTCATCCCATACTT	23S rRNA		1309853	1309874
Primer385	TGTAGCTCAGTTGGTTAGAGCG	tRNA-Ala (TGC)	c	1310818	1310839
Primer386	TAAAATATTGGTGGAGCCAAGC	tRNA-Ile (GAT)		1310670	1310691
Primer387	GAGTACTAGCTGTTGGAGTCGG	16S rRNA	c	1311639	1311660
Primer388	CAAGACCAGGTAAGGTTCTTCG	16S rRNA		1311485	1311506
Primer389	TCAGAATTTGAGAATTAAACTGA	16S rRNA	c	1312461	1312483
Primer390	AGATTCCTACGCGTTACTCACC	16S rRNA		1312343	1312364
Primer391	TACGGTTTTAGCACCTCTTGGT	FTT1289	c	1313302	1313323
Primer392	TAGCAGAGCATAATGGTGCCTA	FTT1289		1313181	1313202
Primer393	TGATTGCCTTGTAGTTTTTGGA	metG	c	1314158	1314179
Primer394	AAAAATGCGAAAGATACTCGTTA	metG		1314054	1314076
Primer395	CCATTTACAGTTAAGAAACCATT	metG	c	1315018	1315040
Primer396	GGGGGTGCTACCCTATATTTTC	ispZ		1318832	1318853
Primer397	ATGGCTCAGTTGCTGGTAATTT	glk1	c	1319729	1319750
Primer398	GCTGTGTAACATTATCCCACCA	FTT1301c		1324845	1324866
Primer399	GTGCCTTACGTACAGCATAACC	FTT1302	c	1325816	1325837
Primer400	CCTCCTCAACATAAAATTAACAA	Intergenic		1331918	1331940
Primer401	CAGAAATAGCTGATGATTCTAGCC	hflB	c	1332982	1333005
Primer402	CATGGATTCACTCGAAACTCAA	FTT1321		1349196	1349217
Primer403	GCTGTTTGTTTACTTGGGAAGC	FTT1322	c	1350074	1350095
Primer404	CCACCTAAGGTTGCTCGATTAC	FTT1328c		1356626	1356647
Primer405	TCAATAGCTGCACAGATTGCTT	gpmI	c	1357602	1357623
Primer406	TTTCTTTAGCTTAAGCCTTGCG	cydC		1365873	1365894
Primer407	GACCGCCTGATAATTGTCTACC	cydC	c	1366822	1366843
Primer408	GATAATCACAATACCCGCAACA	dctA		1367183	1367204
Primer409	AGGTAAAATAGGTGGAGTGGCA	dctA	c	1368074	1368095
Primer410	TGCCACTCCACCTATTTTACCT	dctA		1368074	1368095
Primer411	TCTGGAGCTGAAGCAATAGATG	FTT1339c	c	1368963	1368984
Primer412	CCAACCGTAGGAGATACTCCAG	FTT1339c		1370228	1370249
Primer413	CCATTCAGAGCTATAAAACAAGAA	FTT1341	c	1371404	1371427
Primer414	TGATAGCGCCTATAGATTGGGT	FTT1341		1371534	1371555
Primer415	TGGTTTCCAACCTTCAAAAACT	FTT1342	c	1372466	1372487

Primer416	CATGTGAACTAGCATTAGCAGA	FTT1343c		1374203	1374224
Primer417	ATTTAAAGTGGTACGCGAGCTGAT TTAAAGTGGTACGCGAGCTG	23S rRNA	c	1375180	1375201
Primer418	ATTTAAAGTGGTACGCGAGCTGAT TTAAAGTGGTACGCGAGCTG	23S rRNA		1375072	1375092
Primer419	ACACCAGTGGTTCGTTCACTCACA CCAGTGGTTCGTTCACTC	23S rRNA	c	1376019	1376040
Primer420	ACACCAGTGGTTCGTTCACTCACA CCAGTGGTTCGTTCACTC	23S rRNA		1375894	1375915
Primer421	TGATGGTGATGAGACTTGCTCTTG ATGGTGATGAGACTTGCTCT	23S rRNA	c	1376851	1376872
Primer422	TGATGGTGATGAGACTTGCTCTTG ATGGTGATGAGACTTGCTCT	23S rRNA		1376723	1376744
Primer423	CTTCAATTAACCTTCCAGCACCCT TCAATTAACCTTCCAGCACC	23S rRNA	c	1377681	1377702
Primer424	CTTCAATTAACCTTCCAGCACCCT TCAATTAACCTTCCAGCACC	23S rRNA		1377552	1377573
Primer425	ATGGGGGTTTTACGACCTTACTAT GGGGGTTTTACGACCTTACT	16S rRNA	c	1378515	1378536
Primer426	ATGGGGGTTTTACGACCTTACTAT GGGGGTTTTACGACCTTACT	16S rRNA		1378387	1378408
Primer427	CCCACTTAGCTATGACTTTGGGCC CACTTAGCTATGACTTTGGG	16S rRNA	c	1379343	1379364
Primer428	CCCACTTAGCTATGACTTTGGGCC CACTTAGCTATGACTTTGGG	16S rRNA		1379227	1379248
Primer429	CGATGAAGGACGTGATAATCTGCG ATGAAGGACGTGATAATCTG	pdpA1	c	1380183	1380206
Primer430	CGATGAAGGACGTGATAATCTGCG ATGAAGGACGTGATAATCTG	pdpA1		1380056	1380077
Primer431	TAGATGTTTCAGTTCCCCTCGTTA GATGTTTCAGTTCCCCTCGT	pdpA1	c	1381131	1381154
Primer432	TAGATGTTTCAGTTCCCCTCGTTA GATGTTTCAGTTCCCCTCGT	pdpA1		1380943	1380964
Primer433	GTCAGCTCGTGTTGTGAAATGTGT CAGCTCGTGTTGTGAAATGT	pdpA1	c	1381961	1381982
Primer434	GTCAGCTCGTGTTGTGAAATGTGT CAGCTCGTGTTGTGAAATGT	pdpA1		1381770	1381791
Primer435	GTAAGGGCCATGATGACTTGACGT AAGGGCCATGATGACTTGAC	pdpB1	c	1382823	1382844
Primer436	GTAAGGGCCATGATGACTTGACGT AAGGGCCATGATGACTTGAC	pdpB1		1382679	1382700
Primer437	GTTGGATTAGCTAGTTGGTGGGGT TGGATTAGCTAGTTGGTGGG	pdpB1	c	1383657	1383678
Primer438	GTTGGATTAGCTAGTTGGTGGGGT TGGATTAGCTAGTTGGTGGG	pdpB1		1383553	1383574
Primer439	ATTGTCCAATATTCCCCACTGCAT TGTCCAATATTCCCCACTGC	pdpB1	c	1384526	1384547
Primer440	ATTGTCCAATATTCCCCACTGCAT TGTCCAATATTCCCCACTGC	pdpB1		1384403	1384424
Primer441	CGCAATATCTTTTATATTTTTCGT CGCAATATCTTTTATATTTTTCGT	pdpB1	c	1385351	1385374
Primer442	CGCAATATCTTTTATATTTTTCGT CGCAATATCTTTTATATTTTTCGT	pdpB1		1385232	1385255
Primer443	GCCAAGTTTGCTTGTGAAATTAGC CAAGTTTGCTTGTGAAATTA	FTT1347	c	1386195	1386217
Primer444	GCCAAGTTTGCTTGTGAAATTAGC CAAGTTTGCTTGTGAAATTA	FTT1347		1386083	1386104
Primer445	TTTGCTCAAGTTGTTGAATAAAAA TTTGCTCAAGTTGTTGAATAAAAA	FTT1348	c	1387018	1387040
Primer446	TTTGCTCAAGTTGTTGAATAAAAA TTTGCTCAAGTTGTTGAATAAAAA	FTT1348		1386882	1386904
Primer447	CCCACTAAAAGCACTTTGGACTCC CACTAAAAGCACTTTGGACT	FTT1348	c	1387839	1387860

Primer448	CCCACTAAAAGCACTTTGGACTCC CACTAAAAGCACTTTGGACT	FTT1348		1387711	1387734
Primer449	TTTCTTGAAATTCTGCGATTGATT TCTTGAAATTCTGCGATTGA	FTT1349	c	1388660	1388681
Primer450	TTTCTTGAAATTCTGCGATTGATT TCTTGAAATTCTGCGATTGA	FTT1349		1388556	1388579
Primer451	ACCCTAGTCTGACACATGACAAAC CCTAGTCTGACACATGACAA	FTT1350	c	1389531	1389554
Primer452	ACCCTAGTCTGACACATGACAAAC CCTAGTCTGACACATGACAA	FTT1350		1389420	1389443
Primer453	TGGATATATCGAAGCAAAAGTTTG GATATATCGAAGCAAAAGTT	FTT1350	c	1390376	1390398
Primer454	TGGATATATCGAAGCAAAAGTTTG GATATATCGAAGCAAAAGTT	FTT1351		1390231	1390253
Primer455	AAACCTCAAGCTCAAATCATTGAA ACCTCAAGCTCAAATCATTG	FTT1352	c	1391209	1391230
Primer456	AAACCTCAAGCTCAAATCATTGAA ACCTCAAGCTCAAATCATTG	FTT1352		1391021	1391042
Primer457	TACTTGGGGTCGTATACTTGATTA CTTGGGGTCGTATACTTGAT	FTT1352	c	1392064	1392086
Primer458	TACTTGGGGTCGTATACTTGATTA CTTGGGGTCGTATACTTGAT	FTT1352		1391956	1391979
Primer459	TTTTTGAGTATTTGGATAACAATT TTTGAGTATTTGGATAACAA	FTT1353	c	1392918	1392941
Primer460	TTTTTGAGTATTTGGATAACAATT TTTGAGTATTTGGATAACAA	pdpC1		1392820	1392839
Primer461	TTTAAGAGCACCCTTGTTAGATTT TAAGAGCACCCTTGTTAGAT	pdpC1	c	1393763	1393784
Primer462	TTTAAGAGCACCCTTGTTAGATTT TAAGAGCACCCTTGTTAGAT	pdpC1		1393644	1393665
Primer463	AGAACCCGACTTTATGATTGCTAG AACCCGACTTTATGATTGCT	pdpC1	c	1394610	1394632
Primer464	AGAACCCGACTTTATGATTGCTAG AACCCGACTTTATGATTGCT	pdpC1		1394508	1394530
Primer465	TCATTATCTTTGTTTGGTATAGGA TCATTATCTTTGTTTGGTATAGGA	pdpC1	c	1395464	1395487
Primer466	TCATTATCTTTGTTTGGTATAGGA TCATTATCTTTGTTTGGTATAGGA	pdpC1		1395337	1395358
Primer467	AAAGATTTTTCAAAACCTGAGGAA AAAGATTTTTCAAAACCTGAGGAA	pdpC1	c	1396313	1396335
Primer468	AAAGATTTTTCAAAACCTGAGGAA AAAGATTTTTCAAAACCTGAGGAA	pdpC1		1396190	1396213
Primer469	TCGATTCAGTTGCATTATGAGTGT CGATTCAGTTGCATTATGAGTG	FTT1355	c	1397139	1397160
Primer470	TCGATTCAGTTGCATTATGAGTGT CGATTCAGTTGCATTATGAGTG	FTT1355		1397036	1397057
Primer471	TTCTTATGTCAAAAGCAGACCATT CTTATGTCAAAAGCAGACCA	iglD1	c	1398011	1398032
Primer472	TTCTTATGTCAAAAGCAGACCATT CTTATGTCAAAAGCAGACCA	iglD1		1397900	1397921
Primer473	CATCATAAGCGCTGATACTATGGC ATCATAAGCGCTGATACTATGG	iglC1	c	1398841	1398862
Primer474	CATCATAAGCGCTGATACTATGGC ATCATAAGCGCTGATACTATGG	iglC1		1398727	1398748
Primer475	GCCTTATGAAGAATTTAAAGGTTG CCTTATGAAGAATTTAAAGGTT	iglB1	c	1399676	1399698
Primer476	GCCTTATGAAGAATTTAAAGGTTG CCTTATGAAGAATTTAAAGGTT	iglB1		1399564	1399585
Primer477	AGCTTTTCTTGATTAAACTGTGAG CTTTTCTTGATTAAACTGTG	iglB1	c	1400499	1400520
Primer478	AGCTTTTCTTGATTAAACTGTGAG CTTTTCTTGATTAAACTGTG	iglB1		1400364	1400387
Primer479	CCAAATACTGCTAGACATAGCAAC CCAAATACTGCTAGACATAGCAAC	iglA1	c	1401331	1401354

Primer480	CCAAATACTGCTAGACATAGCAAC CCAAATACTGCTAGACATAGCAAC	iglA1		1401205	1401228
Primer481	ATACCCAAGGAAAGCTAAATGGAT ACCCAAGGAAAGCTAAATGG	pdpD1	c	1402162	1402185
Primer482	ATACCCAAGGAAAGCTAAATGGAT ACCCAAGGAAAGCTAAATGG	pdpD1		1402021	1402042
Primer483	CACTAACGAACAACTTACCAGCTT CACTAACGAACAACTTACCAGCTT	pdpD1	c	1402986	1403009
Primer484	CACTAACGAACAACTTACCAGCTT CACTAACGAACAACTTACCAGCTT	pdpD1		1402872	1402894
Primer485	TTTCGTCATTATGGTTTTCAGTTT TTTCGTCATTATGGTTTTCAGTTT	pdpD1	c	1403843	1403864
Primer486	TTTCGTCATTATGGTTTTCAGTTT TTTCGTCATTATGGTTTTCAGTTT	pdpD1		1403736	1403759
Primer487	TTCTCAATATCCATTCTCCTAGCC TTCTCAATATCCATTCTCCTAGCC	pdpD1	c	1404688	1404709
Primer488	TTCTCAATATCCATTCTCCTAGCC TTCTCAATATCCATTCTCCTAGCC	pdpD1		1404586	1404606
Primer489	CGATGGTATGATATTTTCTCATTC GATGGTATGATATTTTCTCATT	FTT1361c	c	1405555	1405576
Primer490	CGATGGTATGATATTTTCTCATTC GATGGTATGATATTTTCTCATT	FTT1361c		1405445	1405466
Primer491	TCAAAACACCTTTAGCTTTATTAT CAAAACACCTTTAGCTTTATTA	FTT1361c	c	1406378	1406401
Primer492	TCAAAACACCTTTAGCTTTATTAT CAAAACACCTTTAGCTTTATTA	Intergenic		1406266	1406287
Primer493	TGATGGACTAATTCTTGTGAAATG ATGGACTAATTCTTGTGAAA	FTT1362	c	1407240	1407261
Primer494	TGATGGACTAATTCTTGTGAAATG ATGGACTAATTCTTGTGAAA	FTT1362		1407133	1407154
Primer495	CAAGCAAAATATGAAAACGCTGCA AGCAAAATATGAAAACGCTG	isftu1	c	1408097	1408119
Primer496	CAAGCAAAATATGAAAACGCTGCA AGCAAAATATGAAAACGCTG	isftu1		1407974	1407994
Primer497	TGTCAAAAAGATCTTCAAAATAGT GTCAAAAAGATCTTCAAAATAG	treA	c	1408936	1408957
Primer498	TGTCAAAAAGATCTTCAAAATAGT GTCAAAAAGATCTTCAAAATAG	FTT1384c		1429607	1429629
Primer499	TGCAGATACAGATGCTATCAAAGA TGCAGATACAGATGCTATCAAAGA	Intergenic	c	1430882	1430905
Primer500	TGCAGATACAGATGCTATCAAAGA TGCAGATACAGATGCTATCAAAGA	FTT1395c		1443820	1443841
Primer501	CATATTGAAATCTGCTGTTTTAGG CATATTGAAATCTGCTGTTTTAGG	recC	c	1444755	1444776
Primer502	CATATTGAAATCTGCTGTTTTAGG CATATTGAAATCTGCTGTTTTAGG	recC		1445618	1445639
Primer503	TTAATTTATGAAAAATCCGATTAA TTTATGAAAAATCCGA	recC	c	1446528	1446549
Primer504	TTAATTTATGAAAAATCCGATTAA TTTATGAAAAATCCGA	FTT1399		1450410	1450431
Primer505	TTCATAAACTGAGTAAACTGCTTT CATAAACTGAGTAAACTGCT	FTT1401	c	1451337	1451358
Primer506	TTCATAAACTGAGTAAACTGCTTT CATAAACTGAGTAAACTGCT	FTT1417		1468391	1468412
Primer507	AAAATCAAACGCGACATAATTGAA AATCAAACGCGACATAATTG	nusB	c	1469343	1469364
Primer508	AAAATCAAACGCGACATAATTGAA AATCAAACGCGACATAATTG	Intergenic		1475775	1475796
Primer509	TCCATTCGATTAGAAACATTTGCT CCATTCGATTAGAAACATTTGC	FTT1428c	c	1477025	1477046
Primer510	TCCATTCGATTAGAAACATTTGCT CCATTCGATTAGAAACATTTGC	yagD		1482916	1482937
Primer511	AAACTTCTAAAACAGACATCATCA AACTTCTAAAACAGACATCATC	FTT1437c	c	1484066	1484087

Primer512	AAACTTCTAAAACAGACATCATCA AACTTCTAAAACAGACATCATC	Intergenic		1498112	1498135
Primer513	TGAGAATAGGAGTATTTTGCACCA TGAGAATAGGAGTATTTTGCACCA	wbtM	c	1499086	1499109
Primer514	TGAGAATAGGAGTATTTTGCACCA TGAGAATAGGAGTATTTTGCACCA	wzx		1502871	1502892
Primer515	GTAAGAAGAAAGCGGAGACAGGGT AAGAAGAAAGCGGAGACAGG	wzx	c	1503770	1503791
Primer516	GTAAGAAGAAAGCGGAGACAGGGT AAGAAGAAAGCGGAGACAGG	wzx		1503718	1503739
Primer517	TGAGTTTCTGCTCTTCATGTTCAT GAGTTTCTGCTCTTCATGTTCA	wbtI	c	1504669	1504690
Primer518	TGAGTTTCTGCTCTTCATGTTCAT GAGTTTCTGCTCTTCATGTTCA	wbtG		1508410	1508431
Primer519	TGCTCTTAATAGAGCCTCAAAACC TGCTCTTAATAGAGCCTCAAAACC	wzy	c	1509353	1509374
Primer520	TGCTCTTAATAGAGCCTCAAAACC TGCTCTTAATAGAGCCTCAAAACC	wbtA		1516305	1516326
Primer521	AAGTCTTTAACTGATTTGCTGGAA GTCTTTAACTGATTTGCTGG	Intergenic	c	1517427	1517448
Primer522	AAGTCTTTAACTGATTTGCTGGAA GTCTTTAACTGATTTGCTGG	galP1		1527296	1527317
Primer523	ATAGCAGCATGCAAAGATGAGAAT AGCAGCATGCAAAGATGAGA	galT	c	1528178	1528199
Primer524	ATAGCAGCATGCAAAGATGAGAAT AGCAGCATGCAAAGATGAGA	FTT1477c		1529847	1529868
Primer525	CGCAGAATCTAATAGGGCAAATCG CAGAATCTAATAGGGCAAAT	kdsB	c	1530808	1530829
Primer526	CGCAGAATCTAATAGGGCAAATCG CAGAATCTAATAGGGCAAAT	Intergenic		1575969	1575990
Primer527	AGGCTTTTGGAAATCAAATTCAAG GCTTTTGGAAATCAAATTCA	Intergenic	c	1577116	1577139
Primer528	AGGCTTTTGGAAATCAAATTCAAG GCTTTTGGAAATCAAATTCA	FTT1538c		1601033	1601054
Primer529	TCCAATAAGTGCTAAGGCTTTTTC CAATAAGTGCTAAGGCTTTT	FTT1539c	c	1602004	1602025
Primer530	TCCAATAAGTGCTAAGGCTTTTTC CAATAAGTGCTAAGGCTTTT	FTT1547		1609461	1609482
Primer531	GCAACATACTGGCAAACTTCCTGC AACATACTGGCAAACTTCCT	FTT1547	c	1610739	1610760
Primer532	GCAACATACTGGCAAACTTCCTGC AACATACTGGCAAACTTCCT	FTT1558c		1620464	1620485
Primer533	TTGTAGATGGATTTGACTCAGCAT TGTAGATGGATTTGACTCAGCA	proC	c	1621419	1621440
Primer534	TTGTAGATGGATTTGACTCAGCAT TGTAGATGGATTTGACTCAGCA	ispA		1624174	1624195
Primer535	TCGACTCGACATCCACAATACTTC GACTCGACATCCACAATACT	pcs	c	1625125	1625146
Primer536	TCGACTCGACATCCACAATACTTC GACTCGACATCCACAATACT	FTT1564		1626401	1626422
Primer537	TGCAAGAGGTTTGTCAAGAAGATG CAAGAGGTTTGTCAAGAAGA	FTT1565c	c	1627336	1627357
Primer538	TGCAAGAGGTTTGTCAAGAAGATG CAAGAGGTTTGTCAAGAAGA	FTT1579c		1643832	1643853
Primer539	CCATATTGATCAAATTCTGAAACG CCATATTGATCAAATTCTGAAACG	FTT1579c	c	1644790	1644811
Primer540	CCATATTGATCAAATTCTGAAACG CCATATTGATCAAATTCTGAAACG	FTT1579c		1645109	1645130
Primer541	TGTTGATGGTGTCTTAAAGAAAAA TGTTGATGGTGTCTTAAAGAAAAA	FTT1580c	c	1646001	1646022
Primer542	TGTTGATGGTGTCTTAAAGAAAAA TGTTGATGGTGTCTTAAAGAAAAA	Intergenic		1648268	1648289
Primer543	TCTTCTAAAACCCTATCAACACCA TCTTCTAAAACCCTATCAACACCA	FTT1585	c	1649776	1649797

Primer544	TCTTCTAAAACCCTATCAACACCA TCTTCTAAAACCCTATCAACACCA	FTT1594		1659019	1659041
Primer545	TGTTGAGTTTTATGGATAGCCTTG TGTTGAGTTTTATGGATAGCCTTG	FTT1595	c	1659997	1660017
Primer546	TGTTGAGTTTTATGGATAGCCTTG TGTTGAGTTTTATGGATAGCCTTG	FTT1598		1664836	1664859
Primer547	TCGATTCAACCACTATTCCTGTTC GATTCAACCACTATTCCTGT	Intergenic	c	1666014	1666035
Primer548	TCGATTCAACCACTATTCCTGTTC GATTCAACCACTATTCCTGT	FTT1609		1675051	1675072
Primer549	TGATGCTCTAGAAAATACCACTTC TGATGCTCTAGAAAATACCACTTC	FTT1611	c	1675988	1676009
Primer550	TGATGCTCTAGAAAATACCACTTC TGATGCTCTAGAAAATACCACTTC	FTT1613		1676976	1676997
Primer551	GAACCTAGGGATACAGAAGATGCG AACCTAGGGATACAGAAGATGC	FTT1614c	c	1677892	1677913
Primer552	GAACCTAGGGATACAGAAGATGCG AACCTAGGGATACAGAAGATGC	FTT1614c		1677973	1677994
Primer553	TGAAAGCTTTAAGTCTCCTTTTTG AAAGCTTTAAGTCTCCTTTT	cysS	c	1679538	1679559
Primer554	TGAAAGCTTTAAGTCTCCTTTTTG AAAGCTTTAAGTCTCCTTTT	FTT1619		1683196	1683217
Primer555	GGAATAGAGATTCTTATGGTCCTC GGAATAGAGATTCTTATGGTCCTC	Intergenic	c	1684370	1684391
Primer556	GGAATAGAGATTCTTATGGTCCTC GGAATAGAGATTCTTATGGTCCTC	hsdR3		1707462	1707483
Primer557	CAAGTGCTTGGTGGTGGTAATACA AGTGCTTGGTGGTGGTAATA	Intergenic	c	1708631	1708654
Primer558	CAAGTGCTTGGTGGTGGTAATACA AGTGCTTGGTGGTGGTAATA	FTT1652c		1719283	1719304
Primer559	TTGCTCCAGTAGCTGCAAGATTTG CTCCAGTAGCTGCAAGAT	FTT1653	c	1720253	1720274
Primer560	TTGCTCCAGTAGCTGCAAGATTTG CTCCAGTAGCTGCAAGAT	emrA2		1720614	1720636
Primer561	CCTAAATCAACAGGTCGAGAGCCC TAAATCAACAGGTCGAGAGC	FTT1655	c	1721689	1721712
Primer562	CCTAAATCAACAGGTCGAGAGCCC TAAATCAACAGGTCGAGAGC	Intergenic		1738196	1738218
Primer563	GGCTCTTACGATACTTTGTGCCGG CTCTTACGATACTTTGTGCC	napH	c	1739661	1739683
Primer564	GGCTCTTACGATACTTTGTGCCGG CTCTTACGATACTTTGTGCC	FTT1682		1751153	1751174
Primer565	GCTTAAACAAGTTTTAAGATTGCG GCTTAAACAAGTTTTAAGATTGCG	FTT1682	c	1752116	1752137
Primer566	GCTTAAACAAGTTTTAAGATTGCG GCTTAAACAAGTTTTAAGATTGCG	FTT1682		1752472	1752493
Primer567	TCCCATCTAGTGATGTTCCAGATC CCATCTAGTGATGTTCCAGA	FTT1683c	c	1753350	1753371
Primer568	TCCCATCTAGTGATGTTCCAGATC CCATCTAGTGATGTTCCAGA	FTT1691		1760581	1760604
Primer569	AAACCAAAAATGATTCTGCCTGAA ACCAAAAATGATTCTGCCTG	Intergenic	c	1761850	1761871
Primer570	AAACCAAAAATGATTCTGCCTGAA ACCAAAAATGATTCTGCCTG	groL		1765804	1765825
Primer571	AGAACAAAGCGTACAATCATGGAG AACAAAGCGTACAATCATGG	fdh	c	1767018	1767039
Primer572	AGAACAAAGCGTACAATCATGGAG AACAAAGCGTACAATCATGG	fdh		1767573	1767594
Primer417	AAGCGTACAGTTGTTTCATGGATA AGCGTACAGTTGTTTCATGGAT	23S rRNA	c	1768524	1768545
Primer418	AAGCGTACAGTTGTTTCATGGATA AGCGTACAGTTGTTTCATGGAT	23S rRNA		1768416	1768436
Primer419	GCACTTGCATCTGGATGATCTGCA CTTGCATCTGGATGATCT	23S rRNA	c	1769363	1769384

Primer420	GCACTTGCATCTGGATGATCTGCA CTTGCATCTGGATGATCT	23S rRNA		1769238	1769259
Primer421	GGGGTTTGGTAAGGGTATAGGA	23S rRNA	c	1770195	1770216
Primer422	TCAAAATCTTTAGCTGTGGCTTC	23S rRNA		1770067	1770088
Primer423	TGTTCAGAACTAATTCATGCTGCT	23S rRNA	c	1771025	1771046
Primer424	CCAATGATTAAAATACCGCCAT	23S rRNA		1770896	1770917
Primer425	AACAGGCTTGGCAAAAAGTAAT	16S rRNA	c	1771859	1771880
Primer426	TCCCCAATCGTAGTTCAAATTC	16S rRNA		1771731	1771752
Primer427	CGCTACTTGATTCAGAGCCTTT	16S rRNA	c	1772687	1772708
Primer428	AGTTGGTCCAGTATTTGCATTG	16S rRNA		1772571	1772592
Primer429	CTTAACACTGGGACATTCGTGA	pdpA2	c	1773527	1773550
Primer430	GATGTATCGCCAAGTAAAAGGC	pdpA2		1773400	1773421
Primer431	ACCGCTAGAGCTCGAAATAATG	pdpA2	c	1774475	1774498
Primer432	CTCTCTTCGGAGTAGCAAGCAT	pdpA2		1774287	1774308
Primer433	GGACAAACTGTCTACTATGCCG	pdpA2	c	1775305	1775326
Primer434	CCTAAAGTTGTCGACTTACCCG	pdpA2		1775114	1775135
Primer435	GGTGCAAGTATTTTGCTGATGA	pdpB2	c	1776167	1776188
Primer436	AAAACCTAAATAACTTTGATATTG	pdpB2		1776023	1776044
Primer437	AAGATGTTTAGACAAACTATTGAG	pdpB2	c	1777001	1777022
Primer438	GCAATTTCCATACTTCCTGTCC	pdpB2		1776897	1776918
Primer439	AAGTCATGATGCATTTGGTCTG	pdpB2	c	1777870	1777891
Primer440	TACCAACATCCAACAACCGTAA	pdpB2		1777747	1777768
Primer441	TGTTTGCCAATTTATGCAGAGT	pdpB2	c	1778695	1778718
Primer442	TCCAAGCATGAATAACATCAGG	pdpB2		1778576	1778599
Primer443	GAGGGGATGGTTTAGGATTTTC	FTT1702	c	1779539	1779561
Primer444	AGATGACAAGCATCTCAGCAAA	FTT1702		1779427	1779448
Primer445	TGAATCTAAATGCTGCAAGAGC	FTT1703	c	1780362	1780384
Primer446	ACATGTCCACTTTGCTCAACAC	FTT1703		1780226	1780248
Primer447	TGGTAAACGATTCGGTCTTCTT	FTT1703	c	1781183	1781204
Primer448	TGCATTTGCCTAACAACATTTC	FTT1703		1781055	1781078
Primer449	CCAGGCTTAAATCAACTCGTCT	FTT1704	c	1782004	1782025
Primer450	AAGCATGTTGTTGGGAATTCAT	FTT1704		1781900	1781923
Primer451	GCTTTCATATGTTAGTGGTTTAGC	FTT1705	c	1782875	1782898
Primer452	CACCATCTTGACGAACCTCATA	FTT1705		1782764	1782787
Primer453	ATGGGATCCTAAGCAAGTAGCA	FTT1705	c	1783720	1783742
Primer454	GATGGCTATTGGGAGACTTCTG	FTT1706		1783575	1783597
Primer455	CAACCACAACCTACCTCGATTA	FTT1707	c	1784553	1784574
Primer456	GGTGGAAGTTTTCCTGTCTTTG	FTT1707		1784365	1784386
Primer457	GCAGCACAGATGGCATTAAATA	FTT1707	c	1785408	1785430
Primer458	TCGAAATTTTTGCTGAAAAGGT	FTT1707		1785300	1785323
Primer459	GAACACGGCACCTAATGATGTA	FTT1708	c	1786262	1786285
Primer460	TCAGTTAGCCAGCAAGAACAAA	pdpC2		1786164	1786183
Primer461	ATGCTATGGAAATGAAGGCTGT	pdpC2	c	1787107	1787128
Primer462	TTTCATTTTCTGCCAAGGTTTT	pdpC2		1786988	1787009
Primer463	TGCCACAGTTGAAGCTTTAGAA	pdpC2	c	1787954	1787976
Primer464	TTAATGGCTGTACGAGGTAGCA	pdpC2		1787852	1787874
Primer465	AAACAATATCTCGTATGGCACG	pdpC2	c	1788808	1788831
Primer466	CGATACTGAATCAAGTCCAGCA	pdpC2		1788681	1788702
Primer467	AATTCTTAGGCATGACAAACCC	pdpC2	c	1789657	1789679
Primer468	CCGCAGTTTCAAAAATTATTTCC	pdpC2		1789534	1789557
Primer469	ATCTGAAACTTCAGCATAACT	FTT1710	c	1790483	1790504
Primer470	TCAGTCTTTTGATGGTATTGATGG	FTT1710		1790380	1790401
Primer471	ATTTATCGATTGACTCGGCATT	iglD2	c	1791355	1791376
Primer472	TTTGCTTTTATAACTGCGTGGA	iglD2		1791244	1791265
Primer473	CGCGATCAACAAGACGTAGTAG	iglC2	c	1792185	1792206
Primer474	TCCTTTCTGGAGAGTTACATGC	iglC2		1792071	1792092
Primer475	ATATCTGGATTACCAGAGGGCA	iglB2	c	1793020	1793042
Primer476	CATTCGTAGGTTCAGCATTTGT	iglB2		1792908	1792929
Primer477	CCTCAATCGGCTTAAATTGTTC	iglB2	c	1793843	1793864
Primer478	AAGCTTGAGAGTTTTTGATTTT	iglB2		1793708	1793731
Primer479	TGCAGCTGAATTTACTGATAGG	iglA2	c	1794675	1794698
Primer480	ATCAGCTGAAATAGTCCCTGCT	iglA2		1794549	1794572

Primer481	CAAAAGAGATTACAGAGTTATTTT	pdpD2	c	1795506	1795529
Primer482	TTACGCTTGAATGAGCAAAAGA	pdpD2		1795365	1795386
Primer483	CAGAAGACCTAATTGAATCGCC	pdpD2	c	1796330	1796353
Primer484	AGATATATAACAAAAGAGCTGAA	pdpD2		1796216	1796238
Primer485	GCCAATAAAAAGAAGAAAATAACC	pdpD2	c	1797187	1797208
Primer486	GCCTAAGCATGTAGATCCAAAGA	pdpD2		1797080	1797103
Primer487	TGCTATCGCTAGTGATTAGTGGA	pdpD2	c	1798032	1798053
Primer488	AAAGCAGATGTTGACAATGCAC	pdpD2		1797930	1797950
Primer489	ATTCCCAGTGGTCAGAAGTGAT	FTT1716c	c	1798899	1798920
Primer490	TCAACGTTTTACTGCCACTAGC	FTT1716c		1798789	1798810
Primer491	TAAATAATTCGGTTGCTGCTGA	FTT1716c	c	1799722	1799745
Primer492	GATAGCTACCAAATTGACAACCTT	Intergenic		1799610	1799631
Primer493	TCACACTTGGAGAAAAACCACC	FTT1717	c	1800584	1800605
Primer494	CTCGTTCAGCTCTACAACATGC	FTT1717		1800477	1800498
Primer495	CCAAAGAATGTAGAGCAAGAGC	isftu1	c	1801441	1801463
Primer496	CACCCGGTCTAAAATCTATTGC	isftu1		1801318	1801338
Primer573	CATACAATCCAAATGGTGCAGT	purL	c	1802283	1802304
Primer574	ACTCAGGATAATCCCCACAGAA	abcZ		1817368	1817389
Primer575	TAGTTTCACGTATTCACCGTGG	FTT1730c	c	1818246	1818267
Primer576	CAAATATGATCTTTTGGAGCGG	FTT1741c		1831541	1831562
Primer577	AGCCAAGTAGTAAGCCCAACAC	FTT1743	c	1832798	1832819
Primer578	TTTACCGGCTCTACAAGAGTTG	gabD2		1844468	1844489
Primer579	AAGTTATGTTAGGGTAGTTTTGT	Intergenic	c	1845645	1845667
Primer580	ATGCCAGCCTTAAGTGGTTAGA	FTT1761		1849769	1849790
Primer581	GCAATGAAATTGGCAACTGTTA	FTT1762c	c	1850709	1850730
Primer582	TTTAGCTTTGGATACGGTTGGT	FTT1779		1868481	1868502
Primer583	CATTTCTTTCAAAACCTTTGGC	Intergenic	c	1869759	1869780
Primer584	TTCCAAAATTGGATAATTGGTTC	FTT1783		1873321	1873343
Primer585	GGCTACCAGGAGGGTTAGAAAA	FTT1786	c	1874298	1874319
Primer586	ATTACTTGGTTGGGAGATGCTG	FTT1786		1874463	1874484
Primer587	TGGTGTTGCAAGTTTAGTTTGG	FTT1787c	c	1875385	1875406
Primer588	AAAATTTGGGGCATCTGTATTG	iciA		1876086	1876107
Primer589	CTCAATAGCATCCAGCCTTTCT	Intergenic	c	1877014	1877035
Primer590	TGTATCTTTTGGCGATTTAGCA	Intergenic		1876928	1876949
Primer591	TTCTGAGGAAATGATGCTGTTG	FTT1789	c	1877868	1877889
Primer592	AGATAGTATTTCTGGTGGGGCA	FTT1791		1878601	1878622
Primer593	GTTTAGGATACGCTTTCATGGC	pepN	c	1880040	1880061
Primer594	GTTCAAGGTTACGTTCACCCAT	trpC		1883679	1883700
Primer595	AGGTGAGTTACCTGATGCGATT	FTT1796c	c	1884628	1884649
Primer596	CCCAAATTTGTTGCCACTTATC	FTT1796c		1884603	1884624
Primer597	AGCTATATTTGCAGGTGGTGGT	msrA2	c	1885511	1885532
Primer598	TAGAATCTAAAACCACGCCTGC	trpE		1888656	1888677
Primer599	TGATCATCAAGAGCAGAAAACG	trpE	c	1889545	1889566
Primer600	TTTCAAAGCTAGCAACAACATCA	trpE		1889640	1889662
Primer601	GTTTTCCCCTTTTTACTTCTGA	Intergenic	c	1890618	1890639
Primer602	TACAATGTTCACATGGCTGACA	rng		1891141	1891162
Primer603	TGGTTACGTGAAGGTCAAGAAG	rng	c	1892062	1892083

* C strand represents complementary strand.

REFERENCES

Anthony, L.S., Cowley, S.C., Mdluli, K.E., and Nano, F.E. (1994) Isolation of a Francisella tularensis mutant that is sensitive to serum and oxidative killing and is avirulent in mice: correlation with the loss of MinD homologue expression. *FEMS Microbiol Lett* **124**: 157-165.

Barnes, W.M. (1994) PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proc Natl Acad Sci U S A* **91**: 2216-2220.

Baron, C., D, O.C., and Lanka, E. (2002) Bacterial secrets of secretion: EuroConference on the biology of type IV secretion processes. *Mol Microbiol* **43**: 1359-1365.

Baron, G.S., and Nano, F.E. (1998) MglA and MglB are required for the intramacrophage growth of Francisella novicida. *Mol Microbiol* **29**: 247-259.

Bartel, P.L., Chien, C.-T., Sternglanz, R., and Fields, S. (1993) Using the two-hybrid system to detect protein-protein interactions. In *Cellular Interactions in Development: A Practical Approach*. Hartley, D.A. (Oxford University Press, Oxford) pp. 153–179.

Behr, M.A., Wilson, M.A., Gill, W.P., Salamon, H., Schoolnik, G.K., Rane, S., and Small, P.M. (1999) Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284**: 1520-1523.

Ben Zakour, N., Gautier, M., Andonov, R., Lavenier, D., Cochet, M.F., Veber, P., Sorokin, A., and Le Loir, Y. (2004) GenoFrag: software to design primers optimized for whole genome scanning by long-range PCR amplification. *Nucleic Acids Res* **32**: 17-24.

Binnewies, T.T., Motro, Y., Hallin, P.F., Lund, O., Dunn, D., La, T., Hampson, D.J., Bellgard, M., Wassenaar, T.M., and Ussery, D.W. (2006) Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Funct Integr Genomics* **6**: 165-185.

Bladergroen, M.R., Badelt, K., and Spaink, H.P. (2003) Infection-blocking genes of a symbiotic Rhizobium leguminosarum strain that are involved in temperature-dependent protein secretion. *Mol Plant Microbe Interact* **16**: 53-64.

Blanc-Potard, A.B., and Groisman, E.A. (1997) The Salmonella selC locus contains a pathogenicity island mediating intramacrophage survival. *Embo J* **16**: 5376-5385.

Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., and Shao, Y. (1997) The complete genome sequence of Escherichia coli K-12. *Science* **277**: 1453-1474.

Blocker, A., Jouihri, N., Larquet, E., Gounon, P., Ebel, F., Parsot, C., Sansonetti, P., and Allaoui, A. (2001) Structure and composition of the Shigella flexneri "needle complex", a part of its type III secreton. *Mol Microbiol* **39**: 652-663.

Blocker, A., Komoriya, K., and Aizawa, S. (2003) Type III secretion systems and bacterial flagella: insights into their function from structural similarities. *Proc Natl Acad Sci U S A* **100**: 3027-3030.

Boddicker, J.D., Ledeboer, N.A., Jagnow, J., Jones, B.D., and Clegg, S. (2002) Differential binding to and biofilm formation on, HEp-2 cells by Salmonella enterica serovar Typhimurium is dependent upon allelic variation in the fimH gene of the fim gene cluster. *Mol Microbiol* **45**: 1255-1265.

Brotcke, A., Weiss, D.S., Kim, C.C., Chain, P., Malfatti, S., Garcia, E., and Monack, D.M. (2006) Identification of MglA-regulated genes reveals novel virulence factors in Francisella tularensis. *Infect Immun* **74**: 6642-6655.

Brown, J.R. (2008) *Comparative Genomics: Basic and Applied Research.* Boca Raton: CRC Press, Taylor & Francis Group.

Butcher, P.D. (2004) Microarrays for Mycobacterium tuberculosis. *Tuberculosis (Edinb)* **84**: 131-137.

Carlson, S.A., and Jones, B.D. (1998) Inhibition of Salmonella typhimurium invasion by host cell expression of secreted bacterial invasion proteins. *Infect Immun* **66**: 5295-5300.

Casjens, S. (2003) Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol* **49**: 277-300.

Chaudhuri, R.R., Khan, A.M., and Pallen, M.J. (2004) coliBASE: an online database for Escherichia coli, Shigella and Salmonella comparative genomics. *Nucleic Acids Res* **32**: D296-299.

Chaudhuri, R.R., and Pallen, M.J. (2006) xBASE, a collection of online databases for bacterial comparative genomics. *Nucleic Acids Res* **34**: D335-337.

Chaudhuri, R.R., Ren, C.P., Desmond, L., Vincent, G.A., Silman, N.J., Brehm, J.K., Elmore, M.J., Hudson, M.J., Forsman, M., Isherwood, K.E., Gurycova, D., Minton, N.P., Titball, R.W., Pallen, M.J., and Vipond, R. (2007) Genome sequencing shows that European isolates of Francisella tularensis subspecies tularensis are almost identical to US laboratory strain Schu S4. *PLoS ONE* **2**: e352.

Chaudhuri, R.R., Loman, N.J., Snyder, L.A., Bailey, C.M., Stekel, D.J., and Pallen, M.J. (2008) xBASE2: a comprehensive resource for comparative bacterial genomics. *Nucleic Acids Res* **36**: D543-546.

Cheng, S., Fockler, C., Barnes, W.M., and Higuchi, R. (1994) Effective amplification of long targets from cloned inserts and human genomic DNA. *Proc Natl Acad Sci U S A* **91**: 5695-5699.

Christie, G.E., and Calendar, R. (1990) Interactions between satellite bacteriophage P4 and its helpers. *Annu Rev Genet* **24**: 465-490.

Christie, P.J., and Vogel, J.P. (2000) Bacterial type IV secretion: conjugation systems adapted to deliver effector molecules to host cells. *Trends Microbiol* **8**: 354-360.

Christie, P.J., Atmakuri, K., Krishnamoorthy, V., Jakubowski, S., and Cascales, E. (2005) Biogenesis, architecture, and function of bacterial type IV secretion systems. *Annu Rev Microbiol* **59**: 451-485.

Christopher, G.W., Agan, M.B., Cieslak, T.J., and Olson, P.E. (2005) History of U.S. military contributions to the study of bacterial zoonoses. *Mil Med* **170**: 39-48.

Cornelis, G.R., and Wolf-Watz, H. (1997) The Yersinia Yop virulon: a bacterial system for subverting eukaryotic cells. *Mol Microbiol* **23**: 861-867.

Crago, A.M., and Koronakis, V. (1998) Salmonella InvG forms a ring-like multimer that requires the InvH lipoprotein for outer membrane localization. *Mol Microbiol* **30**: 47-56.

Creasey, E.A., Delahay, R.M., Daniell, S.J., and Frankel, G. (2003) Yeast two-hybrid system survey of interactions between LEE-encoded proteins of enteropathogenic Escherichia coli. *Microbiology* **149**: 2093-2106.

Daefler, S., and Russel, M. (1998) The Salmonella typhimurium InvH protein is an outer membrane lipoprotein required for the proper localization of InvG. *Mol Microbiol* **28**: 1367-1380.

Dahlstrand, S., Ringertz, O., and Zetterberg, B. (1971) Airborne tularemia in Sweden. *Scand J Infect Dis* **3**: 7-16.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23**: 324-328.

Darwin, K.H., and Miller, V.L. (2000) The putative invasion protein chaperone SicA acts together with InvF to activate the expression of Salmonella typhimurium virulence genes. *Mol Microbiol* **35**: 949-960.

Das, S., and Chaudhuri, K. (2003) Identification of a unique IAHP (IcmF associated homologous proteins) cluster in Vibrio cholerae and other proteobacteria through in silico analysis. *In Silico Biol* **3**: 287-300.

de Bruin, O.M., Ludu, J.S., and Nano, F.E. (2007) The Francisella pathogenicity island protein IglA localizes to the bacterial cytoplasm and is needed for intracellular growth. *BMC Microbiol* **7**: 1.

Delepelaire, P. (2004) Type I secretion in gram-negative bacteria. *Biochim Biophys Acta* **1694**: 149-161.

Deng, W., Li, Y., Vallance, B.A., and Finlay, B.B. (2001) Locus of enterocyte effacement from Citrobacter rodentium: sequence analysis and evidence for horizontal transfer among attaching and effacing pathogens. *Infect Immun* **69**: 6323-6335.

Dennis, D.T., Inglesby, T.V., Henderson, D.A., Bartlett, J.G., Ascher, M.S., Eitzen, E., Fine, A.D., Friedlander, A.M., Hauer, J., Layton, M., Lillibridge, S.R., McDade, J.E., Osterholm, M.T., O'Toole, T., Parker, G., Perl, T.M., Russell, P.K., and Tonat,

K. (2001) Tularemia as a biological weapon: medical and public health management. *Jama* **285**: 2763-2773.

Dobrindt, U., and Hacker, J. (2001) Whole genome plasticity in pathogenic bacteria. *Curr Opin Microbiol* **4**: 550-557.

Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* **2**: 414-424.

Donnenberg, M.S., and Whittam, T.S. (2001) Pathogenesis and evolution of virulence in enteropathogenic and enterohemorrhagic Escherichia coli. *J Clin Invest* **107**: 539-548.

Donnenberg, M.S. (2002) *Escherichia coli: Virulence Mechanisms of a Versatile Pathogen*: Academic Press, Amsterdam, The Netherlands.

Dorrell, N., Mangan, J.A., Laing, K.G., Hinds, J., Linton, D., Al-Ghusein, H., Barrell, B.G., Parkhill, J., Stoker, N.G., Karlyshev, A.V., Butcher, P.D., and Wren, B.W. (2001) Whole genome comparison of Campylobacter jejuni human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Res* **11**: 1706-1715.

Elliott, S.J., Hutcheson, S.W., Dubois, M.S., Mellies, J.L., Wainwright, L.A., Batchelor, M., Frankel, G., Knutton, S., and Kaper, J.B. (1999) Identification of CesT, a chaperone for the type III secretion of Tir in enteropathogenic Escherichia coli. *Mol Microbiol* **33**: 1176-1189.

Ellis, J., Oyston, P.C., Green, M., and Titball, R.W. (2002) Tularemia. *Clin Microbiol Rev* **15**: 631-646.

Escobar-Paramo, P., Clermont, O., Blanc-Potard, A.B., Bui, H., Le Bouguenec, C., and Denamur, E. (2004) A specific genetic background is required for acquisition and expression of virulence factors in Escherichia coli. *Mol Biol Evol* **21**: 1085-1094.

Farlow, J., Wagner, D.M., Dukerich, M., Stanley, M., Chu, M., Kubota, K., Petersen, J., and Keim, P. (2005) Francisella tularensis in the United States. *Emerg Infect Dis* **11**: 1835-1841.

Filloux, A., Hachani, A., and Bleves, S. (2008) The bacterial type VI secretion machine: yet another player for protein transport across membranes. *Microbiology* **154**: 1570-1583.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., and et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**: 496-512.

Folkesson, A., Lofdahl, S., and Normark, S. (2002) The Salmonella enterica subspecies I specific centisome 7 genomic island encodes novel protein families present in bacteria living in close contact with eukaryotic cells. *Res Microbiol* **153**: 537-545.

Forsman, M., Sandstrom, G., and Jaurin, B. (1990) Identification of Francisella species and discrimination of type A and type B strains of F. tularensis by 16S rRNA analysis. *Appl Environ Microbiol* **56**: 949-955.

Fouts, D.E., Mongodin, E.F., Mandrell, R.E., Miller, W.G., Rasko, D.A., Ravel, J., Brinkac, L.M., DeBoy, R.T., Parker, C.T., Daugherty, S.C., Dodson, R.J., Durkin, A.S., Madupu, R., Sullivan, S.A., Shetty, J.U., Ayodeji, M.A., Shvartsbeyn, A., Schatz, M.C., Badger, J.H., Fraser, C.M., and Nelson, K.E. (2005) Major structural differences and novel potential virulence mechanisms from the genomes of multiple campylobacter species. *PLoS Biol* **3**: e15.

Francetic, O., and Pugsley, A.P. (2005) Towards the identification of type II secretion signals in a nonacylated variant of pullulanase from Klebsiella oxytoca. *J Bacteriol* **187**: 7045-7055.

Frankel, G., Phillips, A.D., Rosenshine, I., Dougan, G., Kaper, J.B., and Knutton, S. (1998) Enteropathogenic and enterohaemorrhagic Escherichia coli: more subversive elements. *Mol Microbiol* **30**: 911-921.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, R.D., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Saudek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J.F., Dougherty, B.A., Bott, K.F., Hu, P.C., Lucier, T.S., Peterson, S.N., Smith, H.O., Hutchison, C.A., 3rd, and Venter, J.C. (1995) The minimal gene complement of Mycoplasma genitalium. *Science* **270**: 397-403.

Fraser, C.M., Eisen, J., Fleischmann, R.D., Ketchum, K.A., and Peterson, S. (2000) Comparative genomics and understanding of microbial biology. *Emerg Infect Dis* **6**: 505-512.

Fraser, C.M., Eisen, J.A., Nelson, K.E., Paulsen, I.T., and Salzberg, S.L. (2002) The value of complete microbial genome sequencing (you get what you pay for). *J Bacteriol* **184**: 6403-6405; discusion 6405.

Fraser, G.M., and Hughes, C. (1999) Swarming motility. *Curr Opin Microbiol* **2**: 630-635.

Freeman, V.J. (1951) Studies on the virulence of bacteriophage-infected strains of Corynebacterium diphtheriae. *J Bacteriol* **61**: 675-688.

Frenay, H.M., Bunschoten, A.E., Schouls, L.M., van Leeuwen, W.J., Vandenbroucke-Grauls, C.M., Verhoef, J., and Mooi, F.R. (1996) Molecular typing of methicillin-

resistant Staphylococcus aureus on the basis of protein A gene polymorphism. *Eur J Clin Microbiol Infect Dis* **15**: 60-64.

Galan, J.E. (1999) Interaction of Salmonella with host cells through the centisome 63 type III secretion system. *Curr Opin Microbiol* **2**: 46-50.

Galan, J.E., and Collmer, A. (1999) Type III secretion machines: bacterial devices for protein delivery into host cells. *Science* **284**: 1322-1328.

Geitz, R.D., and Schiestl, R.H. (1995) Transforming yeast with DNA. *Methods Mol Cell Biol* **5**: 255-269.

Gentschev, I., Dietrich, G., and Goebel, W. (2002) The E. coli alpha-hemolysin secretion system and its use in vaccine development. *Trends Microbiol* **10**: 39-45.

Ghosh, P. (2004) Process of protein transport by the type III secretion system. *Microbiol Mol Biol Rev* **68**: 771-795.

Gliatto, J.M., Rae, J.F., McDonough, P.L., and Dasbach, J.J. (1994) Feline tularemia on Nantucket Island, Massachusetts. *J Vet Diagn Invest* **6**: 102-105.

Golovliov, I., Ericsson, M., Sandstrom, G., Tarnvik, A., and Sjostedt, A. (1997) Identification of proteins of Francisella tularensis induced during growth in macrophages and cloning of the gene encoding a prominently induced 23-kilodalton protein. *Infect Immun* **65**: 2183-2189.

Golovliov, I., Baranov, V., Krocova, Z., Kovarova, H., and Sjostedt, A. (2003a) An attenuated strain of the facultative intracellular bacterium Francisella tularensis can escape the phagosome of monocytic cells. *Infect Immun* **71**: 5940-5950.

Golovliov, I., Sjostedt, A., Mokrievich, A., and Pavlov, V. (2003b) A method for allelic replacement in Francisella tularensis. *FEMS Microbiol Lett* **222**: 273-280.

Gray, C.G., Cowley, S.C., Cheung, K.K., and Nano, F.E. (2002) The identification of five genetic loci of Francisella novicida associated with intracellular growth. *FEMS Microbiol Lett* **215**: 53-56.

Gresham, D., Dunham, M.J., and Botstein, D. (2008) Comparing whole genomes using DNA microarrays. *Nat Rev Genet* **9**: 291-302.

Gundogdu, O., Bentley, S.D., Holden, M.T., Parkhill, J., Dorrell, N., and Wren, B.W. (2007) Re-annotation and re-analysis of the Campylobacter jejuni NCTC11168 genome sequence. *BMC Genomics* **8**: 162.

Gurycova, D. (1998) First isolation of Francisella tularensis subsp. tularensis in Europe. *Eur J Epidemiol* **14**: 797-802.

Gutacker, M.M., Smoot, J.C., Migliaccio, C.A., Ricklefs, S.M., Hua, S., Cousins, D.V., Graviss, E.A., Shashkina, E., Kreiswirth, B.N., and Musser, J.M. (2002) Genome-wide analysis of synonymous single nucleotide polymorphisms in Mycobacterium tuberculosis complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics* **162**: 1533-1543.

Hacker, J., Blum-Oehler, G., Muhldorfer, I., and Tschape, H. (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol Microbiol* **23**: 1089-1097.

Hacker, J., and Kaper, J.B. (2000) Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol* **54**: 641-679.

Hacker, J., Blum-Oehler, G., Hochhut, B., and Dobrindt, U. (2003) The molecular basis of infectious diseases: pathogenicity islands and other mobile genetic elements. A review. *Acta Microbiol Immunol Hung* **50**: 321-330.

Hall, N. (2007) Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol* **210**: 1518-1525.

Hansen-Wester, I., and Hensel, M. (2001) Salmonella pathogenicity islands encoding type III secretion systems. *Microbes Infect* **3**: 549-559.

Harris, S. (1992) Japanese biological warfare research on humans: a case study of microbiology and ethics. *Ann N Y Acad Sci* **666**: 21-52.

Hartleib, S., Prager, R., Hedenstrom, I., Lofdahl, S., and Tschape, H. (2003) Prevalence of the new, SPI1-like, pathogenicity island ETT2 among Escherichia coli. *Int J Med Microbiol* **292**: 487-493.

Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., Ohtsubo, E., Baba, T., Wanner, B.L., Mori, H., and Horiuchi, T. (2006) Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. *Mol Syst Biol* **2**: 2006 0007.

Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C.G., Ohtsubo, E., Nakayama, K., Murata, T., Tanaka, M., Tobe, T., Iida, T., Takami, H., Honda, T., Sasakawa, C., Ogasawara, N., Yasunaga, T., Kuhara, S., Shiba, T., Hattori, M., and Shinagawa, H. (2001) Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* **8**: 11-22.

Henderson, I.R., Navarro-Garcia, F., Desvaux, M., Fernandez, R.C., and Ala'Aldeen, D. (2004) Type V protein secretion pathway: the autotransporter story. *Microbiol Mol Biol Rev* **68**: 692-744.

Hendrix, R.W. (2003) Bacteriophage genomics. *Curr Opin Microbiol* **6**: 506-511.

Herzer, P.J., Inouye, S., Inouye, M., and Whittam, T.S. (1990) Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of Escherichia coli. *J Bacteriol* **172**: 6175-6181.

Hofreuter, D., Tsai, J., Watson, R.O., Novik, V., Altman, B., Benitez, M., Clark, C., Perbost, C., Jarvie, T., Du, L., and Galan, J.E. (2006) Unique features of a highly pathogenic Campylobacter jejuni strain. *Infect Immun* **74**: 4694-4707.

Holland, I.B., Schmitt, L., and Young, J. (2005) Type 1 protein secretion in bacteria, the ABC-transporter dependent pathway (review). *Mol Membr Biol* **22**: 29-39.

Hollis, D.G., Weaver, R.E., Steigerwalt, A.G., Wenger, J.D., Moss, C.W., and Brenner, D.J. (1989) Francisella philomiragia comb. nov. (formerly Yersinia philomiragia) and Francisella tularensis biogroup novicida (formerly Francisella novicida) associated with human disease. *J Clin Microbiol* **27**: 1601-1608.

Hornbeck, P., Winston, S.E., and Fuller, S.A. (2001) Enzyme-linked immunosorbent assays (ELISA). *Curr Protoc Mol Biol* **Chapter 11**: Unit11 12.

Hueck, C.J. (1998) Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiol Mol Biol Rev* **62**: 379-433.

Ide, T., Laarmann, S., Greune, L., Schillers, H., Oberleithner, H., and Schmidt, M.A. (2001) Characterization of translocation pores inserted into plasma membranes by type III-secreted Esp proteins of enteropathogenic Escherichia coli. *Cell Microbiol* **3**: 669-679.

Johansson, A., Ibrahim, A., Goransson, I., Eriksson, U., Gurycova, D., Clarridge, J.E., 3rd, and Sjostedt, A. (2000) Evaluation of PCR-based methods for discrimination of Francisella species and subspecies and development of a specific PCR that

distinguishes the two major subspecies of Francisella tularensis. *J Clin Microbiol* **38**: 4180-4185.

Johansson, A., Farlow, J., Larsson, P., Dukerich, M., Chambers, E., Bystrom, M., Fox, J., Chu, M., Forsman, M., Sjostedt, A., and Keim, P. (2004) Worldwide genetic relationships among Francisella tularensis isolates determined by multiple-locus variable-number tandem repeat analysis. *J Bacteriol* **186**: 5808-5818.

Johnson, J.R., and Russo, T.A. (2002) Extraintestinal pathogenic Escherichia coli: "the other bad E coli". *J Lab Clin Med* **139**: 155-162.

Jores, J., Rumer, L., and Wieler, L.H. (2004) Impact of the locus of enterocyte effacement pathogenicity island on the evolution of pathogenic Escherichia coli. *Int J Med Microbiol* **294**: 103-113.

Juhas, M., Crook, D.W., and Hood, D.W. (2008) Type IV secretion systems: tools of bacterial horizontal gene transfer and virulence. *Cell Microbiol* **10**: 2377-2386.

Kaper, J.B., Nataro, J.P., and Mobley, H.L. (2004) Pathogenic Escherichia coli. *Nat Rev Microbiol* **2**: 123-140.

Karlyshev, A.V., Pallen, M.J., and Wren, B.W. (2000) Single-primer PCR procedure for rapid identification of transposon insertion sites. *Biotechniques* **28**: 1078, 1080, 1082.

Keim, P., Price, L.B., Klevytska, A.M., Smith, K.L., Schupp, J.M., Okinaka, R., Jackson, P.J., and Hugh-Jones, M.E. (2000) Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within Bacillus anthracis. *J Bacteriol* **182**: 2928-2936.

Kim, Y.K., and McCarter, L.L. (2004) Cross-regulation in Vibrio parahaemolyticus: compensatory activation of polar flagellar genes by the lateral flagellar regulator LafK. *J Bacteriol* **186**: 4014-4018.

Kimbrough, T.G., and Miller, S.I. (2000) Contribution of Salmonella typhimurium type III secretion components to needle complex formation. *Proc Natl Acad Sci U S A* **97**: 11008-11013.

Kimura, F. (1983) [Study on the central neural mechanism for gonadotropin secretory rhythm]. *Horumon To Rinsho* **31**: 281-293.

Klein, J.R., Fahlen, T.F., and Jones, B.D. (2000) Transcriptional organization and function of invasion genes within Salmonella enterica serovar Typhimurium pathogenicity island 1, including the prgH, prgI, prgJ, prgK, orgA, orgB, and orgC genes. *Infect Immun* **68**: 3368-3376.

Kubori, T., Matsushima, Y., Nakamura, D., Uralil, J., Lara-Tejero, M., Sukhan, A., Galan, J.E., and Aizawa, S.I. (1998) Supramolecular structure of the Salmonella typhimurium type III protein secretion system. *Science* **280**: 602-605.

Kubori, T., Sukhan, A., Aizawa, S.I., and Galan, J.E. (2000) Molecular characterization and assembly of the needle complex of the Salmonella typhimurium type III protein secretion system. *Proc Natl Acad Sci U S A* **97**: 10225-10230.

Laemmli, U.K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**: 680-685.

Lai, X.H., Golovliov, I., and Sjostedt, A. (2004) Expression of IglC is necessary for intracellular growth and induction of apoptosis in murine macrophages by Francisella tularensis. *Microb Pathog* **37**: 225-230.

Larsson, P., Oyston, P.C., Chain, P., Chu, M.C., Duffield, M., Fuxelius, H.H., Garcia, E., Halltorp, G., Johansson, D., Isherwood, K.E., Karp, P.D., Larsson, E., Liu, Y., Michell, S., Prior, J., Prior, R., Malfatti, S., Sjostedt, A., Svensson, K., Thompson, N., Vergez, L., Wagg, J.K., Wren, B.W., Lindler, L.E., Andersson, S.G., Forsman, M., and Titball, R.W. (2005) The complete genome sequence of Francisella tularensis, the causative agent of tularemia. *Nat Genet* **37**: 153-159.

Lauriano, C.M., Barker, J.R., Yoon, S.S., Nano, F.E., Arulanandam, B.P., Hassett, D.J., and Klose, K.E. (2004) MglA regulates transcription of virulence factors necessary for Francisella tularensis intraamoebae and intramacrophage survival. *Proc Natl Acad Sci U S A* **101**: 4246-4249.

Lawrence, J.G., Hendrix, R.W., and Casjens, S. (2001) Where are the pseudogenes in bacterial genomes? *Trends Microbiol* **9**: 535-540.

Le Fleche, P., Hauck, Y., Onteniente, L., Prieur, A., Denoeud, F., Ramisse, V., Sylvestre, P., Benson, G., Ramisse, F., and Vergnaud, G. (2001) A tandem repeats database for bacterial genomes: application to the genotyping of Yersinia pestis and Bacillus anthracis. *BMC Microbiol* **1**: 2.

Lecointre, G., Rachdi, L., Darlu, P., and Denamur, E. (1998) Escherichia coli molecular phylogeny using the incongruence length difference test. *Mol Biol Evol* **15**: 1685-1695.

Lindgren, H., Golovliov, I., Baranov, V., Ernst, R.K., Telepnev, M., and Sjostedt, A. (2004) Factors affecting the escape of Francisella tularensis from the phagolysosome. *J Med Microbiol* **53**: 953-958.

Lindsay, J.A., Moore, C.E., Day, N.P., Peacock, S.J., Witney, A.A., Stabler, R.A., Husain, S.E., Butcher, P.D., and Hinds, J. (2006) Microarrays reveal that each of the ten

dominant lineages of Staphylococcus aureus has a unique combination of surface-associated and regulatory genes. *J Bacteriol* **188**: 669-676.

Lindstedt, B.A. (2005) Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria. *Electrophoresis* **26**: 2567-2582.

Liu, S.L., and Sanderson, K.E. (1995) Genomic cleavage map of Salmonella typhi Ty2. *J Bacteriol* **177**: 5099-5107.

Lizardi, P.M. (2008) Next-generation sequencing-by-hybridization. *Nat Biotechnol* **26**: 649-650.

Lostroh, C.P., and Lee, C.A. (2001) The Salmonella pathogenicity island-1 type III secretion system. *Microbes Infect* **3**: 1281-1291.

Ludu, J.S., de Bruin, O.M., Duplantis, B.N., Schmerk, C.L., Chou, A.Y., Elkins, K.L., and Nano, F.E. (2008) The Francisella pathogenicity island protein PdpD is required for full virulence and associates with homologues of the type VI secretion system. *J Bacteriol* **190**: 4584-4595.

Macnab, R.M. (2003) How bacteria assemble flagella. *Annu Rev Microbiol* **57**: 77-100.

Mahillon, J., and Chandler, M. (1998) Insertion sequences. *Microbiol Mol Biol Rev* **62**: 725-774.

Mahillon, J., Leonard, C., and Chandler, M. (1999) IS elements as constituents of bacterial genomes. *Res Microbiol* **150**: 675-687.

Makino, S., Tobe, T., Asakura, H., Watarai, M., Ikeda, T., Takeshi, K., and Sasakawa, C. (2003) Distribution of the secondary type III secretion system locus found in enterohemorrhagic Escherichia coli O157:H7 isolates among Shiga toxin-producing E. coli strains. *J Clin Microbiol* **41**: 2341-2347.

Mardis, E.R. (2008a) The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**: 133-141.

Mardis, E.R. (2008b) Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9**: 387-402.

Marguerat, S., Wilhelm, B.T., and Bahler, J. (2008) Next-generation sequencing: applications beyond genomes. *Biochem Soc Trans* **36**: 1091-1096.

Mazars, E., Lesjean, S., Banuls, A.L., Gilbert, M., Vincent, V., Gicquel, B., Tibayrenc, M., Locht, C., and Supply, P. (2001) High-resolution minisatellite-based typing as a portable approach to global analysis of Mycobacterium tuberculosis molecular epidemiology. *Proc Natl Acad Sci U S A* **98**: 1901-1906.

McClelland, M., Florea, L., Sanderson, K., Clifton, S.W., Parkhill, J., Churcher, C., Dougan, G., Wilson, R.K., and Miller, W. (2000) Comparison of the Escherichia coli K-12 genome with sampled genomes of a Klebsiella pneumoniae and three salmonella enterica serovars, Typhimurium, Typhi and Paratyphi. *Nucleic Acids Res* **28**: 4974-4986.

Miller, J.H., (ed) (1972) *Experiments in molecular genetics*: Cold Spring Harbor, N.Y.

Miller, V.L., and Mekalanos, J.J. (1984) Synthesis of cholera toxin is positively regulated at the transcriptional level by toxR. *Proc Natl Acad Sci U S A* **81**: 3471-3475.

Morozova, O., and Marra, M.A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics*.

Mougous, J.D., Cuff, M.E., Raunser, S., Shen, A., Zhou, M., Gifford, C.A., Goodman, A.L., Joachimiak, G., Ordonez, C.L., Lory, S., Walz, T., Joachimiak, A., and Mekalanos, J.J. (2006) A virulence locus of Pseudomonas aeruginosa encodes a protein secretion apparatus. *Science* **312**: 1526-1530.

Mushegian, A.R., and Koonin, E.V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A* **93**: 10268-10273.

Nachamkin, I., Allos, B.M., and Ho, T. (1998) Campylobacter species and Guillain-Barre syndrome. *Clin Microbiol Rev* **11**: 555-567.

Nano, F.E., Zhang, N., Cowley, S.C., Klose, K.E., Cheung, K.K., Roberts, M.J., Ludu, J.S., Letendre, G.W., Meierovics, A.I., Stephens, G., and Elkins, K.L. (2004) A Francisella tularensis pathogenicity island required for intramacrophage growth. *J Bacteriol* **186**: 6430-6436.

Nano, F.E., and Schmerk, C. (2007) The Francisella pathogenicity island. *Ann N Y Acad Sci* **1105**: 122-137.

Nataro, J.P., and Kaper, J.B. (1998) Diarrheagenic Escherichia coli. *Clin Microbiol Rev* **11**: 142-201.

Newell, D.G., Shreeve, J.E., Toszeghy, M., Domingue, G., Bull, S., Humphrey, T., and Mead, G. (2001) Changes in the carriage of Campylobacter strains by poultry carcasses during processing in abattoirs. *Appl Environ Microbiol* **67**: 2636-2640.

Ochman, H., and Selander, R.K. (1984) Standard reference strains of Escherichia coli from natural populations. *J Bacteriol* **157**: 690-693.

Ochman, H., and Jones, I.B. (2000) Evolutionary dynamics of full genome content in Escherichia coli. *Embo J* **19**: 6637-6643.

Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299-304.

Ohnishi, M., Kurokawa, K., and Hayashi, T. (2001) Diversification of Escherichia coli genomes: are bacteriophages the major contributors? *Trends Microbiol* **9**: 481-485.

Ohnishi, M., Terajima, J., Kurokawa, K., Nakayama, K., Murata, T., Tamura, K., Ogura, Y., Watanabe, H., and Hayashi, T. (2002) Genomic diversity of enterohemorrhagic Escherichia coli O157 revealed by whole genome PCR scanning. *Proc Natl Acad Sci U S A* **99**: 17043-17048.

Okabe, M., Yakushi, T., Kojima, M., and Homma, M. (2002) MotX and MotY, specific components of the sodium-driven flagellar motor, colocalize to the outer membrane in Vibrio alginolyticus. *Mol Microbiol* **46**: 125-134.

Ou, H.Y., Chen, L.L., Lonnen, J., Chaudhuri, R.R., Thani, A.B., Smith, R., Garton, N.J., Hinton, J., Pallen, M., Barer, M.R., and Rajakumar, K. (2006) A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria. *Nucleic Acids Res* **34**: e3.

Oyston, P.C., Sjostedt, A., and Titball, R.W. (2004) Tularaemia: bioterrorism defence renews interest in Francisella tularensis. *Nat Rev Microbiol* **2**: 967-978.

Oyston, P.C. (2008) Francisella tularensis: unravelling the secrets of an intracellular pathogen. *J Med Microbiol* **57**: 921-930.

Page, A.L., and Parsot, C. (2002) Chaperones of the type III secretion pathway: jacks of all trades. *Mol Microbiol* **46**: 1-11.

Pallen, M.J., Beatson, S.A., and Bailey, C.M. (2005) Bioinformatics, genomics and evolution of non-flagellar type-III secretion systems: a Darwinian perspective. *FEMS Microbiol Rev* **29**: 201-229.

Pallen, M.J., and Wren, B.W. (2007) Bacterial pathogenomics. *Nature* **449**: 835-842.

Parkhill, J., Wren, B.W., Mungall, K., Ketley, J.M., Churcher, C., Basham, D., Chillingworth, T., Davies, R.M., Feltwell, T., Holroyd, S., Jagels, K., Karlyshev, A.V., Moule, S., Pallen, M.J., Penn, C.W., Quail, M.A., Rajandream, M.A.,

Rutherford, K.M., van Vliet, A.H., Whitehead, S., and Barrell, B.G. (2000) The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences. *Nature* **403**: 665-668.

Parkhill, J., Wren, B.W., Thomson, N.R., Titball, R.W., Holden, M.T., Prentice, M.B., Sebaihia, M., James, K.D., Churcher, C., Mungall, K.L., Baker, S., Basham, D., Bentley, S.D., Brooks, K., Cerdeno-Tarraga, A.M., Chillingworth, T., Cronin, A., Davies, R.M., Davis, P., Dougan, G., Feltwell, T., Hamlin, N., Holroyd, S., Jagels, K., Karlyshev, A.V., Leather, S., Moule, S., Oyston, P.C., Quail, M., Rutherford, K., Simmonds, M., Skelton, J., Stevens, K., Whitehead, S., and Barrell, B.G. (2001) Genome sequence of Yersinia pestis, the causative agent of plague. *Nature* **413**: 523-527.

Parkhill, J., Sebaihia, M., Preston, A., Murphy, L.D., Thomson, N., Harris, D.E., Holden, M.T., Churcher, C.M., Bentley, S.D., Mungall, K.L., Cerdeno-Tarraga, A.M., Temple, L., James, K., Harris, B., Quail, M.A., Achtman, M., Atkin, R., Baker, S., Basham, D., Bason, N., Cherevach, I., Chillingworth, T., Collins, M., Cronin, A., Davis, P., Doggett, J., Feltwell, T., Goble, A., Hamlin, N., Hauser, H., Holroyd, S., Jagels, K., Leather, S., Moule, S., Norberczak, H., O'Neil, S., Ormond, D., Price, C., Rabbinowitsch, E., Rutter, S., Sanders, M., Saunders, D., Seeger, K., Sharp, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Unwin, L., Whitehead, S., Barrell, B.G., and Maskell, D.J. (2003) Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. *Nat Genet* **35**: 32-40.

Perna, N.T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., Posfai, G., Hackett, J., Klink, S.,

Boutin, A., Shao, Y., Miller, L., Grotbeck, E.J., Davis, N.W., Lim, A., Dimalanta, E.T., Potamousis, K.D., Apodaca, J., Anantharaman, T.S., Lin, J., Yen, G., Schwartz, D.C., Welch, R.A., and Blattner, F.R. (2001) Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature* **409**: 529-533.

Petrosino, J.F., Xiang, Q., Karpathy, S.E., Jiang, H., Yerrapragada, S., Liu, Y., Gioia, J., Hemphill, L., Gonzalez, A., Raghavan, T.M., Uzman, A., Fox, G.E., Highlander, S., Reichard, M., Morton, R.J., Clinkenbeard, K.D., and Weinstock, G.M. (2006) Chromosome rearrangement and diversification of Francisella tularensis revealed by the type B (OSU18) genome sequence. *J Bacteriol* **188**: 6977-6985.

Pohlner, J., Halter, R., and Meyer, T.F. (1987) Neisseria gonorrhoeae IgA protease. Secretion and implications for pathogenesis. *Antonie Van Leeuwenhoek* **53**: 479-484.

Pukatzki, S., Ma, A.T., Sturtevant, D., Krastins, B., Sarracino, D., Nelson, W.C., Heidelberg, J.F., and Mekalanos, J.J. (2006) Identification of a conserved bacterial protein secretion system in Vibrio cholerae using the Dictyostelium host model system. *Proc Natl Acad Sci U S A* **103**: 1528-1533.

Pupo, G.M., Lan, R., and Reeves, P.R. (2000) Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A* **97**: 10567-10572.

Rambow-Larsen, A.A., and Weiss, A.A. (2004) Temporal expression of pertussis toxin and Ptl secretion proteins by Bordetella pertussis. *J Bacteriol* **186**: 43-50.

Rao, P.S., Yamada, Y., Tan, Y.P., and Leung, K.Y. (2004) Use of proteomics to identify novel virulence determinants that are required for Edwardsiella tarda pathogenesis. *Mol Microbiol* **53**: 573-586.

Rappuoli, R., and Nabel, G. (2001) Vaccines: ideal drugs for the 21st century? *Curr Opin Investig Drugs* **2**: 45-46.

Raskin, D.M., Seshadri, R., Pukatzki, S.U., and Mekalanos, J.J. (2006) Bacterial genomics and pathogen evolution. *Cell* **124**: 703-714.

Ren, C.P., Chaudhuri, R.R., Fivian, A., Bailey, C.M., Antonio, M., Barnes, W.M., and Pallen, M.J. (2004) The ETT2 gene cluster, encoding a second type III secretion system from Escherichia coli, is present in the majority of strains but has undergone widespread mutational attrition. *J Bacteriol* **186**: 3547-3560.

Ren, C.P., Beatson, S.A., Parkhill, J., and Pallen, M.J. (2005) The Flag-2 locus, an ancestral gene cluster, is potentially associated with a novel flagellar system from Escherichia coli. *J Bacteriol* **187**: 1430-1440.

Riley, M., Abe, T., Arnaud, M.B., Berlyn, M.K., Blattner, F.R., Chaudhuri, R.R., Glasner, J.D., Horiuchi, T., Keseler, I.M., Kosuge, T., Mori, H., Perna, N.T., Plunkett, G., 3rd, Rudd, K.E., Serres, M.H., Thomas, G.H., Thomson, N.R., Wishart, D., and Wanner, B.L. (2006) Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. *Nucleic Acids Res* **34**: 1-9.

Riley, R.L., Mills, C.C., Nyka, W., Weinstock, N., Storey, P.B., Sultan, L.U., Riley, M.C., and Wells, W.F. (1995) Aerial dissemination of pulmonary tuberculosis. A two-year study of contagion in a tuberculosis ward. 1959. *Am J Epidemiol* **142**: 3-14.

Rozen, S., and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**: 365-386.

Sahin, O., Morishita, T.Y., and Zhang, Q. (2002) Campylobacter colonization in poultry: sources of infection and modes of transmission. *Anim Health Res Rev* **3**: 95-105.

Salama, N., Guillemin, K., McDaniel, T.K., Sherlock, G., Tompkins, L., and Falkow, S. (2000) A whole-genome microarray reveals genetic diversity among Helicobacter pylori strains. *Proc Natl Acad Sci U S A* **97**: 14668-14673.

Sandkvist, M. (2001a) Biology of type II secretion. *Mol Microbiol* **40**: 271-283.

Sandkvist, M. (2001b) Type II secretion and pathogenesis. *Infect Immun* **69**: 3523-3535.

Sandstrom, G. (1994) The tularaemia vaccine. *J Chem Technol Biotechnol* **59**: 315-320.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., and Smith, M. (1977a) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**: 687-695.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977b) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**: 5463-5467.

Santic, M., Molmeret, M., Klose, K.E., Jones, S., and Kwaik, Y.A. (2005) The Francisella tularensis pathogenicity island protein IglC and its regulator MglA are essential for modulating phagosome biogenesis and subsequent bacterial escape into the cytoplasm. *Cell Microbiol* **7**: 969-979.

Saunders, N.J., Jeffries, A.C., Peden, J.F., Hood, D.W., Tettelin, H., Rappuoli, R., and Moxon, E.R. (2000) Repeat-associated phase variable genes in the complete genome sequence of Neisseria meningitidis strain MC58. *Mol Microbiol* **37**: 207-215.

Schoolnik, G.K. (2002a) Microarray analysis of bacterial pathogenicity. *Adv Microb Physiol* **46**: 1-45.

Schoolnik, G.K. (2002b) Functional and comparative genomics of pathogenic bacteria. *Curr Opin Microbiol* **5**: 20-26.

Schulein, R., Guye, P., Rhomberg, T.A., Schmid, M.C., Schroder, G., Vergunst, A.C., Carena, I., and Dehio, C. (2005) A bipartite signal mediates the transfer of type IV secretion substrates of Bartonella henselae into human cells. *Proc Natl Acad Sci U S A* **102**: 856-861.

Schuster, S.C. (2008) Next-generation sequencing transforms today's biology. *Nat Methods* **5**: 16-18.

Segal, G., Purcell, M., and Shuman, H.A. (1998) Host cell killing and bacterial conjugation require overlapping sets of genes within a 22-kb region of the Legionella pneumophila genome. *Proc Natl Acad Sci U S A* **95**: 1669-1674.

Segal, G., Feldman, M., and Zusman, T. (2005) The Icm/Dot type-IV secretion systems of Legionella pneumophila and Coxiella burnetii. *FEMS Microbiol Rev* **29**: 65-81.

Serres, M.H., Gopal, S., Nahum, L.A., Liang, P., Gaasterland, T., and Riley, M. (2001) A functional update of the Escherichia coli K-12 genome. *Genome Biol* **2**: RESEARCH0035.

Sexton, J.A., Miller, J.L., Yoneda, A., Kehl-Fie, T.E., and Vogel, J.P. (2004) Legionella pneumophila DotU and IcmF are required for stability of the Dot/Icm complex. *Infect Immun* **72**: 5983-5992.

Sheikh, J., Dudley, E.G., Sui, B., Tamboura, B., Suleman, A., and Nataro, J.P. (2006) EilA, a HilA-like regulator in enteroaggregative Escherichia coli. *Mol Microbiol* **61**: 338-350.

Shendure, J., and Ji, H. (2008) Next-generation DNA sequencing. *Nat Biotechnol* **26**: 1135-1145.

Shrivastava, S., and Mande, S.S. (2008) Identification and functional characterization of gene components of Type VI Secretion system in bacterial genomes. *PLoS ONE* **3**: e2955.

Siguier, P., Filee, J., and Chandler, M. (2006a) Insertion sequences in prokaryotic genomes. *Curr Opin Microbiol* **9**: 526-531.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006b) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* **34**: D32-36.

Snyder, L.A., Loman, N., Pallen, M.J., and Penn, C.W. (2009) Next-Generation Sequencing-the Promise and Perils of Charting the Great Microbial Unknown. *Microb Ecol* **57**: 1-3.

Sobhanifar, S. (2003) Yeast Two Hybrid Assay: A Fishing Tale. In *Special section on techniques, Pathology, University of British Columbia*.

Sreenu, V.B., Alevoor, V., Nagaraju, J., and Nagarajaram, H.A. (2003) MICdb: database of prokaryotic microsatellites. *Nucleic Acids Res* **31**: 106-108.

Stabler, R.A., Hinds, J., Witney, A.A., Isherwood, K., Oyston, P., Titball, R., Wren, B., Hinchliffe, S., Prentice, M., Mangan, J.A., and Butcher, P.D. (2003) Construction of a Yersinia pestis microarray. *Adv Exp Med Biol* **529**: 47-49.

Stavrum, R., Valvatne, H., Bo, T.H., Jonassen, I., Hinds, J., Butcher, P.D., and Grewal, H.M. (2008) Genomic diversity among Beijing and non-Beijing Mycobacterium tuberculosis isolates from Myanmar. *PLoS ONE* **3**: e1973.

Stein, M., Rappuoli, R., and Covacci, A. (2000) Tyrosine phosphorylation of the Helicobacter pylori CagA antigen after cag-driven host cell translocation. *Proc Natl Acad Sci U S A* **97**: 1263-1268.

Stewart, B.J., and McCarter, L.L. (2003) Lateral flagellar gene system of Vibrio parahaemolyticus. *J Bacteriol* **185**: 4508-4518.

Sukhan, A., Kubori, T., and Galan, J.E. (2003) Synthesis and localization of the Salmonella SPI-1 type III secretion needle complex proteins PrgI and PrgJ. *J Bacteriol* **185**: 3480-3483.

Svensson, K., Larsson, P., Johansson, D., Bystrom, M., Forsman, M., and Johansson, A. (2005) Evolution of subspecies of Francisella tularensis. *J Bacteriol* **187**: 3903-3908.

Tarnvik, A. (1989) Nature of protective immunity to Francisella tularensis. *Rev Infect Dis* **11**: 440-451.

Tettelin, H., Saunders, N.J., Heidelberg, J., Jeffries, A.C., Nelson, K.E., Eisen, J.A., Ketchum, K.A., Hood, D.W., Peden, J.F., Dodson, R.J., Nelson, W.C., Gwinn, M.L., DeBoy, R., Peterson, J.D., Hickey, E.K., Haft, D.H., Salzberg, S.L., White, O., Fleischmann, R.D., Dougherty, B.A., Mason, T., Ciecko, A., Parksey, D.S., Blair, E., Cittone, H., Clark, E.B., Cotton, M.D., Utterback, T.R., Khouri, H., Qin, H., Vamathevan, J., Gill, J., Scarlato, V., Masignani, V., Pizza, M., Grandi, G., Sun, L., Smith, H.O., Fraser, C.M., Moxon, E.R., Rappuoli, R., and Venter, J.C. (2000) Complete genome sequence of Neisseria meningitidis serogroup B strain MC58. *Science* **287**: 1809-1815.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N.,

Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R., and Fraser, C.M. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* **102**: 13950-13955.

Thanassi, D.G. (2002) Ushers and secretins: channels for the secretion of folded proteins across the bacterial outer membrane. *J Mol Microbiol Biotechnol* **4**: 11-20.

Thomas, N.A., Deng, W., Puente, J.L., Frey, E.A., Yip, C.K., Strynadka, N.C., and Finlay, B.B. (2005) CesT is a multi-effector chaperone and recruitment factor required for the efficient type III secretion of both LEE- and non-LEE-encoded effectors of enteropathogenic Escherichia coli. *Mol Microbiol* **57**: 1762-1779.

Titball, R.W., Johansson, A., and Forsman, M. (2003) Will the enigma of Francisella tularensis virulence soon be solved? *Trends Microbiol* **11**: 118-123.

Titball, R.W., and Oyston, P.C. (2003) A vaccine for tularaemia. *Expert Opin Biol Ther* **3**: 645-653.

Titball, R.W., and Petrosino, J.F. (2007) Francisella tularensis genomics and proteomics. *Ann N Y Acad Sci* **1105**: 98-121.

Tobe, T., Beatson, S.A., Taniguchi, H., Abe, H., Bailey, C.M., Fivian, A., Younis, R., Matthews, S., Marches, O., Frankel, G., Hayashi, T., and Pallen, M.J. (2006) An extensive repertoire of type III secretion effectors in Escherichia coli O157 and the role of lambdoid phages in their dissemination. *Proc Natl Acad Sci U S A* **103**: 14941-14946.

Torres-Cruz, J., and van der Woude, M.W. (2003) Slipped-strand mispairing can function as a phase variation mechanism in Escherichia coli. *J Bacteriol* **185**: 6990-6994.

Touchman, J.W., Wagner, D.M., Hao, J., Mastrian, S.D., Shah, M.K., Vogler, A.J., Allender, C.J., Clark, E.A., Benitez, D.S., Youngkin, D.J., Girard, J.M., Auerbach, R.K., Beckstrom-Sternberg, S.M., and Keim, P. (2007) A North American Yersinia pestis draft genome sequence: SNPs and phylogenetic analysis. *PLoS ONE* **2**: e220.

Tucker, S.C., and Galan, J.E. (2000) Complex function for SicA, a Salmonella enterica serovar typhimurium type III secretion-associated chaperone. *J Bacteriol* **182**: 2262-2268.

Tukel, C., Akcelik, M., de Jong, M.F., Simsek, O., Tsolis, R.M., and Baumler, A.J. (2007) MarT activates expression of the MisL autotransporter protein of Salmonella enterica serotype Typhimurium. *J Bacteriol* **189**: 3922-3926.

van Belkum, A., Scherer, S., van Alphen, L., and Verbrugh, H. (1998) Short-sequence DNA repeats in prokaryotic genomes. *Microbiol Mol Biol Rev* **62**: 275-293.

van Belkum, A. (1999a) Short sequence repeats in microbial pathogenesis and evolution. *Cell Mol Life Sci* **56**: 729-734.

van Belkum, A. (1999b) The role of short sequence repeats in epidemiologic typing. *Curr Opin Microbiol* **2**: 306-311.

van Belkum, A. (2007) Tracing isolates of bacterial species by multilocus variable number of tandem repeat analysis (MLVA). *FEMS Immunol Med Microbiol* **49**: 22-27.

van Ham, R.C., Hart, H., Mes, T.H., and Sandbrink, J.M. (1994) Molecular evolution of noncoding regions of the chloroplast genome in the Crassulaceae and related species. *Curr Genet* **25**: 558-566.

Vogel, J.P., Andrews, H.L., Wong, S.K., and Isberg, R.R. (1998) Conjugative transfer by the virulence system of Legionella pneumophila. *Science* **279**: 873-876.

Wainwright, L.A., and Kaper, J.B. (1998) EspB and EspD require a specific chaperone for proper secretion from enteropathogenic Escherichia coli. *Mol Microbiol* **27**: 1247-1260.

Wassenaar, T.M., and Blaser, M.J. (1999) Pathophysiology of Campylobacter jejuni infections of humans. *Microbes Infect* **1**: 1023-1033.

Weissman, S.J., Moseley, S.L., Dykhuizen, D.E., and Sokurenko, E.V. (2003) Enterobacterial adhesins and the case for studying SNPs in bacteria. *Trends Microbiol* **11**: 115-117.

Welch, R.A., Burland, V., Plunkett, G., 3rd, Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J., Stroud, D., Mayhew, G.F., Rose, D.J., Zhou, S., Schwartz, D.C., Perna, N.T., Mobley, H.L., Donnenberg, M.S., and Blattner, F.R. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proc Natl Acad Sci U S A* **99**: 17020-17024.

Whipp, M.J., Davis, J.M., Lum, G., de Boer, J., Zhou, Y., Bearden, S.W., Petersen, J.M., Chu, M.C., and Hogg, G. (2003) Characterization of a novicida-like subspecies of Francisella tularensis isolated in Australia. *J Med Microbiol* **52**: 839-842.

Whittam, T.S., Wolfe, M.L., Wachsmuth, I.K., Orskov, F., Orskov, I., and Wilson, R.A. (1993) Clonal relationships among Escherichia coli strains that cause hemorrhagic colitis and infantile diarrhea. *Infect Immun* **61**: 1619-1629.

Williams, M.D., Ouyang, T.X., and Flickinger, M.C. (1994) Starvation-induced expression of SspA and SspB: the effects of a null mutation in sspA on Escherichia coli protein synthesis and survival during growth and prolonged starvation. *Mol Microbiol* **11**: 1029-1043.

Witney, A.A., Marsden, G.L., Holden, M.T., Stabler, R.A., Husain, S.E., Vass, J.K., Butcher, P.D., Hinds, J., and Lindsay, J.A. (2005) Design, validation, and application of a seven-strain Staphylococcus aureus PCR product microarray for comparative genomics. *Appl Environ Microbiol* **71**: 7504-7514.

Wold, B., and Myers, R.M. (2008) Sequence census methods for functional genomics. *Nat Methods* **5**: 19-21.

Wren, B.W. (2000) Microbial genome analysis: insights into virulence, host adaptation and evolution. *Nat Rev Genet.* **1**: 30-39.

Yang, F., Yang, J., Zhang, X., Chen, L., Jiang, Y., Yan, Y., Tang, X., Wang, J., Xiong, Z., Dong, J., Xue, Y., Zhu, Y., Xu, X., Sun, L., Chen, S., Nie, H., Peng, J., Xu, J., Wang, Y., Yuan, Z., Wen, Y., Yao, Z., Shen, Y., Qiang, B., Hou, Y., Yu, J., and Jin, Q. (2005) Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res* **33**: 6445-6458.

Yarwood, J.M., and Schlievert, P.M. (2003) Quorum sensing in Staphylococcus infections. *J Clin Invest* **112**: 1620-1625.

Zhang, L., Chaudhuri, R.R., Constantinidou, C., Hobman, J.L., Patel, M.D., Jones, A.C., Sarti, D., Roe, A.J., Vlisidou, I., Shaw, R.K., Falciani, F., Stevens, M.P., Gally, D.L., Knutton, S., Frankel, G., Penn, C.W., and Pallen, M.J. (2004) Regulators encoded in the Escherichia coli type III secretion system 2 gene cluster influence expression of genes within the locus for enterocyte effacement in enterohemorrhagic E. coli O157:H7. *Infect Immun* **72**: 7282-7293.

Ziebuhr, W., Ohlsen, K., Karch, H., Korhonen, T., and Hacker, J. (1999) Evolution of bacterial pathogenesis. *Cell Mol Life Sci* **56**: 719-728.

PUBLICATIONS