Genomic diversity of ten *Escherichia coli* strains associated with bloodstream infections

Thesis submitted for the degree of Doctor of Philosophy
at the University of Leicester

By

Ali Salman Bin Thani
Department of Infection, Immunity and Inflammation
University of Leicester

December 2007

Dedication

To my little daughter and my wife

Statement of Originality

The accompanying thesis submitted for the degree of PhD entitled "Genomic diversity of ten *Escherichia coli* strains associated with bloodstream infections" is based on work conducted by the author in the Department of Infection, Immunity and Inflammation at the University of Leicester mainly during the period between January 2004 and December 2007.

All the work recorded in this thesis is original unless otherwise acknowledged in the text or by references.

None of the work has been submitted for another degree in this or any other University.

Signed:_____ Date:_____

Acknowledgements

It is time to thank all the people who made this thesis possible by their treasured involvement. I would like to take this opportunity to acknowledge **Dr Kumar Rajakumar** for his supervision, time, support and assistance throughout my PhD.

I gratefully thank **Prof. Mike Barer** for clarifying, discussing with me some of the points in the experiments and for his hints and advices which I have greatly profited from.

I am indebted to **Dr Hong Yu Ou, Dr Rebecca Smith and Dr Natalie Garton** for their continuous help whenever I needed it. My technical discussions with you improved my ability to analyse the results and overcome difficulties. I am also grateful to **Dr Katrina Levi** for her help with the optimization of the PFGE conditions.

Big thanks go to **James Lonnen, Ewan, Barbara, Rob, Farha and Jon** for their advice, help, support, friendship and making my labwork easier by being around. It would not be the same without you all.

I do not forget to thank my **sponsor** (University of Bahrain) for supporting me, sponsoring me throughout my studying and for giving me the great opportunity to take my PhD from the University of Leicester.

Lots of thanks for **Ewan Harrison**, **Hind Abdulmajed** and **Rasheedah Ahmad** for their great assistance by going through some of the chapters and proofread them for me and for their kindness throughout my stay here in Leicester.

Finally, I would like to express my deepest gratitude for the constant support, understanding, encouragement and love I received from my wife **Amal** and my hearted thanks to my smile: my daughter **Reem** who enlightened my life and made a big difference in it in her own way.
Thank you all.

Abstract

Genomic diversity of ten *Escherichia coli* strains associated with bloodstream infections

Ali Salman Bin Thani

Escherichia coli are usually regarded as a harmless human colonic flora. However, pathogenic strains of *E. coli* have been associated with infections that could range from infected mucosal surfaces by intestinal pathogenic *E. coli* to the more severe cases of disseminated infections throughout the body by the extraintestinal pathogroups. The main focus of this project was to investigate the genomic contents of pathogenic bloodstream infection (BSI)-associated *E. coli* strains. This is because the genome contents of the *E. coli* BSI-associated isolates have not been well studied, with only few reports indicating that the pathogenincity of these strains could be attributed to horizontally acquired DNAs known as genomic islands (GEIs).

The genomic contents of 10 clinical BSI-associated *E. coli* strains, isolated at the Leicester Royal Infirmary were investigated in this study. The first approach used to investigate the genomic contents of these strains was by interrogating the downstream ends of tRNA genes for their GEI contents by the sequential PCR strategy tRIP-PCR (tRNA interrogation for pathogenicity islands) followed by the SGSP-PCR (single genome specific primer-PCR). In this approach the flanking regions of the tRNA sites were used to first screen the tRNA genes for their GEIs followed by amplifying the boundaries of the identified GEIs. In the second approach termed Microarray-Assisted mobilome Prospecting (MAmP), the physical genome size of the tested strains obtained by the pulsed-field gel electrophoresis (PFGE) is compared to the sum total of the bits of the genome detected or visualized by the array. The difference between the two measurements is used to estimate the size of the novel, non-microarray-represented mobile genome (mobilome) present in the tested strains.

Remarkably, despite only studying 10 *E. coli* strains, associated with a single disease type the tRIP-PCR method has identified at least 3 GEIs that contain novel sequences, and 46 GEIs, resembling uropathogenic *E. coli* CFT073-like entities. One particular strain E105 had 13 tRNA sites occupied with GEIs. On the other hand, an average novel, non-microarray-borne mobilome of (219 kb /strain) was obtained by the MAmP which, corresponds with previous studies.

The strategies used in this study had proved successful in addressing and identifying mobilome-rich strains. Therefore, using such approaches in combination with whole genome sequencing progects could prioritize the strains and the genomic regions that need to be sequenced. Such prioritization would avoid sequencing of hundreds of isolates to identify their novel gene pool and would reduce the cost of genomic sequencing. Moreover, applying such approaches for the identification of new virulence genes and/or pathogenic mechanisms could lead to significant improvements in the treatment of *E. coli* infections.

TABLE OF CONTENTS

List of Figures

List of Figures

List of Figures

Abbreviations

ABC	ATP-binding cassette
AR-PCR	Arbitrary primed PCR
Bfp	Bundle-forming pili
BHI	Brain heart infusion
bp	Base pairs
BSI	Bloodstream Infection(s)
CDS	Protein coding sequences
CGH	Comparative genomic hybridization
CGI	Comparative genomic indexing
CTAB	Cetyl trimethyl ammonium bromide
D primer (D#)	Downstream primer
DAEC	Diffusely adherent *E. coli*
DNA	Deoxyribonucleic acid
Dnd	DNA degradation
dNTP	Deoxyribonucleotide triphosphate
DR	Directly repeated
eae	*E. coli* attachment-effacement
EAggEC	Enteroaggregative *E. coli*
EAST	Enteroaggregative heat-stable toxin
ECOR	*E. coli* collection of reference strains
EHEC	Enterohemorragic *E. coli*
EIEC	Enteroinvasive *E. coli*
EPEC	Enteropathogenic *E. coli*
ETEC	Enterotoxigenic *E. coli*
ExPEC	Extraintestinal pathogenic *E. coli*
GAP	Gamma amino propylsilane-coated
GEI	Genomic island
HGT	Horizontal gene transfer
HPI	High pathogenicity island
ICs	Inferred contigs
IFR	Institute of food research
III	Infection, Immunity and Inflammation
In	Natural log
IS	Insertion sequence
kb	Kilobase (1000bp)
LB	Luria-Bertani
LEE	Locus of enterocyte effacement
LOWESS	Locally weighted linear regression
LPS	Lipopolysaccharide
LT	Heat-labile toxin
MAmP	Microarray-assisted mobilome prospecting
Mb	Megabase (1000,000bp)
MGE	Mobile genetic element
Mobilome	Mobile genome
MRSA	Methicillin-resistant *S. aureus*

Abbreviations

MVG	Microarray visualized genome
NHE	Normal hydrogen electrode
NPV	Negative predictive value
nt	nucleotide
ORF	Open reading frame
*ori*C	Origin of chromosomal replication
PAIs	Pathogenicity islands
PCR	Polymerase Chain Reaction
PFGE	Pulsed-field gel electrophoresis
pfm	Pulse field mapping
PMT	Photomultiplier tube
PPV	Positive predictive value
rDNA	Ribosomal DNA
RFLP	Restriction fragment length polymorphism
rRNA	Ribosomal RNA
SD	Standard deviation
SDS	Sodium dodecyl sulphate
SGSP-PCR	Single Genome Specific PCR
SNPs	Single nucleotide polymorphisms
ST	Heat-stable toxin
STa	Methanol-soluble heat-stable toxin
STb	Methanol-insoluble heat-stable toxin
STEC	Shiga toxin-producing *E. coli*
Stx	Shiga toxin
T1SS	Type 1 secretion system
tDNA	Transfer DNA
T_m	Melting temperature
tmRNA	Transfer messenger RNA
tRIP	tRNA site Interrogation for PAIs
tRNA	Transfer RNA
U primer (U#)	Upstream primer
UHL	University Hospital of Leicester
UPEC	Uropathogenic *E. coli*
UTI	Urinary tract infection(s)

Chapter 1 Introduction

1.1. The history and classification of *Escherichia coli*

Escherichia coli are facultative anaerobic bacteria usually regarded as a harmless human colonic flora. However, under certain conditions e.g., debilitation, normal "non-pathogenic" strains of *E. coli* can cause infection. These could range from infected mucosal surfaces by intestinal pathogenic *E. coli* to the more severe cases of disseminated infections throughout the body by the extraintestinal pathogroups. Pathogenic *E. coli* strains are associated with generally three clinical syndromes: (i) urinary tract infection, (ii) sepsis/meningitis, and (iii) enteric/diarrhoeal disease (Nataro and Kaper., 1998; Willenbrock *et al.*, 2006).

1.1.1. Isolation and Identification

E. coli belongs to the family *Enterobacteriaceae* and the tribe *Escherichia*, which contains mostly motile gram-negative bacilli (Edwards and Ewing., 1972; Bettelheim., 1994). The organism is recovered from clinical specimens on either general or selective media (MacConkey or eosin methylene-blue agar) at 37°C under aerobic conditions. *Enterobacteriaceae* are identified via biochemical reactions. With these tests about 90% of *E. coli* strains are lactose positive and 99% are indole positive. The indole test is considered as the single best test to differentiate *E. coli* from other members of the *Enterobacteriaceae* (Nataro and Kaper., 1998).

1.1.2. Serotyping

The scheme proposed by Kauffman in 1944 and used to classify *E. coli* according to their surface antigen profiles has been modified and is currently used to serotype *E. coli* strains on the basis of their O (somatic), H (flagellar), and K (capsular) surface antigen profiles (Edwards and Ewing., 1972; Lior., 1996). Agglutination tests are used to determine the presence of these antigens. A combination of O and H antigens is used to define the "serotype" of an isolate. Some serogroups are associated with certain clinical syndromes, and therefore, are used as chromosomal markers that correlate with specific virulent clones (Whittam *et al*., 1993).

1.1.3. Phenotypic assays based on virulence characteristics

With only some exceptions as with serotype O157:H7 (a marker for virulent enterohemorrhagic *E. coli*), identification of pathogenic *E. coli* strains by serotypic markers is rarely sufficient to reliably identify a strain as pathogenic. In addition the test has limited sensitivity and specificity, and is regarded still as tedious and expensive. On the other hand, *in vitro* phenotypic assays are found to be more reliable in the identification of pathogenic strains. These usually correlate with the presence of specific virulence traits. HEp-2 adherence assay is considered a very useful phenotypic assay for the diagnosis of diarrhoeagenic *E. coli*. The test was first described in 1979 by Cravioto *et al* (Cravioto *et al*., 1979; Donnenberg and Nataro., 1995) and still remains the "gold standard" for the diagnosis of enteroaggregative *E. coli* EAggEC and diffusely adherent *E. coli* (DAEC) (Nataro and Kaper., 1998). *E. coli* strains are divided into several disease phenotypes, including nonpathogenic *E. coli*, enterohemorragic *E. coli* (EHEC), enteropathogenic *E. coli* (EPEC), enterotoxigenic *E. coli* (ETEC), enteroinvasive *E. coli* (EIEC), and urinary tract infectious or uropathogenic *E. coli* (UPEC) (Fukiya *et al.*, 2004).

1.1.4. Molecular detection methods

Substantial progress has been made in the development of nucleic acid-based probe technologies and PCR methods. Two methods are used for nucleic acid probe specimen preparation. In the first method purified cultures are inoculated onto agar plates to produce "colony" blots. The bacterial growth is lysed, denatured, and hybridized with the probe *in situ.* An alternative method is the stool blot method. In this technique, stool samples are spotted directly onto nitrocellulose filters overlaid onto an agar plate (Lanata *et al*., 1985) and treated in the same way as with the colony blots. The advantages of this method include (i) no isolation of the microbe is needed from the stool sample, and (ii) a better sensitivity could be achieved if the pathogenic strain represents a minority member of the flora. However, the use of stool blots does not result in a pure bacterial culture which is required for verification of phenotypes (Nataro and Kaper., 1998).

PCR represents a major advance in molecular diagnostics of pathogenic microorganisms. This is because PCR has increased the sensitivity for *in situ* detection of target templates. Problems associated with decreased detection of *E. coli* within stool samples by PCR due to inhibitory substances within stools (Stacy-Phipps *et al*., 1995); had been solved by the use of several methods that remove such inhibitors, including Sepharose spin column chromatography and adsorption of nucleic acids onto glass resin (Stacy-Phipps *et al*., 1995; Lou *et al*., 1997).

1.2. Diseases associated with *E. coli* pathogenic groups

As mentioned above pathogenic *E. coli* strains could range from intestinal pathogens into the more disseminated infections by the extraintestinal pathogroups. At least six pathotypes are identified under the intestinal pathogroup: enterotoxigenic *E. coli* (ETEC), enteroaggregative *E. coli* (EAggEC), diffusely adhering *E. coli* (DAEC), enteropathogenic *E. coli* (EPEC), enterohaemorrhagic *E. coli* (EHEC), and enteroinveasive *E. coli* (EIEC). These six different pathotypes are classified upon certain characteristics such as patterns of bacterial attachment to host cells, effects of attachment on the host cells, production of toxins, and invasiveness (Salyers and Whitt., 2002). A specific type of tissue culture called HEp-2 is used to classify intestinal pathogenic *E. coli* according to their adherence to the host cells. The bacterial strain to be classified is incubated with HEp-2 cells for 3 hours before the tissue culture is washed with buffered saline to remove the nonadherent bacteria. Two types of adherence could be recognized (Figure 1.1). The first is called localized adherence and is characterized by the formation of isolated patches of adherent cells, in which each patch is one or two bacteria thick. This type of adherence is associated with ETEC, EHEC, and EPEC strains. Aggregative adherence, many cells thick, is found in EAggEC. The last class is the diffuse adherence in which bacterial cells are dispersed over the surface of the monolayer and is associated with DAEC. The tissue culture test is also used to determine the invasiveness of a strain (Salyers and Whitt., 2002).

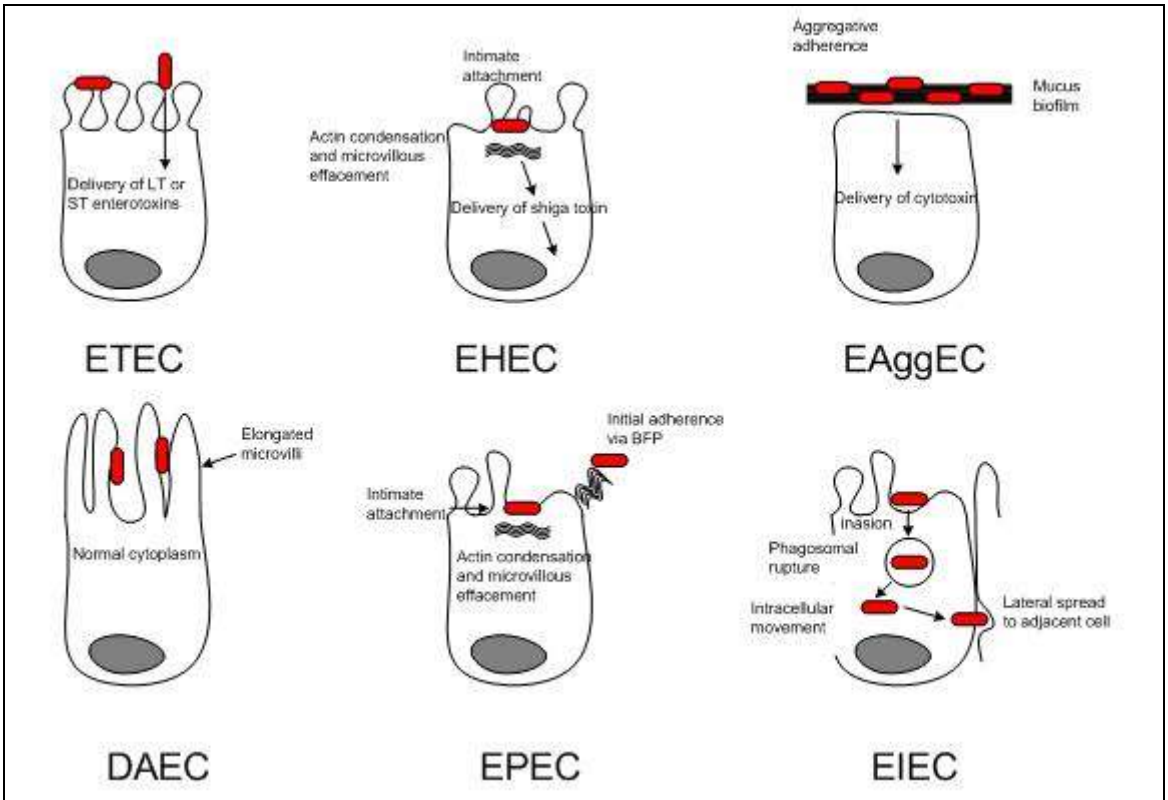


Figure 1.1. Pathogenic schemes of diarrhoeagenic *E. coli*. The interaction of the six pathotypes of diarrhoeagenic *E. coli* strains with eukaryotic cells is schematically represented. Modified from (Nataro and Kaper., 1998).

The pathogenesis of ETEC strain is characterized by the production of toxins that affects the mucosal cells and subsequently causes diarrhoea. ETEC in young children is often considered fatal. However, death among adults is rare, as most adults will develop some immunity to the pathogen especially in epidemic areas. The adult version of the disease is known as the traveller's diarrhoea as most travellers from ETEC free areas are susceptible. The microbe produces two types of enterotoxins. The first is the heat-labile toxin (LT) and the second is heat-stable toxin (ST). The stability of the toxin is defined by the retention of the toxin activity after incubation at 100°C for 30 minutes. Two types of the LT toxins are defined, LT-I and LT-II. Because LT-I shares 75% of its amino acid identity with the cholera toxin it is also known as choleralike toxin. The toxin is involved in the activation of the chloride ion channels of the intestinal cells. The increase in chloride secretion and inhibition of NaCl uptake causes ion imbalance, which consequently leads to loss of water from tissue and diarrhoea (Salyers and Whitt., 2002).

In contrast to LT, the ST represents a family of small peptide toxins that could be classified into two main groups: methanol-soluble ST (STa) and methanol-insoluble ST (STb). The small size of these toxins explains why they are heat-stable. Unlike large proteins small peptides would not unfold at the same extent as large proteins do when inactivated by high temperatures. The mechanism by which the ST toxins could cause diarrhoea is not fully understood, however, some studies suggest that STa works like a hormone analog to guanylin. Guanylin is involved in stimulating the intestinal cells to release water and therefore keeping the mucin layer wet. However, the actions of the STa are more drastic than those caused by guanylin and it is unclear what benefits the bacteria would get of such interaction (Salyers and Whitt., 2002).

Opposite to ETEC, EAggEC strains tend to clump into small aggregates and hence they are called EAggEC, however, both strains are considered non-invasive. Two main features characterize the EAggEC pathogroup: its ability to enhance the mucus secretion from the mucosa and subsequently the formation of bacterium-mucus biofilm and changes to the mucosal surface, which is characterized by the shortening of the villi and the formation of edema by infiltration of the submucosa. Moreover, the EAggEC are involved in the production of ST-like toxin and hemolysin. The ST-like toxin is called enteroaggregative ST (EAST) and its role in the pathogenesis of the EAggEC is to be elucidated. The other toxin associated with EAggEC strains is called exotoxin, which is similar to a hemolysin produced by the urinary tract infections *E. coli*. The exotoxin is involved in the formation of pores on the surface of the HEp-2 cells and increasing the intracellular uptake of Ca^{2+} ions. The Ca^{2+} uptake have different effects on the cellular functions e.g., Ca^{2+} dependent phosphorylation of cell proteins. However, the significant of these effects to the pathogen are still unknown (Salyers and Whitt., 2002).

Like EAggEC, adherence of EPEC strains to mucosal cells results in histological changes causing the effacing of the host cell microvilli. Therefore, this process is called attaching and effacing. The effacing of the host cell microvilli is presumed to be the

results of extensive rearrangement of the host cell actin at the attachment site of the bacteria (Salyers and Whitt., 2002).

The diarrhoeal process is thought to be the result of three stages; the first called non-intimate binding, and involves the association of the bacteria with the host cell by specific pili called bundle-forming pili (Bfp). Next, attachment of the bacteria to the host cell will trigger a signal transduction event, activating the host cell tyrosine kinases which lead to increased Ca^{2+} intracellular levels. Then, the close association of the bacteria with the host cell (intimate binding), would subsequently lead to the actin rearrangement at the vicinity of the bacteria. The genes encoding proteins involved in the attaching-effacing process are called *eae* for *E. coli* attachment-effacement. One of these proteins (intimin) encoded by *eaeA*, is very essential for the actin rearrangements that leads to the effacement of the host microvilli (Salyers and Whitt., 2002).

It is thought that diarrhoea caused by EPEC strains is the result of the mucosal cells incapability to absorb water from the surrounding due to damaged cell surface. A possible reason for such damage could be attributed to the increased Ca^{2+} intracellular levels, which activate the actin-depolymerizing enzymes responsible for cytoskeletal maintenance (Salyers and Whitt., 2002).

Little is known about the interaction of EHEC strains to the host cell; however, the adherence of the EHEC strains involves the same type of actin rearrangements observed with EPEC strains. The EHEC strains produce a toxin that is identical to the Shiga toxin (Stx), responsible for the dysentery (a type of diarrhoea in which stools contain blood and mucus), associated with *Shigella* species. Two types of Stx could be recognized Stx1 and Stx2. The receptors for these Stx are located on the kidney cells as well as on intestinal cells. Colonization of the intestinal mucosa and the dissemination of the bacteria to the kidney could results in acute kidney failure and kidney haemorrhages associated with EHEC strains. An interesting finding about the Stx is that the gene encoding the toxin is present on temperate bacteriophage. The presence of the Stx gene on a bacteriophage could be responsible for the transduction of the gene between diarrhoeal strains of *E. coli* (Salyers and Whitt., 2002).

Pathogenesis of EIEC resembles that observed with *Shigella* spp. as both strains invade the colonic epithelium and produce enterotoxins involved in the diarrhoea. The pathogenesis involve the following steps: penetration of the epithelial cell, followed by lysis of the endocytic vacuole, then the bacteria will start intracellular multiplication, and a directional movement through the cytoplasm before the extension into adjacent epithelial cells (Nataro and Kaper., 1998).

Very little is known about the pathogenesis of the DAEC. A study by Yamamato *et al* (1994) showed that DAEC could induce the HEp-2 cells to produce a finger like projections. These projections are thought to provide a protection for the embedded bacteria from gentamicin.

Infections caused by the extraintestinal pathogenic *E. coli* ExPEC include urinary tract infections, meningitis, and sepsis. In urinary tract infections (UTI) two routes of disseminations could be attributed to the disease. The first is by the community-acquired UTI route, in which colonization of the colon by a UPEC strain is considered as a reservoir for the urinary tract infections. Community-acquired UTIs are also known for being ascending infections. This is because the pathogen starts by infecting the urethra causing urethritis (inflammation of the urethra) and then could ascend to the bladder leading to cystitis (inflammation of the bladder). A more advanced stage of UTI would involve the infection of the kidneys (pyelonephritis). Infection of kidneys is considered to be a more serious disease than urethritis or cystitis as it involves more tissue invasion and due to the vascularized structure of the kidneys it is possible that a potential bacterial leak could lead to a bloodstream infection (Salyers and Whitt., 2002).

The second possible source for the UTI infections is by a UPEC strain acquired in the hospital (nosocomial infections). These infections are usually associated with indwelling urinary catheters. Removal of the catheters is often followed by bacterial clearance from the bladder by normal urine flushes. However, UPEC strains are able to adhere to the bladder cells and therefore colonize the bladder. Patients receiving painkillers during their treatment might develop more complications as the bacteria

might not be detected until late stage of infection e.g., “kidney infection, which is a common starting point for septicaemia in hospitalized patients” (Salyers and Whitt., 2002).

Beside the ascending route of UPEC strains that could lead to the colonization of the kidneys and bloodstream infections (BSI), other possible routes for septicaemia include people with perforated intestinal tracts during an accident or during appendicitis or surgery (Salyers and Whitt., 2002).

1.3. Bacterial genome structure

Many studies were conducted to explore the sizes and organization of bacterial genomes. The smallest of these is for *Mycoplasma genitalium* (580kb) and the largest is for *Myxococcus xanthus* (9.5Mb) (Fonstein *et al*., 1995). Variation in genome sizes between bacterial genera and species included the presence of different numbers and combinations of circular and linear chromosomes and extra-chromosomal entities (Dobrindt and Hacker., 2001). New terminologies were discovered after studying the bacterial genome organization e.g., ‘the minimal gene set’ and ‘the 70% hurdle’ (Mushegian and Koonin., 1996; Koonin., 2000; Bork., 2000). Studies done by Mushegian and Koonin (1996) and Koonin (2000) were aiming to define the minimal gene set that is shared by most of the known bacterial genomes and at the same time were sufficient to support a functional cell. Koonin (2000) concluded that such minimal gene set is dependent on the growth conditions specific to the individual bacterial species. While studies done by Bork (2000) had showed that only the function of less than 70% of the proteins can be predicted with certainty.

In general, the bacterial chromosome consists of a backbone (core) genome (normally 70%−80%) that is required for cellular processes and therefore considered conserved within members of the same family. However, the ability of bacterial strains to colonise a broad spectrum of different environments necessitates the presence of different subsets of an extensive repertoire of metabolic and regulatory genes that allow the bacteria to survive at such niches. The nature of such flexible region was found to be

mobile. Transferring between bacterial strains and accounting for differences between pathotypes. Genes encoded by the flexible genome could range from small accessory genetic elements, such as transposons, insertion sequence elements, prophages and plasmids, up to large genomic islands of 200 kb in size (Hacker and Kaper., 2000; Ochman *et al*., 2000; Dougan *et al.*, 2001; Hacker and Kaper., 2002; Dobrindt *et al*., 2004; Dobrindt., 2005; Medini *et al.*, 2005). Introduction of such elements into the bacterial chromosome is carried via horizontal gene transfer (HGT) and involves vectors such as bacteriophages or plasmids (Acheson *et al*., 1998). Though, three main processes are involved in the transfer of mobile genetic elements, transformation, transduction, and conjugation (Abe *et al*., 1999; Dougan *et al.*, 2001).

Comparative analysis using multigenomic approaches was applied to identify and differentiate these mobile genetic elements associated with virulence determinants from the backbone genome. Such methods included suppressive subtractive hybridization (Bonacorsi *et al*., 2000; Janke *et al*., 2001; Allen *et al*., 2001; Blanc-Potard *et al*., 2002; Miyazaki *et al*., 2002; Stocki *et al*., 2002; Dozois *et al*., 2003; Sorsa *et al*., 2004; Schouler *et al*., 2004; Mokady *et al*., 2005), screening of genomic libraries (Dobrindt *et al*., 2002; Janka *et al*., 2002), array-based DNA–DNA hybridization (Ochman and Jones., 2000; Anjum *et al*., 2003; Dobrindt *et al*., 2003; Fukiya *et al*., 2004), and determination of complete genome sequences (Blattner *et al*., 1997; Hayashi *et al*., 2001; Perna *et al*., 2001; Welch *et al*., 2002; Dobrindt., 2005).

Analysis of these comparative approaches uses features of the genome sequence such as the GC content and base composition to elucidate differences between the core and mobile genomes. This is because the GC content is usually similar between closely related organisms and the base composition is often homogeneous over the entire bacterial chromosome. Therefore, regions with atypical base compositions (dinucleotide bias) or codon usage patterns could indicate evidence for horizontal gene transfer. However such approaches may fail to detect ancient horizontal transfer events as these could adopt more genomic features resembling their host or the acquisition of regions from organisms with similar sequence compositions (Lawrence and Ochman., 1997; Gal-Mor and Finlay., 2006).

Other approaches include the search for genes that are often associated with HGT events such as mobility genes, integrases, transposases, phage genes, or genes with unusual similarity to phylogenetically distant species. Screening of tRNA genes for genomic islands has also proved to be successful, examples include the identification of genomic regions specific to *Salmonella enterica* serovars Typhimurium and Typhi (Hansen-Wester and Hensel., 2002) and the detection of genomic islands in four *E. coli* and *Shigella* strains by *in silico* approach (Ou *et al*., 2006). Other *in silico* approaches would use tools and programs such 'IslandPath' in a direct comparative genomics of evolutionary related species to identify unique regions. These tools incorporates multiple characteristics of PAIs such as atypical sequence composition and HGT associated genes for island detection (Hsiao *et al*., 2003; Gal-Mor and Finlay., 2006).

Variation in genome sizes between different *E. coli* isolates could reach up to 1 Mb. Such heterogeneity in genome size is thought to be due to size variation in the flexible gene pool. For example, genome comparisons between the enterohemorrhagic *E. coli* O157:H7 and K12 strains had revealed that both strains contained a conserved backbone genome of 4.1 megabases (Mb) interrupted by strain-specific islands measured 1.4 Mb for O157:H7 and 0.5 Mb for K12, respectively (Hayashi *et al*., 2001 and Perna *et al*., 2001). Such variation in genome sizes expresses the adaptability of *E. coli* and its success to spread and survive in different environments (Dougan *et al.*, 2001; Fux *et al.*, 2005).

1.4. Genomic islands

Genomic islands (GEIs) are associated with various symbiotic or pathogenic functions of bacteria. Depending on the genes they carry, they may be called symbiosis, metabolic, or resistance islands. For example, the symbiotic lifestyle observed between rhizobia and its plant host is thought to be encoded by GEIs present in two large plasmids pSymA and pSymB harboured by *Sinorhizobium meliloti* strain 1021 (Capela *et al*., 2001; Barnett *et al*., 2001; Galibert *et al*., 2001; and Dobrindt *et al*., 2004). The pSymA encodes for proteins required for nodulation, colonization of a low-oxygen environment (nodules) and metabolism of nitrogen-containing compounds, while, pSymB is involved in the metabolism of organic compounds and production of

polysaccharides required for colonization (Capela *et al*., 2001; Barnett *et al*., 2001; Galibert *et al*., 2001; and Dobrindt *et al*., 2004). Some of the most important resistance islands that have been characterized are the 'SCC*mec*' islands, ranging in size between 20 kb to 60 kb. These were involved in the evolution of the clinically important strains of methicillin-resistant *S. aureus* (MRSA) (Ito *et al*., 1999; Ito *et al*., 2001; Hiramatsu *et al*., 2002).

On the other hand, GEIs that contain large continuous blocks of virulence genes are referred to as 'pathogenicity islands' (PAIs) (Hacker and Kaper., 2000). PAI was first described by Jörg Hacker after discovering two large unstable regions on the chromosome of uropathogenic *Escherichia coli* (UPEC) (Hacker *et al*., 1990; Blum *et al*., 1994). Currently, PAI is used to describe regions that are present in the genomes of pathogens but absent in the non-pathogenic strains of closely related species (Gal-Mor and Finlay., 2006).

1.5. Structure of the PAI

Except for carrying genes encoding virulence factors, PAIs share the following features with other types of genomic islands (Figure 1.2):

PAIs are often integrated downstream of tRNA genes. The 3' end of tRNA genes is identical to the attachment sites of bacteriophages, and therefore these sites are regarded as target sites for integration of certain plasmids and phages in different bacteria (Reiter *et al*., 1989). 75% of the identified GEIs are associated with tRNA genes emphasizing the relationship between the genomic islands and other foreign DNA entities such as bacteriophages and plasmids, for example the high pathogenicity island HPI of *Yersinia* carries a phage P4 integrase homologue (Buchrieser *et al*., 1998; Hou., 1999). During the integration process small directly repeated (DR) sequences (16 and 20 bp) are generated via recombination with the host genome. After the integration process the PAIs are flanked by these (DR) sequences.

The sizes of PAIs are usually between ≥ 10–200 kb, with smaller DNA segments being called pathogenicity islets. The mosaic structure of PAIs consist of DNA regions that differ from the core genome in G+C content and codon usage, reflecting the generation

of PAIs by horizontal gene transfer in a multistep process. This mosaic-like structure carries cryptic or functional genes encoding mobility factors such as integrases, transposases, and insertion sequence (IS). All of which are involved in the instability of the PAIs and their subsequent deletion (Hacker and Kaper., 2000).

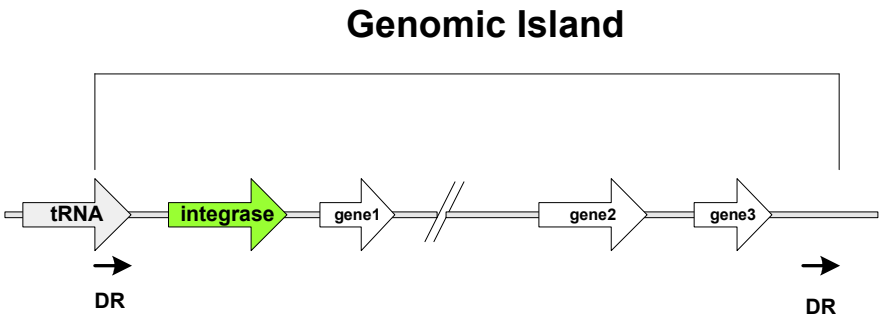


Figure 1.2. Typical structure of genomic island. The black arrows refer to the direct repeats (DR).

Some of the most important virulence factors encoded on PAIs are listed below:

1.5.1. Adhesins

These are attachment factors enabling the microbes to attach to specific eukaryotic receptors. In uropathgenic *E. coli* UPEC, P fimbriae binds to galactose–α1-4–galactose-specific receptor molecules present on uroepithelial cells. The P-fimbrial genes (*pap* or *prs*) are often linked to gene clusters *hly* and *cfn*, which encode the UPEC-specific toxins α-hemolysin and cytotoxic necrotizing factor 1 respectively (Blum *et al*., 1994). Other attachment factors include the S fimbriae; a protein which binds to sialic acid-specific receptors on uroepithelial cells and on brain cells, and therefore it has been associated with pathogenic strains causing UPEC, sepsis and meningitis (Morschhauser *et al*., 1994).

One of the most important adherence factors is intimin. It is encoded by the *eae* gene, which is present in the locus of enterocyte effacement (LEE) PAI of the enteropathogenic (EPEC) and enterohemorrhagic *E. coli* (EHEC) strains. Both strains are involved in a remarkable colonization mechanism, in which the epithelial cells brush border is effaced after bacterial attachment, leading to subsequent intimate adherence to the epithelial cells. This adherence is mediated by intimin, which binds to

a translocated intimin receptor (Tir) protein. The (Tir) receptor is translocated from the bacterium to the host cell by type III secretion system. Both factors (Tir) and the type III secretion system are encoded on the LEE PAI (Kenny *et al*., 1997).

1.5.2. Secretion Systems

Five different secretion systems were discovered in gram negative bacteria (I, II, III. IV, V). These sophisticated mechanisms are involved in the extracellular secretion or the direct translocation of virulence factors to the surface of the host cell. Type III and IV systems are considered more closely associated with PAI extracellular and translocation mechanisms than the other three systems (Schmidt and Hensel., 2004).

Type I secretion system (T1SS) contains three main parts, an ATP-binding cassette (ABC) transporter protein located within the inner membrane, a periplasmic protein, and an outer membrane protein that forms the secretion pore. *hly* operon is a T1SS encoded by a PAI in UPEC, and it is responsible for synthesis, activation, and transport of α-hemolysin virulence factors (Schmidt and Hensel., 2004).

Examples of type II secretion systems are found on both the mobile genomes and on the core backbone DNA of non-pathogenic strains. The type II system encoded by the pO157 virulence plasmid of EHEC is thought to be involved in the virulence of this strain (Burland *et al*., 1998), however the mechanism of such system is still unknown. Another example is the cryptic type II system encoded in *E. coli* K-12 chromosome (Francetic and Pugsley., 1996), but, the functions and mechanisms of this system is also unknown.

Most of type III secretion systems encoded proteins are located in the inner membrane of pathogenic *E. coli* strains such as EPEC and EHEC (Hueck., 1998). Some of these are ATPase which might provide energy for the secretion or assembly process. Others such as EscC of EPEC and EHEC form channels in the inner effector proteins into the host cell by filamentous surface structures that work like a needle and syringe (Wattiau *et al*., 1996).

Type V secretion systems also known as autotransporters are a group of transport proteins, which are synthesized in the form of a single preproprotein (Henderson *et al*., 1998). After being secreted into the periplasm the transporter domains of the proprotein form a β-barrel structure in the outer membrane allowing the passenger domain of the proprotein to pass through the pore. The passenger domains are released into the extracellular region by a proteolytic cleavage e.g., release of immunoglobulin G proteases and the VacA toxin are examples of such pathogenic factors. Examples of the type V secretion systems include the LPA and the EspC PAI of pathogenic *E. coli* (Kaper *et al*., 1999; Schmidt and Hensel., 2004)

1.5.3. Toxins

Different mechanisms are involved in the functions of toxins encoded by PAIs. The pore-forming toxins of UPEC are called α-hemolysins (HlyA). These are associated with lysis of erythrocytes and other eukaryotic cells after insertion into the eukaryotic cell membrane. Other types of toxins work like protease/mucinase, for example, the Pic protein of *S. flexneri* which degrades gelatine and mucin. On the other hand, many toxins work by modifying host cell components e.g., the cytotoxic necrotizing factor 1 modifies the host RhoA protein, a small GTPase by deamidation. The gene encoding the cytotoxic necrotizing factor 1 is part of PAIs of UPEC, where it forms a gene cluster with *hly* and *pap* genes (Hacker and Kaper., 2000).

1.5.4. Iron Uptake Systems

The iron uptake system plays an important role in the fitness and adaptability of non-pathogenic microbes to particular environments. However, the same system is regarded as part of the pathogenic repertoire of pathogenic strains, as the acquisition of iron is a prerequisite for the infection process. Two different strategies are described for the uptake of iron ions. In the first, bacterial cells express receptors for iron uptake (e.g., heme, hemoglobin, lactoferrin, and transferrin), while in the second method low molecular weight compounds with high affinity for iron binding known as siderophores are secreted by these microbes. Two different siderophore systems can be recognized according to their structures, the yersiniabactin (phenolate) and the aerobactin (hydroxamate) (Hacker and Kaper., 2000).

1.6. Evolution of PAIs

1.6.1. Acquisition of PAIs

Exchange of genetic information by horizontal gene transfer (HGT) contributes to the evolution of bacterial cells as evident from the comparison of pathogenic and non-pathogenic strains of *E. coli* that share only 40% of their genome (Welch *et al*., 2002), indicating that 60% of the genome has been acquired by HGT. The HGT is defined as the transfer of genetic material between cells uncoupled with cell division (Lawrence., 2005). Three different mechanisms are involved in the HGT. The first is natural transformation in which bacterial cells take up free DNA from the surrounding environments. The second process is conjugation, whereby a conjugative plasmid is transferred through a tube-like structure (pilus) from the donor bacterium into a recipient cell. The last method; transduction is the transfer of genes between bacterial cells by viruses (Jain *et al*., 2002).

According to the model of Hacker and Carniel (2001), PAI formation can occur by the following five main steps:

i. The acquisition of virulence gene(s) from the environment (flexible gene pool) by HGT using one of the three mechanisms described above.

ii. Integration into the chromosome by means of site-specific recombination mediated by an integrase or recombinase.

iii. A mobile genetic element might then develop into a PAI by acquisition or loss of other genetic elements or by genetic rearrangements. Such changes, might add to the PAI stability e.g., deletion of plasmid origin of replication (*ori*) or genes that are involved in mobilization and/or self-transfer of plasmids (*tra*, *mob*) or bacteriophages (*int*).

iv. Expression of the acquired genes contributing to the fitness of the bacteria and regulation of these virulence genes.

v. Further recombination, insertion, or excision events could result in regain of mobile genetic elements, which could consequently lead to excision of the entire island and its transfer to another recipient.

1.6.2. Origins of PAIs

Specific ecological niches colonized by diverse bacterial species are presumed to be the likely source of such horizontal gene pool with its unique and novel DNA sequences. These include aquatic systems, soil, mixed microbial biofilms, or the rumens and guts of animals (Gal-Mor and Brett Finlay., 2006).

Naked DNA fragments are found in such environments in the form of plasmids, chromosomal DNA fragments, phages, integrating conjugative elements and other mobile genetic elements. All of which are depicted to be transferred by one of the three methods previously described (transformation, conjugation and transduction), with recent analysis indicating that phage transduction is regarded as the predominant mechanism of the three (Canchaya *et al*., 2003). Therefore, phages are proposed to be a versatile carrier of new genetic information into the bacterial genome, or as a way of rearranging existing genetic information in new and unique combinations (Gal-Mor and Brett Finlay., 2006).

1.7. Distribution of PAIs in *E. coli* strains

PAIs are usually associated with certain bacterial pathogens, or to a particular strain or serovar. For example, PAI I_{536} and PAI III_{536} are specific to uropathogenic *E. coli* strain 536 only; whereas PAI_{CFT073} is found exclusively in UPEC strain CFT073 (Guyer *et al*., 1998). PAI I_{536} (77 kb) is located directly downstream of *selC*, and is associated with a phage P4-like integrase gene. While, PAI III_{536} (68 kb) is integrated downstream of *thrW* tRNA gene. This PAI begins with an integrase gene homologous to that of phage Sfx, which recognizes *thrW* as an integration site. PAI I_{CFT073} (58 kb), carries an alpha-hemolysin operon (*hly*), a *pap* operon encoding P-fimbriae, and genes related to iron transport systems and putative carbohydrate transport systems. PAI I_{CFT073} was found in

CFT073 a UPEC strain isolated from the blood and urine of a woman with pyelonephritis (Mobley *et al*., 1990; Welch *et al*., 2002).

On the other hand, some PAIs are rather promiscuous. The high-pathogenicity island (HPI) discovered in pathogenic *Yersinia* spp (Carniel *et al*., 1996) has been also found in many other bacterial strains like enteroaggregative *E. coli* (EAggEC), enteroinvasive (EIEC), enterotoxic (ETEC), enteropathogenic (EPEC) (Schubert *et al*., 1998), and Shiga toxin-producing (STEC) *E. coli*, extraintestinal *E. coli*, and *Salmonella enterica* strains (Karch *et al*., 1999; Oelschlaeger *et al*., 2003; Schubert *et al*., 2002). HPI integrates into one of the three *asnT* tRNA genes, a feature of this GEI that might be caused by an integrase specific for *asnT* (Schubert *et al*., 1999). The HPI contributes to the fitness of *E. coli* by iron acquisition. Interestingly, PAI IV_{536} (30.2 kb) is also associated with the *asnT* tRNA gene and is considered to be the core element of the HPI of pathogenic *Yersinia* spp. (Dobrindt *et al*., 2002). Another example is the locus of enterocyte effacement (LEE) PAI, which is present in different enteropathogenic *E. coli* (EPEC), enterohaemorrhagic *E. coli* (EHEC) strains, as well as in *C. rodentium*. This PAI is responsible for causing infant diarrhoea in developing and industrialized countries. In E2348/69 EPEC strain the LEE is 36 kb in size with 41 ORF. The PAI is usually associated with the following tRNA sites *selC*, *pheU*, or *pheV*. The mechanism by which the LEE is colonizing the host epithelial cells requires the functions of the following proteins (all of which are encoded by the LEE PAI): The central portion of the LEE encodes intimin (Eae), which mediates intimate attachment of the bacterial cell to the host epithelial cell. Intimin attaches to intimin receptor called Tir. This protein is chaperoned and translocated into the host cell by the type III secretion system. The adherence of EPEC to the epithelial cells results in reduced transepithelial electrical resistance, which consequently increases intestinal permeability. Increased permeability as a result of decreased transepithelial electrical resistance has been found to be associated with diarrhoea. It is proposed that the loss of epithelial cell microvilli due to the bacterial effacing action could lead to decreased absorption and account for the persistent diarrhoea (Canil *et al*., 1993; Spitz *et al*., 1995).

Certain pathogenic species possess more than one PAI in their genome. For example, *S. enterica* carries twelve characterized PAIs, while UPEC 536 and *S. flexneri* harbour at least four PAIs. On the other hand, other bacterial pathogens such as *Mycobacterium* spp., *Chlamydia* spp. and the spirochetes possess none or very few PAIs in their genomes. The reasons for such absence of PAIs from the genomes of these microbes could be related to the lifestyle of these pathogens; being restricted and highly specialized to a specific host environment (Moran., 2002). This would lead to reduction in the genome size and consequently they will lose the ability to replicate outside the host. Such reduction could be the reason for the deletion of major portions of horizontally acquired DNA elements. Moreover, such obligate intracellular lifestyle would mean that access to an environmental flexible gene pool would be consequently reduced (Schmidt and Hensel., 2004).

1.8. Why study BSI-associated *E. coli* strains?

Despite their clinical importance as pathogenic strains, the genome contents of *E. coli* BSI isolates have not been well studied, with only few reports indicating that the pathogenicity of these strains could be attributed to PAIs. For example *E. coli* strain AL862, a BSI strain isolated from a cancer patient was found to harbour two genomic islands at the *pheV* and *pheR* tRNA genes. PAI I_{AL862} (61 kb) possesses AfaE-VIII adhesin, encoded by the *afa-8* operon, as the only virulence factor encoded by this island and a P4-like integrase gene. PAI II_{AL862} is also 61 kb long, and it also contains the *afa-8* gene cluster (Lalioui and Le Bouguenec., 2001).

1.8.1. Disease and epidemiology

Because the main focus of this project is to investigate the genomic content of pathogenic bloodstream infection (BSI)-associated *E. coli* strains, this section as well as the next section, will describe the disease and the pathogenesis of *E. coli* strains isolated from patients with sepsis.

Among the most frequent BSI, gram-negative bacteria represent a serious health problem and are usually associated with high morbidity and mortality (Andremont *et*

al., 1996). In the United States, about 250,000 patients will develop a hospital-acquired infection attributed to bloodstream infections when admitted to the hospital, and 35% of them are expected to be within the mortality rate. In the UK, 60% of the *bacteraemia* cases are caused by infections from the *Enterobacteriaceae* family (Reacher *et al.*, 2000; Hilali *et al.*, 2000; Correa and Pittet., 2000; Hautala *et al.*, 2005).

Septicaemia is caused by the presence of bacteria (*Bacteraemia*) or bacterial products (*endotoxaemia*) such as lipopolysaccharide (LPS) in the bloodstream (Greenwood *et al.*, 2002). The disease is usually associated with extreme inflammations that damage the lungs, liver, kidneys, and cardiovascular system, thus leading to multiple organ failure. This damage is primarily due to activated neutrophils and the cytokines, chemokines, and other pro-inflammatory mediators released from endothelial, epithelial, and other cell populations activated by microbial products, such as LPS (Guha and Mackman., 2001; Strassheim *et al.*, 2002). Gram-negative bacteria that cause bloodstream infections are transferred from intestinal flora (Tancrede and Andremont., 1985) to the blood by a mechanism called bacterial translocation (Berg., 1995). This mechanism is associated with a prior intestinal colonization and overgrowth of the translocating bacteria, immunosuppression of the host, and alterations of the intestinal mucosa (Berg., 1995), all of which may occur in cancer patients. *Escherichia coli* is the most common gram-negative species isolated from cancer patients with *bacteraemia* (Tancrede and Andremont., 1985; Andremont *et al*., 1996).

1.8.2. Pathogenesis and virulence traits associated with BSI-associated *E. coli* strains

Phenotypic and genotypic characterization of *E. coli* strains isolated from the blood of patients were found to express virulence traits such as alpha-hemolysin encoded by *hly* operon, which generates cation-selective pores in eukaryotic cell membranes and causes target cell lysis (Aumont *et al*., 1989; Cherifi *et al*., 1990; Opal *et al*., 1990; Blanco *et al*., 1992), aerobactin (iron binding protein) encoded by the *aero* operon (Aumont *et al*., 1989; Cherifi *et al*., 1990; Opal *et al*., 1990), and other virulence traits e.g., *pap*, the operon which encodes pyelonephritis adhesin pili, and *sfa*, the operon which encodes S fimbria adhesins (Opal *et al*., 1990; Maslow *et al*., 1995). According to the *E. coli*

Collection of Reference ECOR strains (Ochman and Selander., 1984) the population structure of *E. coli*, is thought to be clonal. In ECOR, five major clonal groups are observed, A, B1, B2, D, and E (Herzer *et al*., 1990). Most *E. coli* blood isolates belong to the B2 ECOR group (Picard and Goullet., 1988). Maslow *et al* (1995) and Boyd and Hartl (1998) had found that a set of virulence genes comprising *pap*, *hly*, and *sfa* were distributed in strains that belong to the B2 and D ECOR groups and in a cluster of strains causing bloodstream infections. Moreover, *E. coli* isolates from cancer patients with *bacteraemia* were found to carry these virulence genes *papC*, *hlyC*, and *cnf1* indicating that such traits are associated with the pathogenicity of BSI isolates (Hilali *et al*, 2000).

1.9. Aims and Objectives

This project will focus on investigating the genomic contents of pathogenic bloodstream infection (BSI)-associated *E. coli* strains. This is because the genome contents of the *E. coli* BSI-associated strains have not been well studied, with only few reports indicating that the pathogenicity of these strains could be attributed to horizontally acquired DNAs e.g., genomic islands (GEIs).

We aim to investigate the genomic contents of BSI-associated *E. coli* strains by using two approaches. In the first approach we will test the hypothesis that GEIs are usually associated with tRNA genes. For this approach 15 tRNA and one tmRNA[a] genes were identified as integration hot spots and were selected for interrogation by the sequential PCR strategy tRIP-PCR (tRNA interrogation for pathogenicity islands) followed by the SGSP-PCR (single genome specific primer-PCR). In this approach the flanking regions of the tRNA sites were used to, first screen the tRNA genes for their GEIs followed by amplifying the boundaries of the identified GEIs. The PCR amplicons will be analysed by the NCBI GeneBank Blast similarity searches for identified GEI sequences as well as for novel DNA sequences. In the second approach termed Microarray-Assisted mobilome Prospecting (MAmP), we will obtain the physical genome size of the tested

[a] Transfer messenger RNA is involved in resolving stalled ribosomes during translation and adding a peptide tag to the incomplete protein for proteolysis (William., 2002)

strains by the pulsed-field gel electrophoresis (PFGE) and subtract from it the size of the genome obtained by the microarray-visualized genome (MVG) technique to estimate the size of the novel, non-microarray-represented mobile genome (mobilome) in the tested strains. This information will assist in the identification of strains that are rich in novel mobilome and therefore being targeted for further studies.

We predict that the data generated will improve our understanding about the pathogenicity of the *E. coli* BSI-associated strains and the role of the identified GEIs in the infection.

Chapter 2 Materials and Methods

2.1. Bacterial strains, plasmids and growth conditions

Table 2.1 shows the bacterial strains and the plasmids used. All strains used in this study were *Escherichia coli.* Strains were routinely grown at 37ºC in either 2YT or LB medium (Ausubel *et al.*, 1997) with the addition of the required antibiotics when necessary.

Table 2.1. *Escherichia coli* strains

Bacterial strain	Characteristics/genotype[a]	Reference[b]
E102	BSI isolate (bacteremia), urine negative	UHL isolate
E103	BSI isolate (bacteremia), associated with diarrhea	UHL isolate
E104	BSI isolate (bacteremia)	UHL isolate
E105	BSI isolate (bacteremia)	UHL isolate
E106	BSI isolate (bacteremia), urine negative	UHL isolate
E107	BSI isolate (bacteremia)	UHL isolate
E108	BSI isolate (bacteremia), urine negative	UHL isolate
E109	BSI isolate (bacteremia)	UHL isolate
E110	BSI isolate (bacteremia), urine positive[c]	UHL isolate
E111	BSI isolate (bacteremia), urine negative	UHL isolate
E215	DH5α harboring Tn*5*Map	This study
E217	K12 harboring Tn*5*Map	This study
E218	E102 harboring Tn*5*Map	This study
E220	E103 harboring Tn*5*Map	This study
E222	E104 harboring Tn*5*Map	This study
E223	E105 harboring Tn*5*Map	This study
E224	E107 harboring Tn*5*Map	This study
E225	E108 harboring Tn*5*Map	This study
E226	E110 harboring Tn*5*Map	This study
E227	E111 harboring Tn*5*Map	This study
K12 (MG1655)	F^- λ^- *ilvG- rfb*-50 *rph*-1	Blattner *et al*, 1997
CFT073	Pyelonephritis isolate, P1 *pap*, P2 *pap*	Welch *et al*, 2002
EDL933	Stx_1, stx_2, wild-type enterohemorrhagic O157:H7	Perna *et al*, 2001
DH5α	F^- *endA1 hsdR17* (r_K^- m_K^+) *supE44 thi-1* λ– *recA1 gyrA96 relA1 deoR* Δ(*lacZYA-argF*)-U169 *ϕ*80d*lacZ*ΔM15	Grant *et al*, 1990
SM10*λpir*	*thi thr leu tonA lacY supE recA*::RP4-2Tc::Mu Km *λpir*	Prof Estelle Jumas-Bilak, Universite Montrpellier
SY327*λpir*	Δ(*lac-pro*) *argE* (Am) *rif nalA recAλpir*	Prof Estelle Jumas-Bilak
Plasmids		
pBluescript II KS ($^+$) vector	*lacZ* (Am), cloning vector	Stratagene
pGF2	π-dependent suicide vector containing Tn5Map	Prof Estelle Jumas-Bilak, Universite Montrpellier
pGEM®-T Easy	*lacZ* (Am), cloning vector	Promega

[a]BSI, bloodstream infection.
[b]UHL, University Hospital of Leicester.
[c]The *E. coli* strain was isolated from two specimens' blood and urine.

2.2. Storage of bacterial strains

Bacterial strains were stored at –20°C and -80°C in 30% glycerol medium (100 ml) supplemented with 3.7g of brain heart infusion BHI (OXOID).

2.3. API 20E test for the identification of *Escherichia coli*

The API system (bioMerieux, Inc., Hazelwood, MO) for *Enterobacteriaceae* was used for the identification of *Escherichia coli*. Following the manufacturer's instructions the results were read visually after 18-24 hours incubation at 37°C and were compared to the identification manufacturer database provided in the Leicester Royal Infirmary system.

2.4. Antibiotics

After being dissolved in the appropriate solvent the antibiotics were filter sterilized with 0.4μm acrodisc filter from Pall Corporation, USA. The stock was aliquoted and stored in the appropriate temperature. Agar media were cooled to 50°C before the addition of the antibiotic.

Table 2.2. Antibiotics used in this study

Antibiotic	Stock concentration	Working concentration
Ampicillin (Sigma-Aldrich)	100 mg ml^{-1}	100 μg ml^{-1}
Kanamycin (Sigma-Aldrich)	50 mg ml^{-1}	25 μg ml^{-1} or 50 μg ml^{-1}

2.5. Routine Techniques for DNA manipulation

2.5.1. Genomic DNA extraction from *E. coli* strains

The genomic DNA was extracted using a modification of the method described by Ausubel *et al* (1997). 2.5 ml of overnight bacterial culture grown in BHI broth were harvested and the pellet was resuspended in: 567 μl TE [10 mM Tris.Cl pH 7.4, 1 mM EDTA pH 8.0], 30 μl 10% (w/v) SDS, 3 μl 20 mg ml^{-1} proteinase K (Sigma-Aldrich). The mixture was incubated for 45 minutes at 37°C then 2 μl of 10 mg ml^{-1} RNase were added and the mixture was incubated at 65°C for 45 minutes. Another 2 μl of 20mg ml^{-1} proteinase K were added and incubated for 30 minutes at 37°C. The mixture was

incubated at -20°C for 45 minutes followed by 45 minutes at 37°C. 100 µl of 5 M NaCl was added and mixed thoroughly. Then 80 µl of preheated (65°C) CTAB/NaCl solution were added and mixed thoroughly followed by incubation at 65°C for 10 minutes. The DNA was extracted by adding equal volume (~700µl) of chloroform: isoamyl alcohol (24:1) (Sigma), followed by phenol: chloroform: isoamyl alcohol (25:24:1) (Fisher Scientific) extraction (this step was repeated when necessary), and a final chloroform: isoamyl alcohol (24:1) extraction. All of these steps required thorough mixing of the two phases, then separation by centrifugation at 16046 ×g for 10 minutes. After the final extraction, the aqueous phase was removed to a clean tube and the DNA was precipitated by adding 0.7 volume of isopropanol, (this can be left in the -20°C for 2 hours to aid precipitation) then centrifuged at 16046 ×g for 30 minutes. The pellet was washed twice with 70% ethanol, air dried and resuspended in 100 µl sterile distilled water.

2.5.2. Plasmid DNA Extraction from *E. coli* strains

The plasmid DNA extraction protocol was modified from the alkaline lysis method described by Morelle (1989). 2 ml LB broth of bacterial cultures harbouring plasmids were grown overnight at 37^{o}C with aeration and agitation (200 rpm) and supplemented with corresponding antibiotic(s). Then, Cells were harvested by low-speed centrifugation (6076 ×g for 2 minutes) and resuspended in 200 µl lysis buffer [4 mg ml^{-1} lysozyme (Sigma-Aldrich), 50 mM glucose, 25 mM Tris-HCl pH 8.0, 10 mM EDTA]. 1 µl 10 mg ml^{-1} RNase was added and left for 6-10 minutes at room temperature. 400 µl of freshly prepared alkaline solution [0.2 N NaOH, 1% SDS] were added and mixed by inversion, followed by 5 minutes incubation on ice. 300 µl 7.5 M ammonium acetate solution were added and mixed gently for a few seconds. The mixture was incubated at 0^{o}C for 10mins to allow most of the protein, high molecular weight RNA and chromosomal DNA to precipitate and then centrifuged at 16046 ×g for 10mins. The clear supernatant was removed and transferred into clean eppendorf. 0.6 volumes of isopropanol (Sigma-Aldrich) were added and incubated at room temperature for ten minutes. The mixture was centrifuged at 16046 ×g for 10mins. Again the supernatant

was removed and the pellet was washed twice in 70% ethanol. The tubes were left inverted on tissue paper for ~15 minutes at room temperature to allow pellets to dry. The pellets were resuspended in 100µl sterile distilled H_2O.

2.5.3. DNA concentration quantification

DNA concentration was measured by ultraviolet absorbance spectrophotometry at OD_{260nm} at which wavelength absorbance of 1.0 corresponds to 50µg of double-stranded DNA per ml (Stephenson., 2003).

2.5.4. Agarose gel electrophoresis

0.8% (w/v) agarose (BIOLINE) gel was used to separate DNA fragments (>1kb) and 1-1.2% (w/v) was used to separate smaller DNA fragments. The agarose (multipurpose agarose from BIOLINE) was dissolved in the appropriate volume of 1x TAE buffer (50x TAE: 242g Tris-base 57.1 ml glacial acetic acid, 200 ml 0.5 M EDTA pH 8) and ethidium bromide was added to the agarose gel at a final concentration of 0.5 µg ml^{-1}. DNA was mixed with $1/5^{th}$ (v/v) of 6x gel-loading dye buffer (0.09% bromophenol blue, 0.09% xylene cyanol FF, 60% glycerol and 60 mM EDTA- from MBI Fermentas) before loading the samples into the wells of an agarose gel. The electrophoresis conditions were as following: 80-90 volts for 70 minutes or 10-20 volts for overnight runs. 0.5µg of 10kb GeneRuler DNA ladder mix or Lambda DNA/*Hin*dIII marker from MBI Fermentas was used to size the DNA fragments.

2.5.5. DNA restriction digests

Restriction endonucleases used for DNA manipulation were purchased from Roche Products and Promega UK. The restriction digests were carried in a 40 µl total volume reaction with 3-5 units of enzyme and 1µg genomic DNA or 225ng plasmid DNA for 3 hours at 37°C. Deactivation of enzyme was carried as specified by the manufacturer (usually at 65°C for 15 minutes)

2.5.6. Ligation of DNA fragments

The T4 ligase was purchased from Roche Products or Promega UK. The total volume of the reaction was 30 µl with one unit of T4 ligase and a ratio of 1:20 of vector to Genomic DNA. Ligation was carried out at 4^{o}C overnight for sticky end reactions and 20^{o}C for blunt ends. Deactivation of enzyme was carried at 65^{o}C for ten minutes

2.5.7. Construction of dTTP-tailed vectors (pBluescript II KS^{+})

100ng of pBluescript II KS^{+} were digested with *Eco*RV to linearize the plasmid and produce blunt ends. The plasmid was incubated at 70^{o}C for 2 hours with 2 mM of dTTPs plus 1.5 mM of $MgCl_2$ and one unit of Taq DNA polymerase (Marchuk., *et al* 1990).

2.5.8. Recovering PCR product from gel using QIAquick Gel extraction kit protocol from QIAGEN

DNA fragment of interest was excised from agarose gel with a clean, sharp scalpel, weighed in eppendorf tube and 3 volumes QG (guanidine thiocyanate) buffer were added. Incubated at 50^{o}C for 10 minutes or until gel slice has completely dissolved. To help dissolve the gel slice the mixture was vortexed every 2-3 minutes during incubation. After the gel slice has dissolved completely, the colour of the mixture should change to yellow, if it is orange or violet then 10 µl 3 M sodium acetate were added. One gel volume isopropanol was added to sample and mixed. To bind DNA, sample was centrifuged in QIAquick column at 16046 ×g for one minute. 0.5 ml QG buffer added to column and recentrifuged for another one minute to remove any residual traces of agarose. To wash, 0.75 ml PE (80% ethanol) buffer was added to column and centrifuged for one minute. Again the tube was recentrifuged for one min at 16046 ×g to remove residual ethanol. DNA elution was carried in a clean 1.5 ml microcentrifuge tube by adding 50 µl EB (10 mM Tris.Cl pH 8.5) buffer to the centre of the QIAquick membrane. Left to stand for one minute followed by centrifugation for one minute.

2.5.9. Preparing PCR products for sequencing as specified by the sequencing company

0.1 volume of 3 M sodium acetate pH 5.2 were added to the recovered PCR product using the QIAquick Gel extraction kit followed by adding 2 volumes of ethanol. The reaction was mixed and left for 1-2 hours at room temperature. The mixture was centrifuged at 16046 ×g for 15 minutes and the pellet was washed with 70% ethanol and recentrifuged for another five minutes at 16046 ×g and the pellet was dried under vacuum for 10-20 minutes.

2.6. Preparation of electro-competent bacteria

The method is modified from the method described by Smith *et al*., (1990). 5 ml LB broth was inculcated with *E. coli* DH5α cells and grown overnight at 37^{o}C on shaker table. The overnight culture was diluted in 100 ml LB broth and incubated at 37^{o}C with agitation (200 rpm) to an absorbance of 0.5 at 600nm. All subsequent steps were performed on ice. Cells were harvested and washed twice by centrifugation at 1537 ×g for 10 minutes and resuspended in ~250 ml ice cold 10% v/v glycerol. Again the cells were centrifuged at 1537 ×g for ten minutes and resuspended in 2 ml 10% v/v glycerol. Cells were aliquoted (100-200 µl) and snap-freezed with ethanol and dry ice and stored at –80^{o}C for up to 6 months.

2.7. Electroporation of bacteria

An aliquot of the electro-competent bacterial cells (DH5α) was thawed on ice. 20 µl of the cells were dispensed into a pre-chilled microcentrifuge tube and 2 µl salt-free DNA was added to it. The components were gently mixed by pipetting up and down and then kept on ice for 5mins. The bacteria/DNA mixture was transferred into a pre-chilled 1mm electroporation cuvette. Electroporation parameters were: 2.5Kv, 200Ω and 25µF, yielding a time constant of 4.5 – 4.9 msec. Bacteria were resuspended immediately in one ml SOC medium (Appendix) and left to recover for 1hour at 37^{o}C/200 rpm before being plated out on selective media (LB plates containing 40 µg ml^{-1} final concentration of X-Gal and the required antibiotics).

2.8. Transformation efficiency

After the transformation of the 2 μl ligation reaction into the competent cells and adding the mixture to 1 ml SOC medium, 100 μl of this solution were diluted in 900 μl of SOC medium. Then, 100 μl were spread onto a plate containing 100 μg ml^{-1} ampicillin, and the plate was incubated overnight at 37°C. The first step in calculating the transformation efficiency was to determine how many micrograms of plasmid DNA were in the 100 μl sample spread on the ampicillin plate. The original 2 μl ligation reaction contained 0.2 μg of plasmid DNA. To arrive at the amount of DNA in the 100 μl used for spreading, the original concentration (0.2 μg / 2 μl or 0.15 μg μl^{-1}) was multiplied by the dilution and amount plated.

$$x \text{ μg plasmid DNA} = \frac{0.2\mu g \text{ plasmid DNA}}{2\mu l} \times \frac{2\mu l}{1000\mu l} \times \frac{100\mu l \text{ dilution}}{1000\mu l} \times 100\mu l \text{ plated}$$

$$x \text{ μg plasmid DNA} = \frac{4000 \text{ μg plasmid DNA}}{2{,}000{,}000} = 2.0 \times 10^{-3} \text{ μg plasmid DNA}$$

Therefore, the 100 μl volume spread on the ampicillin plate contained 2.0×10^{-3} μg plasmid DNA. Transformation efficiency is then calculated as the number of colonies counted divided by the amount of plasmid DNA contained within the spreading volume.

$$\text{Transformation efficiency} = \frac{222 \text{ transformants}}{2.0 \times 10^{-3} \text{ μg DNA}} = 1.1 \times 10^{5} \text{ transformats / μg DNA}$$

2.9. Conjugation experiment (plate mating)

2ml 2xYT medium (Appendix) of donor & recipient cells were grown overnight. 0.7 ml of both cultures was mixed. The mixture broth was centrifuged and the pellet was washed with 100µl antibiotic- free 2xYT medium and plated onto tryptone soya agar for overnight plate mating at 37°C. The bacteria was harvested by adding 3 ml of 1x M9 minimal medium (Appendix) to the plate and the growth collected in 1.5 ml eppendorf tube by centrifugation and resuspension of the cells using 1x M9 minimal medium free of antibiotics (Rajakumar *et al*., 1997). A serial dilution of the transconjugants 10^{0}-10^{-05} was plated on M9 supplemented with the following (0.5ml 1M $MgSO_4.7H_2O$, 5ml 20% glucose, 50µl 0.5% thiamine, 2.5ml kanamycin final con. 25 µg ml^{-1}).

2.10. The polymerase chain reaction (PCR)

The total volume of the PCR reactions was 20 µl, ~100ng of template DNA was used in the reaction, and primers at concentration of 1 pmoles/µl. The PCR components were as following: 1X PCR reaction buffer (ABgene), 1.5 mM $MgCl_2$, 200 µM dNTPs, and 2 units of Thermoprime Plus DNA Polymerase (Abgene) for normal PCR or Thermo-start plus DNA Polymerase (Abgene) for hotstart PCR.

2.10.1. tRNA site Interrogation for PAIs (tRIP) PCR

The tRIP-PCR is a strategy to screen large number of strains for their genomic islands (GEI) content (Figure 2.1). In tRIP two primers (upstream and downstream) flanking a tRNA gene are used to indicate the presence or absence of an island. A positive tRIP-PCR indicates the absence of an island as the relatively short DNA sequence between the two primers was amplified by the Taq polymerase. On the other hand, a negative tRIP-PCR could indicates the presence of a GEI. The presence of such GEI is further investigated by another PCR strategy called Single Genome Specific Primer-PCR SGSP-PCR. The SGSP-PCR is used to prove the presence of GEI and exclude other possible reasons for the negative tRIP-PCR such as primers miss-pair or the lack/alteration of the conserved flanking regions for that particular tRNA site. The PCR

primers used for the 16-tRNA gene tRIP-PCRs are listed in Table 2.3. Details about the tRIP-PCR primer design and specifity are presented and published by our group Ou *et al*., (2006). The cycle parameters for the tRIP-PCR were as following (manufacturer-ABgene- instructions with modification): After 5 min incubation at 95°C, the amplification was performed for 30 cycles of 94°C for 30 sec, annealing for 30 sec (annealing temperature was 1°C below the lowest T_m for the primers used), and extension at 72°C (extension time was chosen according to the size of the positive control product e.g., for 1kb PCR product 1 minute was set for extension). After the last cycle, samples were incubated for 10 min at 72°C.

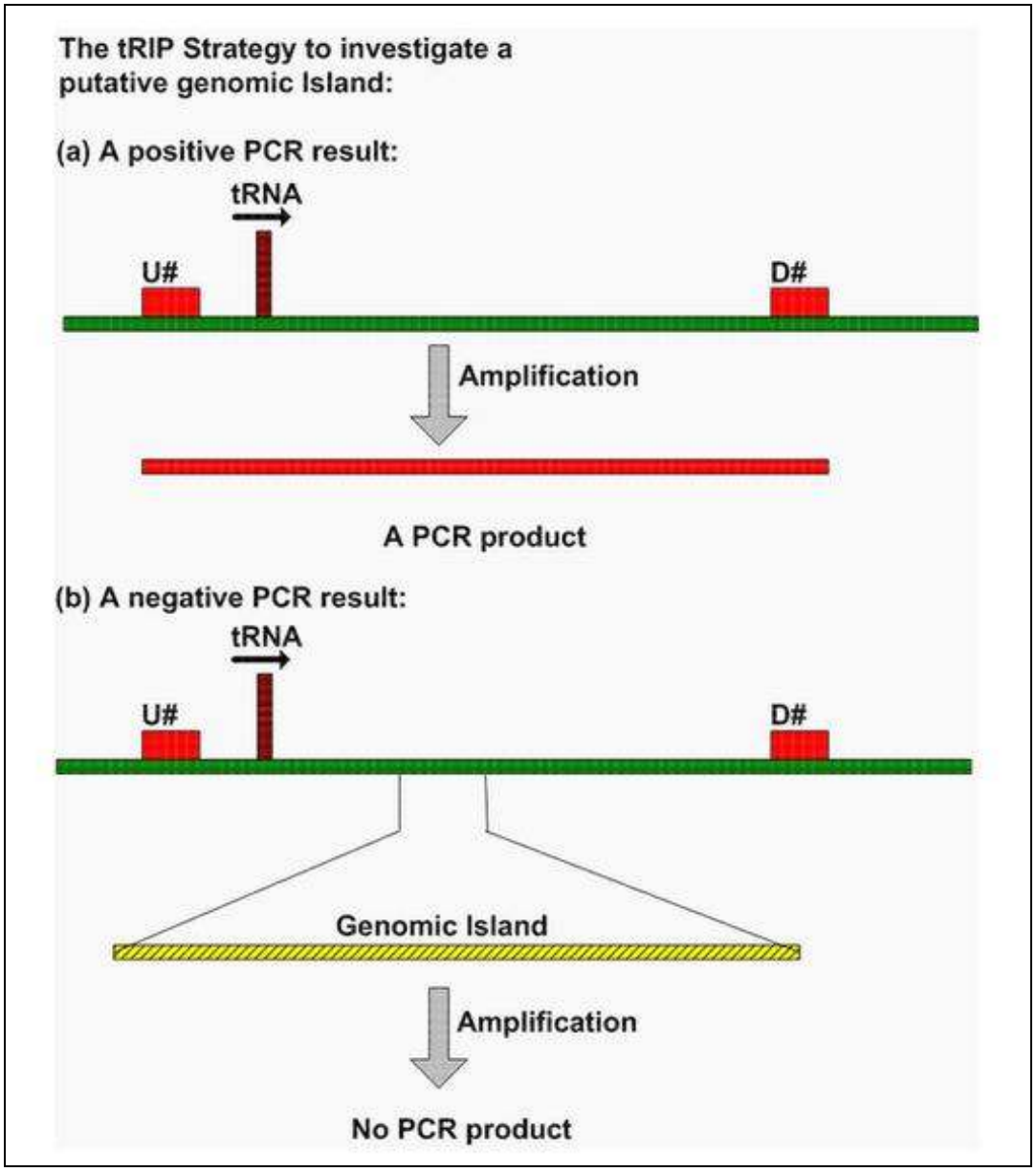


Figure 2.1. The tRIP-PCR strategy used to investigate the insertion of putative genomic islands at tRNA sites. (a) A positive tRIP PCR product indicates the absence of a genomic island in the investigated tRNA site. (b) A negative tRIP PCR product indicates the possibility of a genomic island insertion at this site. U# and D# refer to the upstream and downstream primers used in the tRIP-PCR, respectively.

Table 2.3. Primers used in the tRIP-PCR

Primer name	Primer sequence	Size (bp)
ArgW U	5`TCTGGCCCTTCGCACTACCTACTT 3`	24
ArgW D	5` GCCCGGGCATCAGCAGACATA 3`	21
AsnT U3	5`AGGTTGCTGGCTGGGAACACGAT 3`	23
AsnT U4	5`TGCGTGAGGTTCTGGCTGG 3`	20
AsnT D	5`ACTGGCAACCTGATAACCGACTCCA 3`	25
AsnV U	5` GCCCGGCATAACAAATAATAAAAA 3`	24
AsnV D	5` CGAGAAACCCCGCGTAACTGG 3`	21
AspV U1	5` TTGCGGTGGCGAGGAAAATGTT 3`	22
AspV U2	5`GCTTAAGCGCGATATTCCGAAGAC 3`	24
AspV D1	5` GGTGACAGCCGGGTGATTA 3`	19
AspV D2	5` GCCGCTGGTGTGCTACGACTTAC 3`	23
GlyU U	5`ATGGCGAATTAATCAGCAGTCAGC 3`	24
GlyU D	5`TCCGGGATTATTGTCGCAGTAGTT 3`	24
LeuX U	5`CACCACTTTATCGGCACCCATCG 3`	23
LeuX D	5`GGAGGCCCGCCATGTCACTTT 3`	21
MetV U	5`TAAGGCGCAACGAAGATAACAAAC 3`	24
MetV D	5` CCGGCCAATGCACAGGATA 3`	19
PheU U1	5`GAAACGCAAACCGCCGAACAAAA 3`	23
PheU U2	5`CCCGATCCTGGCCACCCTATTC 3`	22
PheU D1	5` CACGGGGCCGCACGACATT 3`	19
PheU D2	5`GGGCCGCACGACATTTCACG 3`	20
PheV U	5`CCGGATTACGCATCTGTGGCATTA 3`	24
PheV D	5`GCGGCGCGTTTTATTCACTGGT 3`	22
SelC U2	5` CCTTGATGCTATAGGGGTGCTGAGA 3`	25
SelC D5	5`CAATTAGCGTTGAGGGATAGGTGGT 3`	25
SerT U	5`GCACTTTTGGCTGTTTTTCA 3`	20
SerT D	5` TTTACCCATCTTTACGCATTTG 3`	22
SerU U	5` TCCAGGGCCACTTAATCATCGTT 3`	23
SerU D	5` TTGCACCACGAAAATCATCTCAT 3`	23
SerW U	5` GGAGTAATGTGCCGAACCTGT 3`	21
SerW D	5` CACCGATGCGATGGAAGAGAT3`	21
SerX U	5` CAAAGGCCACCAGCATAACAAATC 3`	24
SerX D	5`TTCCCCTCGCCCTAACAGACG 3`	21
SsrA U	5` CCGTACCCGCAAGTTACTTCTCAA 3`	24
SsrA D	5`AGGGGTACTCGATGGCGGTCTATA 3`	24
ThrW U	5` TGACGCATCGCCCGGTTAGTTT 3`	22
ThrW D	5` ACGTCTGCGGTTCGGTGGAGTTT 3`	23

2.10.2. Colony PCR

The same PCR parameters were used as indicated before in tRIP-PCR except for the template DNA that is introduced to the PCR reaction by first making a suspension of a single colony in 50 µl of sterile distilled water. Boiled for 5 minutes and then vortexed and centrifuged before 1 µl of the lysate being introduced into the PCR reaction tube.

2.10.3. Single Genome Specific Primer-PCR (SGSP-PCR) and hot start/ touchdown protocol

Figures 2.2 and 2.3 illustrate in detail the SGSP-PCR strategy. In the SGSP-PCR a panel of eight different restriction enzymes were used to construct the genomic libraries (*Bam*HI, *Eco*RI, *Eco*RV, *Hin*cII, *Hin*dIII, *Kpn*I, *Sal*I, *Pst*I). These genomic libraries consisted of genomic DNA and the vector pBluescript II KS^+ both digested with the same restriction enzyme and ligated with T4 ligase. Amplifying short fragments of the GEI boundaries flanked by the conserved core genome using primers U or D and the vector primers should prove that a negative tRIP-PCR was due to insertion by a GEI or not after sequencing and analysing the SGSP-PCR product.

The cycle parameters for the Hotstart/Touchdown-SGSP PCR were as following (Don *et al*., 1991): After 15 min incubation at 95°C, the amplification was performed for 10 cycles of 94°C for 30 sec, annealing for 30 sec (annealing temperature was set 10°C above the lowest T_m for the primers used and was decreased 1°C every cycle to the touchdown at the lowest T_m), and extension at 72°C for 2.0 min. This was followed by 20 cycles of 94°C for 30 sec, annealing for 30 sec (annealing temperature was set at the lowest T_m), and extension at 72°C for 2.0 min. After the last cycle, samples were incubated for 10 min at 72°C.

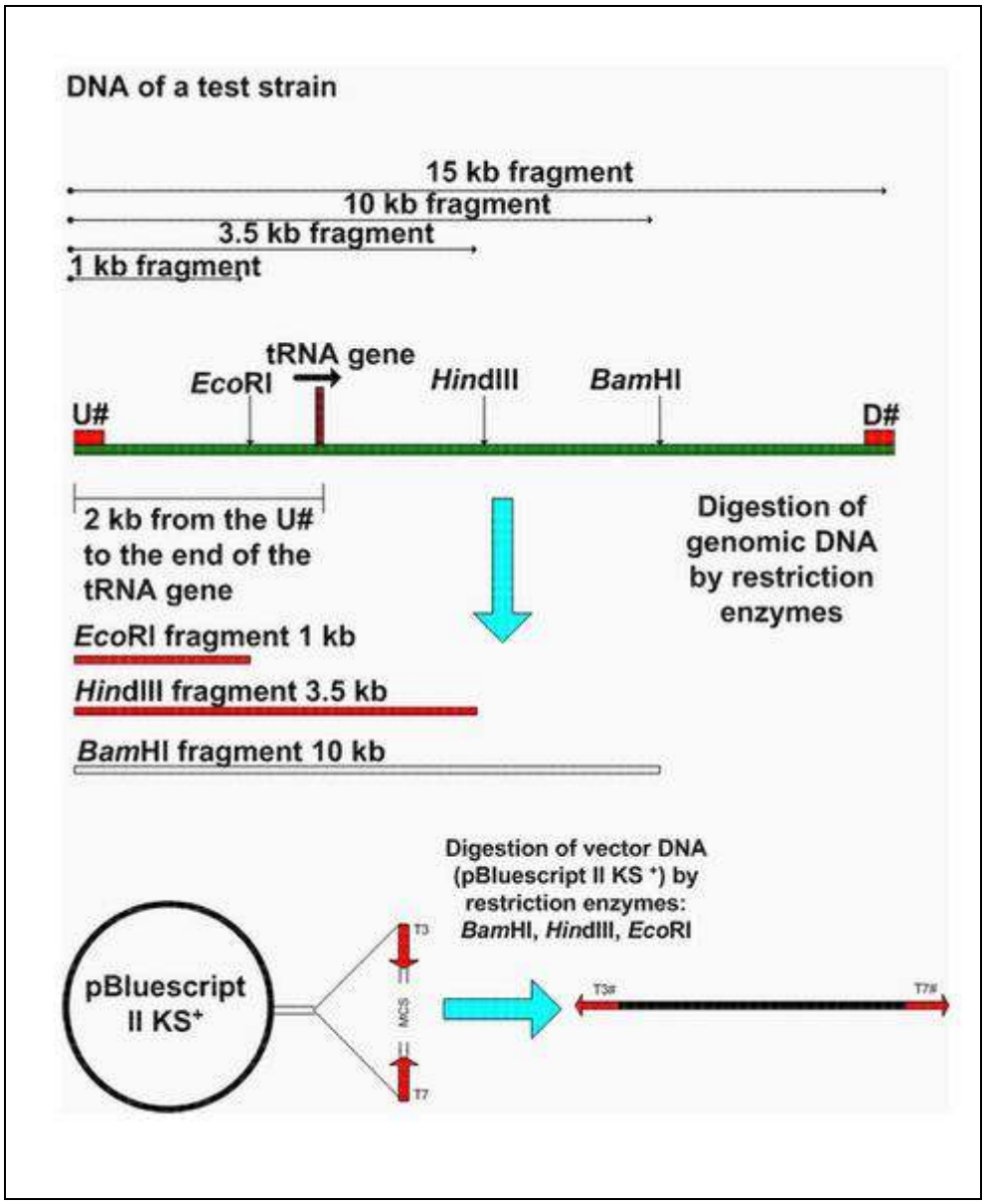


Figure 2.2. The SGSP-PCR strategy. In the SGSP-PCR strategy, different genomic libraries were constructed by the digestion of the genomic DNA and the cloning vector (pBluescript II KS^{+}) with the same restriction enzymes.

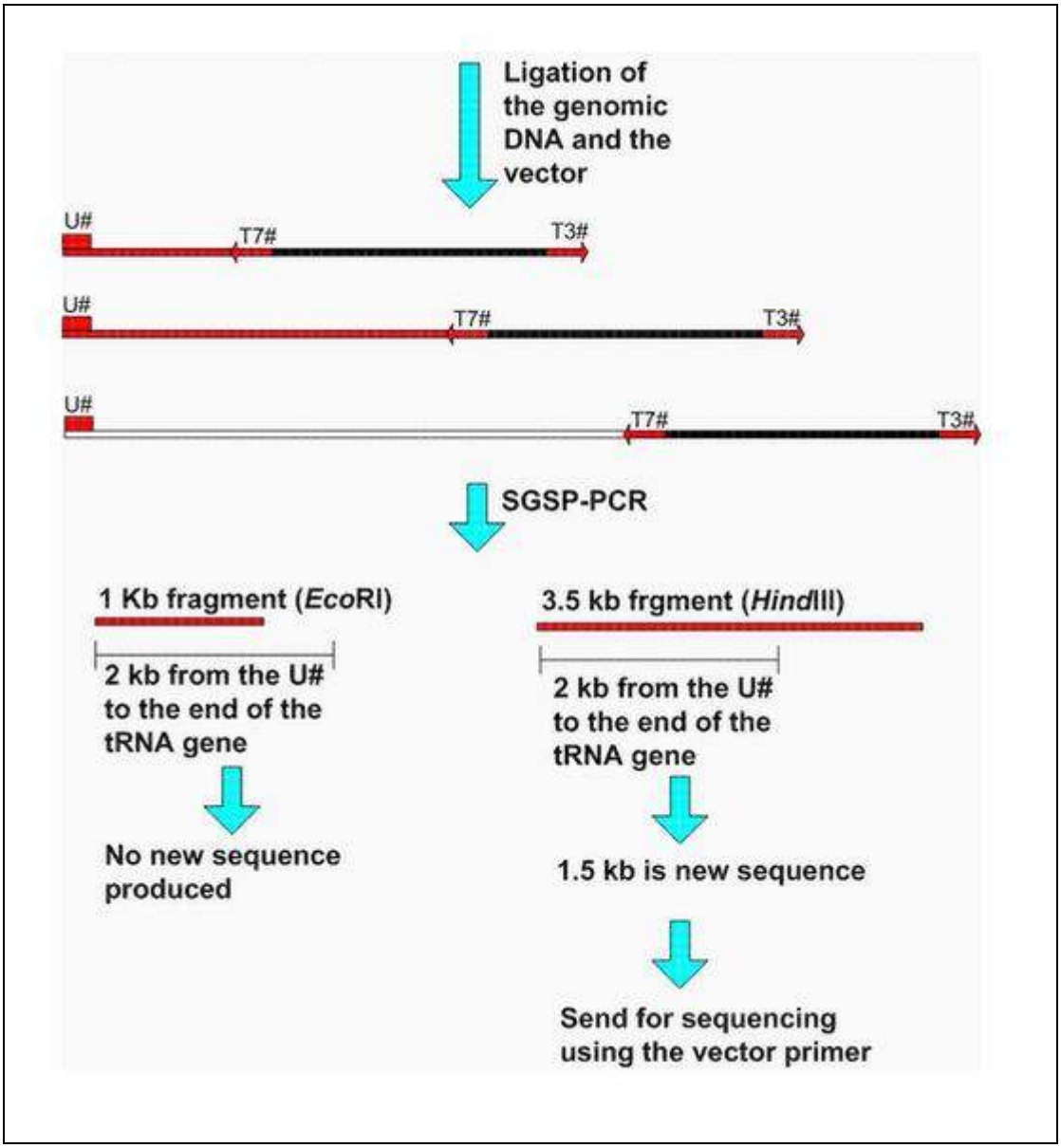


Figure 2.3. Continue of the SGSP-PCR strategy. The restricted fragments from the digested genomic DNA were then ligated to the corresponding vector, which was digested with the same restriction enzyme. GEI DNA sequence can be revealed by PCR amplification using either the U or D primer of the tRNA site and one of the vector primers (e.g., T7). Appropriate SGSP-PCR amplicons that walked into GEI sequence and not just the conserved flanks were sent for sequencing using one of the primers that produced the amplicon.

2.10.4. Hemi-nested PCR (2nd round PCR) protocol

This method was applied to reduce the multiple bands usually produced with SGSP-PCR and was used in the preliminary stages of the project before we introduced the hotstart and the touchdown PCR to the SGSP-PCR and both had prove to eliminate the multiple bands better than the hemi-nested PCR. In the hemi-nested PCR the PCR products from the first round PCR (SGSP-PCR) were stabbed with a micropipette tip and transferred into a 100 µl sterile distilled water and incubated at 75°C for 15 minutes and then 1 µl of the mixture was introduced into a 2nd round PCR (hemi-nested PCR) using the same U or D primer used in the first round PCR but changing the vector primer used in the first round into a nested vector primer in attempt to amplify only the specific target. The cycle parameters for the hemi-nested- PCR were as following: after 5 min incubation at 95°C, the amplification was performed for 30 cycles of 94°C for 30 sec, annealing for 30 sec (annealing temperature was 1°C below the lowest T_m for the primers used), and extension at 72°C (extension time was chosen according to the size of the positive control product). After the last cycle, samples were incubated for 10 min at 72°C.

2.10.5. Integrase PCR

One of the characteristics of a genomic island is the presence of integrase genes, which could be found on the flanks of the island. In this approach primers were designed to amplify within certain integrase genes observed to associate with certain tRNA sites. Either the U or D primers flanking the tRNA site plus the integrase primer were used in one integrase PCR. When a PCR product is amplified, the PCR amplicon was sequenced and if the sequence analysis shows that the sequence belongs to an integrase this prove the presence of a GEI at this tRNA site. Table 2.4 shows the integrase primers used in this study.

Table 2.4. Primers used in the integrase PCR

Primer name	Primer sequence	Size (bp)
LeuX U	5`CACCACTTTATCGGCACCCATCG 3`	23
LeuX D	5`GGAGGCCCGCCATGTCACTTT 3`	23
LeuX-IR	5` GTCCACATWGMCGTTTCAAAA 3`	21
LeuX-IIR	5` AACATAAATGCCGCTGGTTC 3`	20
LeuX-IIIR	5` CTTTGCACTGCATARCGCAT 3`	20
SerW U	5` GGAGTAATGTGCCGAACCTGT 3`	21
SerW D	5` CACCGATGCGATGGAAGAGAT 3`	21
P4I- R (serW)	5` GAAGGGTCAGTCCGGTAATTC 3`	21
SsrA U	5` CCGTACCCGCAAGTTACTTCTCAA 3`	24
SsrA D	5`AGGGGTACTCGATGGCGGTCTATA 3`	24
IntA-R (ssrA)	5` TCAATCAGSCCTGTGTTCTG 3`	22

2.10.6. PCR with arbitrary primers approach (Welsh and McClelland., 1990)

In this approach, another strategy was used to interrogate tRNA sites for their GEI contents. Either the U or D primer was used as the arbitrary primer in the first round PCR. The fingerprints (multiple bands) produced by the PCR with arbitrary primers were reproducible when the PCR conditions were maintained the same. The PCR tube content was then ligated into a T-tailed vector (pBluescript II KS $^{+}$) and the cloned PCR products were introduced into a second round PCR. In this second round PCR the specific target was amplified using one of the vector primers and a nested primer located downstream of the U or upstream of D primer.

The cycle parameters for the PCR with arbitrary primers were as following (Welsh and McClelland., 1990): after 15min incubation at 95°C, the amplification was performed for 2 cycles of 94°C for 2.0 min, 40°C for 2.0 min, and 72°C for 2.0 min. This was followed with 10 cycles of 94°C for 30 sec, 60°C for 30 sec, and 72°C for 2.0 min. Then another 20 µl of a fresh master mixture were added, and a second PCR run was performed: 95°C for 5min, then 30 cycles of: 94°C for 30 sec, 60°C for 30 sec, 72°C for 2.0 min and after the last cycle the samples were incubated at 72°C for 10 min.

2.10.7. Phylogenetic group triplex PCR

The phylogenic group PCR was used as described by Duriez *et al* (2001), to classify the bloodstream infection isolates into the *E. coli* five different phylogenic groups (A, B1, B2, D and E). The phylogenetic group triplex PCR Cycle Parameters were as following: incubation at 94°C for 4min, followed by 30 cycles of: 94°C for 5 sec, 59°C for 10 sec, and 72°C for 5 min.

2.11. Steps for Primer design

The tRNA conserved flanking regions of four different genomes were used in the design of the up and the downstream primers for the tRIP strategy: *Escherichia coli* K12 (MG1655), uropathogenic *Escherichia coli* CFT073, enterohaemorrhagic *Escherichia coli* O157:H7 EDL933, and *Shigella flexneri* 2a Sf301 (Ou *et al*., 2006). To identify the conserved sequences within the four different strains, 2 kb of the upstream and downstream flanking regions were extracted and aligned using ClustalX (Thompson *et al*., 1997), (www-igbmc.u-strasbg.fr/BioInfo/), Figures 2.4 and 2.5.

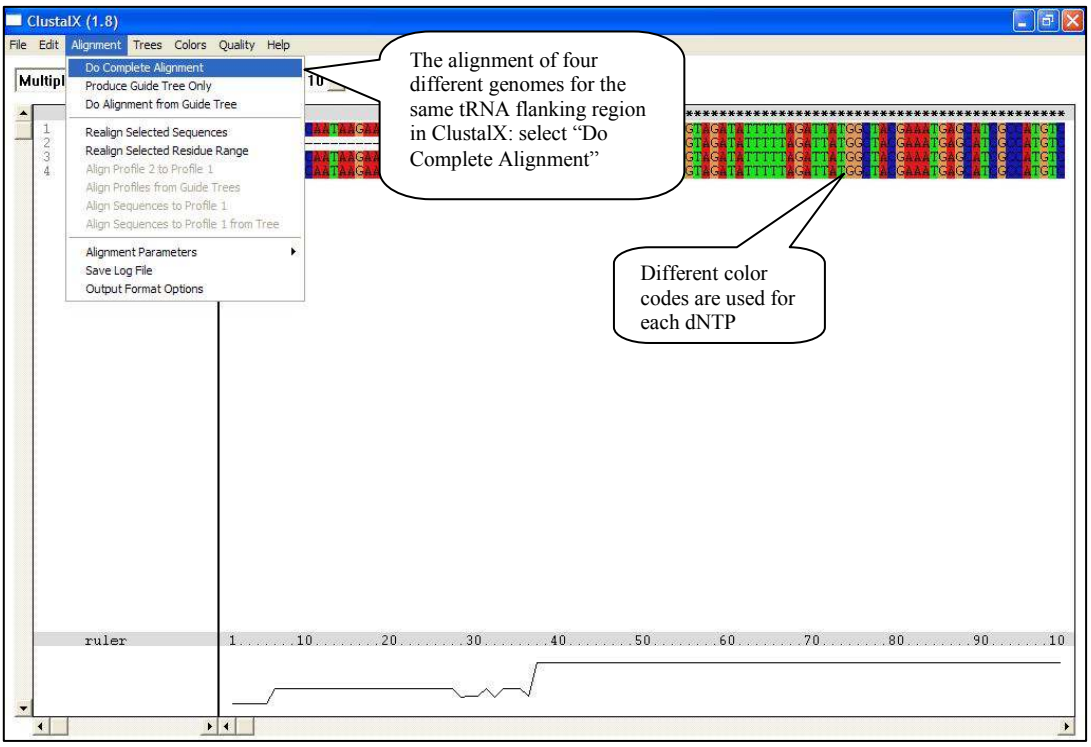


Figure 2.4. Alignment of 4 *E. coli* strain sequences using 2kb of the flanking region upstream or downstream of the tRNA gene (ClustalX).

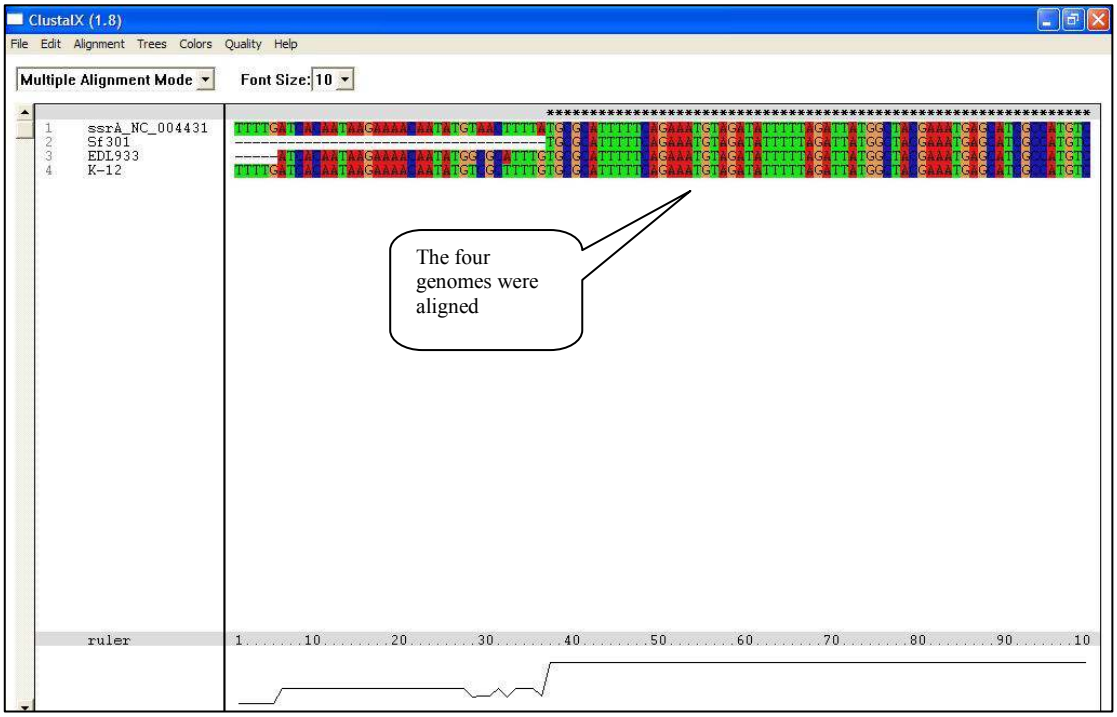


Figure 2.5. Extraction of the conserved flanking region.

The aligned sequences were then submitted to Primaclade (Gadberry *et al*., 2005) online utility (www.umsl.edu/services/kellogg/primclade-html) to design the primer, Figures 2.6 and 2.7.

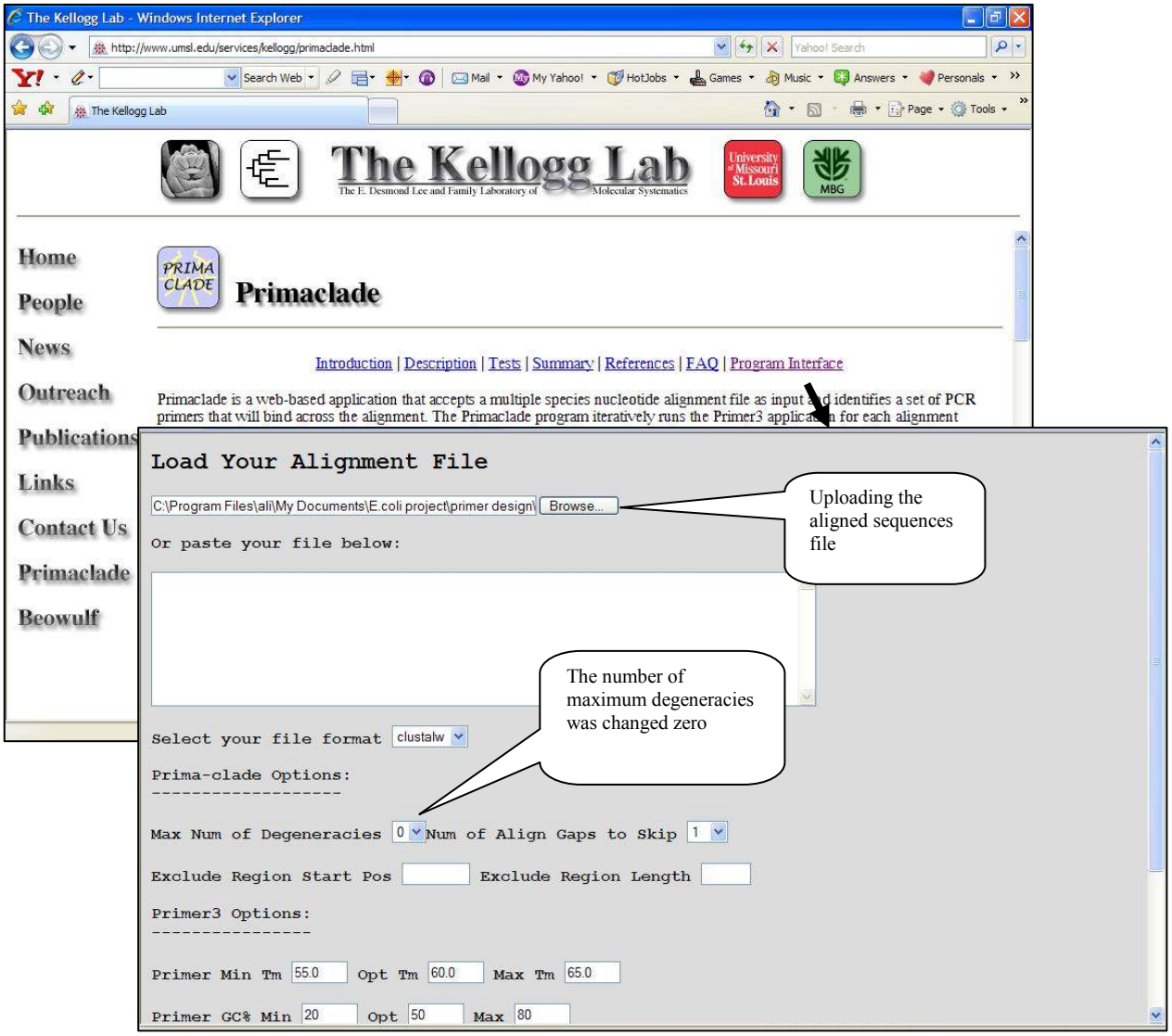


Figure 2.6. Design of the primers using Primaclade with ClustalX derived multiple sequence alignments as inputs

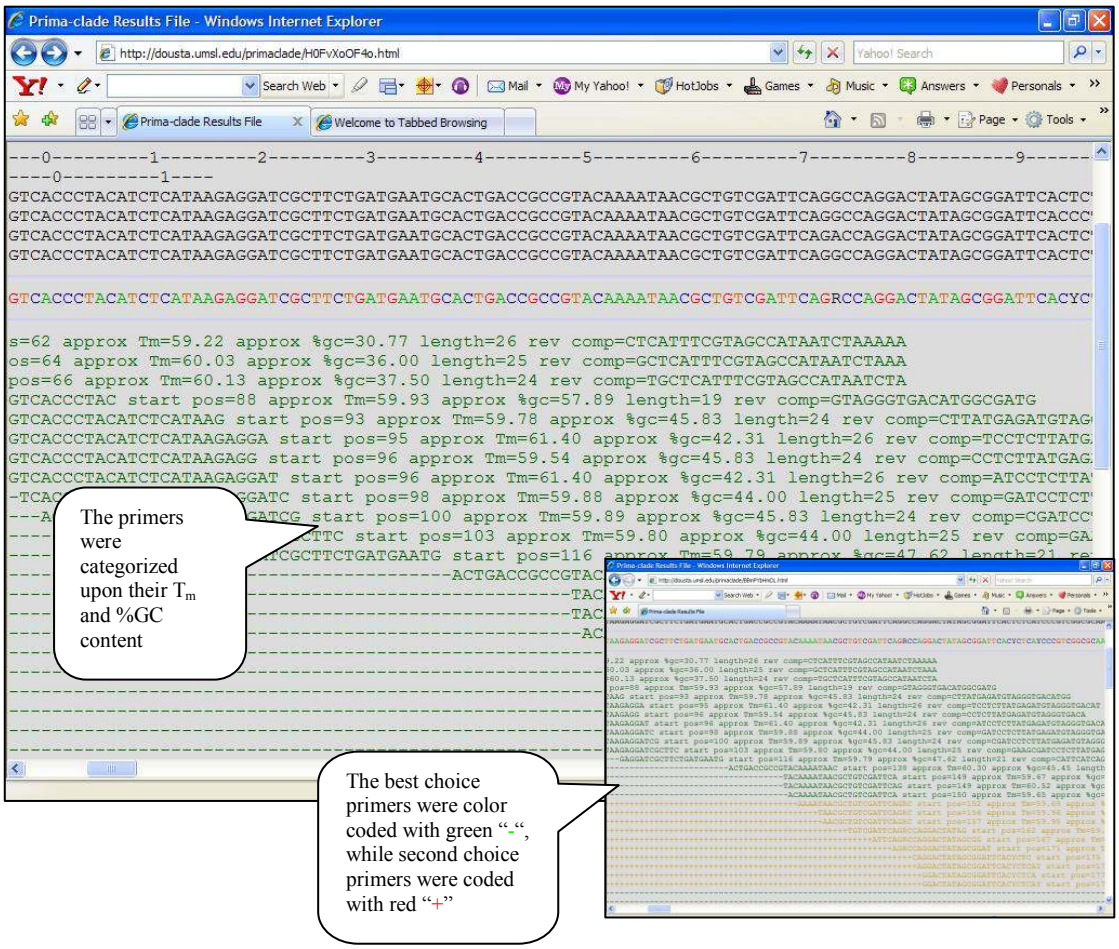


Figure 2.7. Output of the Primaclade, and selection of the possible candidate primers.

The selected primer was then checked with MapDraw from DNASTAR, Inc for the absence of restriction sites before being blasted against the sequenced genomes of *E. coli* and *Shigella* in *coli*BASE (http://colibase.bham.ac.uk) (Chaudhuri *et al*., 2004) to minimize the likelihood of non-specific amplification, Figure 2.8.

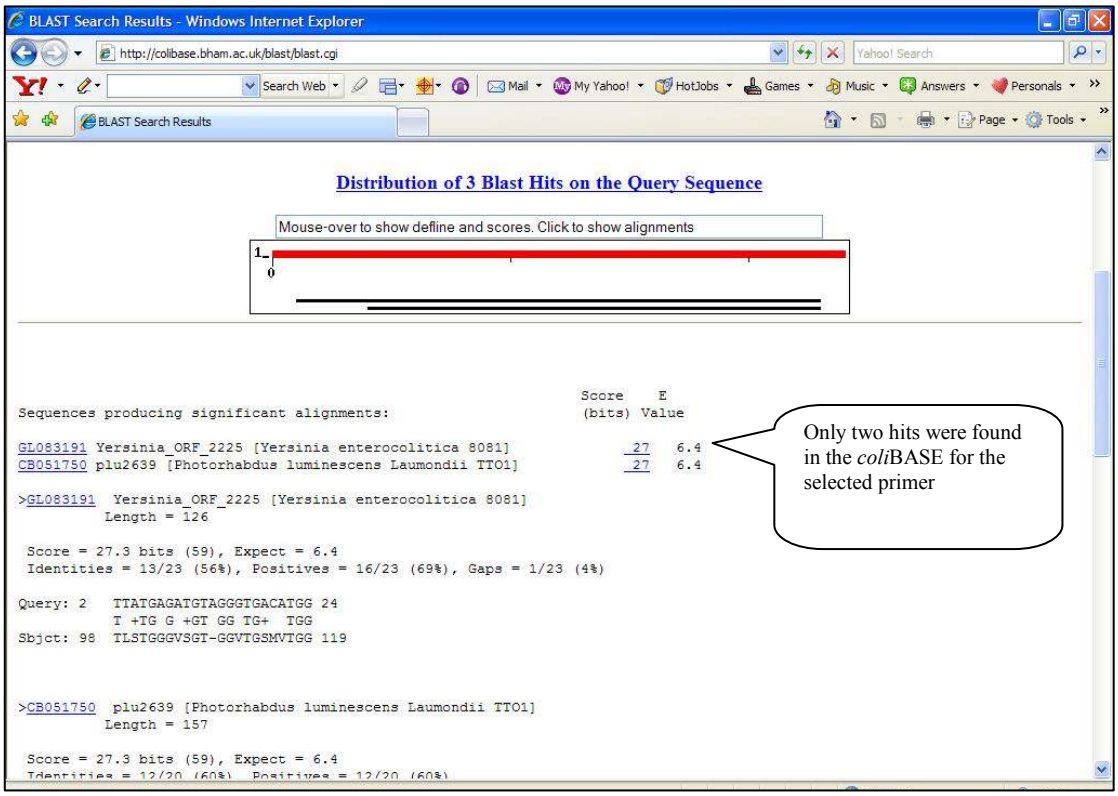


Figure 2.8. BLASTN of the candidate primers againt the *E. coli* and *Shigella* genomes.

2.12. DNA sequencing

The PCR amplicons were sequenced by AGOWA, Germany and MWG Biotech, Germany automated services

2.13. Microarray

The protocol for the microarray and the *ShE.coli* (PCR amplicon-based) metagenome microarray slides used in this study were supplied by the Molecular Microbiology Group, Institute of Food Research (IFR), Norwich Research Park, UK. The construction of the microarray slides were performed by the Molecular Microbiology group in the IFR: The CDS were amplified with specific primer pairs (Sigma-Genosys). Then, DNA from the PCRs was resuspended in a spotting solution containing 50% dimethyl sulfoxide and 0.3× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate). PCR products were spotted onto gamma amino propylsilane-coated GAPS II slides (Corning) by using a Stanford MarkII arrayer. The DNA was UV cross-linked to the slides by using a Stratalinker (Stratagene) at 300 mJ. Slides were washed in a 95ºC water bath for 2 minutes fallowed by 95% ethanol wash for 1 minute and dried by centrifugation at 185 ×g prior its storage at room temperature (Anjum *et al*., 2003; Lucchini *et al*., 2005).

The meta-array probe sequences represented 6239 CDS comprising 4264 *E. coli* K-12 MG1655 CDS, 1101 *E. coli* EDL933 CDS, 516 *S. flexneri* 2a Sf301 CDS and a further 358 virulence-associated *E. coli* CDS derived from strains representative of different pathotypes of *E. coli*, particularly enteropathogenic *E. coli* (EPEC) and enterotoxigenic *E. coli* (ETEC). Of the 358 virulence-associated CDS, 132 were located on *E. coli* chromosomes while the remaining 226 mapped onto a range of plasmids based on published literature and GenBank submissions (Ou *et al.*, 2005).

DAY 1

Genomic DNA digestion

DNA was quantified using a spectrophotometer. Genomic DNA (1μg DNA / array) was digested for 2 hours at 37°C.

DNA labelling

For *ShE.coli* array an equal mix of (MG1655) DNA and EDL933 genomic DNA was used as the comparator strain (reference strain). To each 1μg of DNA, 10 μl of 2.5x Random primers (Invitrogen) was added. Boiled for 5 minutes in a heating block. Immediately kept on ice for 5 minutes to denature the DNA and allow the primers to bind. On ice 2.5 μl 10X dNTPs were added per 1μg of DNA being labelled. 1.5 μl of the appropriate fluorescent dye was added per 1μg of DNA being labelled (Cy3 or Cy5, 1mM Stock from Amersham). 1 μl Kleenow enzyme (Invitrogen) added. Spin briefly and covered with foil and incubated in a static incubator at 37°C overnight.

DAY 2

Clean up of Labelling Reactions

Unincorporated dNTPs were removed from the labelling reactions using Qiagen PCR clean up columns and eluted in 2 x 30 μl Sigma Water. 1μg comparator was added to 1μg test DNA and samples were dried down under vacuum. Each 2μg DNA was allowed to resuspend slowly in 10 μl Sigma Water (kept in the dark to protect the light sensitive fluorescent dyes).

Slide Preparation

Slides were placed with array side up in the Stratalinker. Crosslinked using Auto function e.g., 2x (1200Jx100).

Blocking of Slides

In the fume hood, under nitrogen, 1.5g of succinic anhydride was weighed in a glass universal and dissolved in 300 ml dichloroethane DCE. 3.75 ml of n-methylimidazol was added. The slides were then incubated in the buffer for 1hr with gentle agitation to prevent the build up of bubbles on the surface of the slides.

Washes

The slides were tapped briefly and transferred quickly to 300 ml fresh DCE and incubated for 2-3 minutes. Again the slides were tapped briefly and transferred to a boiling water to denature the DNA. Then immediately the slides were transferred to 96% ethanol for 1 minute and then were centrifuged at 161 ×g for 5 minutes to dry.

Hybridisations

To each 10 µl of resuspended labelling reaction the following was added:
1.5 µl 50X Denhardts solution
2.25 µl 20X SSC (sodium chloride/sodium citrate)
1.125 µl *S. cerevisiae* tRNA (10 µg ul^{-1})
0.375 µl 1M HEPES pH7.0
0.375 µl 10% SDS

The reactions were incubated at 100ºC for 2 minutes and then left to stand on the bench for 5-10 minutes. This was followed by centrifugation at 16046 ×g for 5 minutes and then transferred to a clean microcentrifuge tube. Again the samples were spun for 5 minutes.

The slides were placed in the hybridisation chamber, and 15 µl of hybridisation solution was transferred onto the slide. Covered with a clean coverslip. 20 µl 3xSSC was applied around the edges of the slide away from the array. This is to keep the hybridisation

chamber humidity correct. The hybridisation chambers were closed and placed in a Tupperware box containing water at 63ºC. The slides were then incubated overnight at 63ºC in a pre-warmed hybridisation oven.

DAY 3

Washes

The slides were quickly transferred from the hybridisation chambers and were washed twice for 5 minutes at 63ºC in 1 L 2X SSC, 0.1% SD with vigorous stirring. After that, two washes of 250ml 1X SSC were performed at room temperature for 5 minutes each in a slide box on a platform shaker with blotting between each wash. Again, the slides were washed twice with 250 ml of 0.2X SSC before being transferred to a falcon tubes and centrifuged at 161 ×g for 5 minutes to dry. Stored protected from the light and ready to scan. The processed slides were scanned with a GenePix 4200A scanner (Axon Instruments, Inc) and the data were quantified and post-processed (normalized) by the Bluefuse software (BlueGnome, Ltd)

2.14. Pulsed-Field Gel Electrophoresis (Scott and Pitt., 2004)

The preparation of the high molecular weight DNA samples was adopted from the protocol used by the Queen's Medical Centre, Nottingham, UK.

Day 1

5 ml of LB Broth were inculcated with the *E. coli* strains for overnight culture at 37°C

Day 2

2 ml of the overnight cultures were centrifuged at 16046 ×g for 2 minutes. The pellet was resuspended in 400 µl SE buffer (75 mM NaCl, 25 mM EDTA). Then, the suspensions were mixed with 400 µl 2% pulsed-field certified agarose (BIORAD) prepared in SE buffer and immediately were transferred to mould. The moulds were set

at 4°C for 5 minutes to solidify. Then, plugs were transferred to sterile 5ml bijoux containing 3 ml of Gram-positive lysis buffer (6 mM Tris-HCl, 100 mM EDTA, 1 M NaCl, 0.5% (w/v) Brij58 (Sigma-Aldrich), 0.2% (w/v) sodium deoxycholate, 0.5% (w/v) lauroyl sarcosine, pH 7.5, 0.5 mg ml^{-1} lysozyme) and were incubated overnight at 37°C.

Day 3

The Gram-positive lysis buffer was removed, and 3 ml of Gram-negative lysis buffer (1% (w/v) lauroyl sarcosine, 500 mM EDTA, pH 9.5) were added containing 500 μg ml^{-1} proteinase K, and incubated overnight at 56°C.

Day 4

The plugs were washed in 3 ml TE buffer for 30 minutes at 4°C. This step was repeated three times. The plugs were stored at 4°C for 3-6 months, with regular changes of TE.

2.14.1. Digestion of High molecular weight DNA plugs with I-*Ceu*I and I-*Sce*I

The DNA plugs were washed twice with TE buffer for 30 minutes at 4°C. After that, the plugs were pre-incubated with 2 units of I-*Ceu*I (New England BioLabs) at 4°C for 30 minutes and then incubated at 37°C for overnight digestion. For the I-*Sce*I (BioLabs) digests the same steps were followed except that the DNA plugs were pre-incubated with the enzyme for 100 minutes at 4°C and then digested at 37°C (water bath) for 17-20 minutes.

2.14.2. Pulsed-field gel electrophoresis run conditions

I-*Sce*I run condition:

The same run conditions used by Jumas-Bilak *et al* (1995) for the separation of DNA fragments digested with I-*Sce*I were applied: switch time of 60-130 s at 6 volts/cm in a 0.8% agarose gel for 24 hours. Other run conditions used to separate I-*Sce*I digests were obtained from the BioRad CHEF-DR® II pulsed-field gel electrophoresis instruction manual and are used to separate DNA fragments in the size range of 2.4-3.0 or 3.5-5.7

Mb: switch time of 900-1200 or 1800-3600 s at 3 or 2 volts/cm in a 0.8% agarose gel for 72 or 144 hours respectively.

I-*Ceu*I run condition:

Run conditions for the size range of 40-400 kb: switch time of 20-60 s at 6 volts/cm in a 1.0% agarose gel for 24 hours, with addition of 50 μM thiourea to the Tris-HCl running buffer. Run conditions for the size range of 200-1000 kb: switch time of 70-100 s at 6 volts/cm in a 1.2 % agarose gel for 44 hours, with addition of 50 μM thiourea to the Tris-HCl running buffer. Run conditions for the size range of 1-3.1 Mb: switch time of 250-900 s at 3 volts/cm in a 0.8 % agarose gel for 144 hours, with addition of 50 μM thiourea to the Tris-HCl running buffer. When the addition of 50 μM thiourea or more does not obtain sufficient reduction of the DNA degradation the use of HEPES (16 mM HEPES-NaOH, 16 mM sodium acetate, 0.8 mM EDTA [pH 7.5]) as a running buffer instead of the Tris-HCl was an alternative option.

When HEPES buffer was used instead of Tris-HCl as a running buffer the voltage had to be reduced to 4 V/cm to keep the current within the normal range with HEPES. This is because the HEPES has a higher ionic strength than does 0.5x TBE (Koort *et al*., 2002).

Chapter 3 Interrogation of tRNA sites for their genomic islands by tRIP-PCR

3.1. Introduction

Four different hypotheses have been proposed to explain the association of the tRNA gene 3` site with pathogenicity islands. However, none of these hypotheses can explain all the aspects related to the interaction between the 3` site of the tRNA gene and the integrated pathogenicity islands. In the first hypothesis a specific tRNA is available to read codons carried by a pathogenicity island. These codons are rarely used elsewhere in the genome, making this island associated with the tRNA gene encoding this specific tRNA e.g., expression of the $tRNA^{leuX}$ is involved in the synthesis of many virulence factors present on PAI-I_{536} (Ritter *et al*., 1997). However, the hypothesis cannot explain why other tRNA genes with no preference for specific codons are associated with pathogenicity islands (Hou., 1999). The second hypothesis argues that the presence of multiple copies of tRNA genes in the bacterial genome would assist in the amplification of the virulence factors encoded by the pathogenicity islands (Cheetham and Katz., 1995). However, many other tRNA genes with only one copy in the bacterial genome are regarded as hot spots for the integration of pathogenicity islands (e.g., *selC* and *leuX*) (Hou., 1999). In the third hypothesis, it is thought that the conserved secondary structure of the tRNA genes is involved in the integration and excision of pathogenicity islands (Reiter *et al*., 1989). The 5` and 3` ends of a tRNA gene are complementary to each other forming a pair of inverted repeats. This accordingly, could facilitate the integrase-mediated recombination of a pathogenicity island with its target hotspot. However, these inverted repeats are considered short (only 7 nucleotides) compared to the inverted repeats (over 13 nucleotides) used by the integrases of bacteriophages and PAIs, and therefore, would not favour DNA recombination (Hou., 1999). The fourth hypothesis (the hybrid hypothesis) proposes that the conserved 3` CCA sequence of a tRNA gene provides the initial recognition site for integrase and that a hybrid structure of mature tRNAs hybridized to their own genes is involved in improving the integrase action (Hou., 1999). However, *in vitro* studies have shown that efficient integration of

integrase genes into the tDNA *attB*s sites could be achieved in the absence of tRNA (Williams., 2002).

To get a better understanding of the association between the tRNA genes and the integrated genomic islands more investigation regarding the identity and structure of these genomic islands is required as this could answer many questions related to the interactions between a tRNA and a PAI. In this study we utilized a PCR-based screening method to locate putative genomic islands (GEIs) lying downstream of tRNA genes, known hotspots for the integration of horizontally acquired DNA (Hacker and Kaper., 2000). Sixteen tRNA genes that had been shown to harbour islands in one of four sequenced reference strains (MG1655, CFT073, EDL933, Sf301) were selected for interrogation. The cognate tRNA gene loci in ten *E. coli* BSI strains were investigated by tRNA site Interrogation for Pathogenicity islands (tRIP) PCR (Ou *et al*., 2006) for the presence or absence of GEIs and the results were confirmed by chromosome walking and sequencing. The aim was to identify tRNA-associated GEI repertoires of individual *E. coli* BSI-associated strains and to discover novel GEIs.

As stated in the materials and methods the tRIP-PCR is designed to screen large number of strains for their genomic island (GEI) content by amplifying the flanking regions of a tRNA gene. A positive tRIP-PCR indicates absence of genomic island while a negative tRIP-PCR highlights the possibility of GEI insertion. The tRIP-PCR primers were designed by aligning the conserved flanking regions of four to five reference strains (*E. coli* K-12 MG1655, *E. coli* UPEC CFT073, *E. coli* O157:H7 EDL933, *S. flexneri* 2a SF301 and *S. flexneri* 2a 2457T). Primers selected should produce PCR products in the size range of 1.5 to 4 kb, have a melting temperature (T_m) between 55°C to 65°C, and should not anneal at multiple sites in the reference chromosomes or the vector sequence. For some tRNA sites, the conserved flanking regions in all five-reference strains are separated by sequence of a genomic island. This has been achieved by the tRNAcc method for GEI identification and confirmed by *in silico* PCR (Ou *et al*., 2006). In this case, the primers were designed for a proposed empty site.

3.2. Optimization of the tRIP-PCR

The amount of DNA and the number of units of DNA polymerase added to the PCR reactions were both optimized in order to achieve high throughput PCR amplification and reduce the cost of the reagents used in the PCR reactions. When the number of Taq polymerase units was fixed at 1 U/reaction, the PCR was able to amplify down to 2 ng of template DNA (Figure 3.1). On the other hand, the threshold for the number of Taq units at which a successful PCR run could be achieved was 0.75 U/ reaction (Figure 3.1). The result of the Taq polymerase unit dilution was consistent with three different tRNA sites (*asn*V, *asp*V and *ser*W). Fallowing this experiment ~100 ng of template DNA and 2 units of Taq® DNA polymerase were used for subsequent PCR reactions. The results of the tRIP-PCR for the 16 tRNA sites of the ten clinical strains are summarized in Table 3.1.

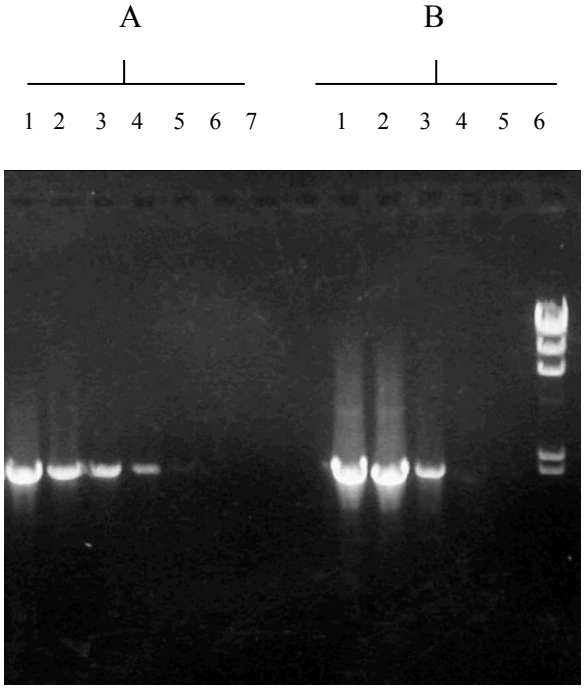


Figure 3.1. Optimization of tRIP-PCR. Panel (A): Dilution of DNA template. Lanes (1 U/reaction of Taq® DNA polymerase was used in all reactions): 1, dilution of 10^{0} (2000 ng); 2, dilution of 10^{-1} (200 ng); 3, dilution of 10^{-2} (20 ng); 4, dilution of 10^{-3} (2 ng); 5, dilution of 10^{-4} (0.2 ng); 6, dilution of 10^{-5} (0.02 ng); 7, negative control. Panel (B): dilution of the Taq® DNA polymerase (ABgene). Lanes (2000 ng of template DNA was used in all reactions): 1, 2U/reaction of Taq® DNA polymerase; 2, 1U/reaction; 3, 0.75U/reaction; 4, 0.5U/reaction; 5, 0.25U/reaction; 6, Lambda DNA/*Hin*dIII marker (Fermentas).

Table 3.1. Matrix of Results for (tRIP-PCR)

Strain	Phylogenetic group	tRNA site[a]															
		pheU	*serT*	*selC*	*glyU*	*serU*	*serW*	*asnV*	*thrW*	*metV*	*serX*	*argW*	*leuX*	*asnT*	*aspV*	*pheV*	*ssrA*
MG1655	A	+	+	+	0.9	+	+	+	–	+	1.2	–	–	–	+	–	–
E102	D	+	+	+	–	1.2	1.8	+	1.7	+	–	–	–	–	–	–	–
E103	D	+	+	+	0.9	1.2	1.8	+	–	+	–	–	–	–	–	–	–
E104	D	+	+	1.5	–	1.2	–	+	1.7	–	–	–	–	–	–	–	–
E105	B2	+	–	–	0.9	–	–	–	–	–	–	–	–	–	–	–	–
E106	B2	+	+	3.2	0.9	1.2	–	–	1.7	–	–	–	–	–	–	–	–
E107	D	+	+	+	–	1.2	+	+	–	+	0.3	–	–	–	–	–	–
E108	B2	+	+	+	0.9	–	–	–	1.7	–	1.2	1.8	1.7	0.5	–	–	–
E109	B2	+	+	–	0.9	–	+	–	1.7	–	1.2	0.5	1.6	0.5	–	–	–
E110	B2	+	+	–	0.9	1.2	+	–	–	–	–	–	–	–	–	–	–
E111	B2	+	+	+	0.9	–	–	+	–	–	–	1.8	–	–	–	–	–
		The data shown below is taken from *in silico* data[b]															
Empty tRNA site		0.6	0.3	1.0	0.7	0.4	1.1	2.0	0.3	1.4	0.4	1.1	1.1	0.5	0.7	1.0	1.0
MG1655	A	0.6	0.3	2.9	12.4	1.8	1.4	2.0	40.2	1.4	0.9	13.7	41.2	10.6	3.1	10.1	30.6
CFT073	D	52.8	0.3	69.6	0.8	23.5	1.4	–	8.0	34.0	–	15.7	17.0	37.8	100.7	128.9	–
EDL933	D	0.6	45.6	44.6	28.4	47.0	89.0	2.0	35.5	1.4	87.9	15.3	45.5	11.6	37.6	–	30.2
Sf.301	N/A[c]	0.6	0.3	–	10.7	22.7	1.5	2.0	–	1.3	0.4	6.5	8.6	5.0	58.4	56.1	4.6

[a]'+', positive tRIP-PCR result with PCR amplicon size equal to the predicted size obtained by the positive control MG1655. PCR amplicons with different size than the positive MG1655 were indicated. '–', negative tRIP-PCR result.
[b]The data was taken from Ou *et al* (2006) and it showed the predicted size between the primer pair. The size was shown to the nearest 0.1 kilobase (kb). '–', denotes no match was found for the primers used.
[c]N/A, 'not available', the PCR result of the phylogenetic grouping for this strain was not performed.

3.3. Summary of the tRIP screening method

Table 3.1 had revealed some important findings regarding the tRNA sites investigated. The following tRNA sites (*argW, metV, serX, asnT, leuX, aspV, pheV* and *ssrA*) had negative tRIP-PCR results for at least 7/10 strains investigated, meaning that these sites could be hot spots for genomic islands and other mobile genetic elements integration. The total number of negative tRIP-PCR was (93/160), and therefore, the percentage of strain-tRNA sites putative positive for genomic islands (GEIs) was 58%. Six distinct major profiles (including MG1655) were obtained using the tRIP-PCR strategy with the following groups (E102), (E104), (E107), (MG1655 and E103), (E105, E106, E110 and E111) and (E108 and E109) having similar tRIP profiles with minor differences (Figure 3.2). In order to achieve such clustering the results of the 11 original profiles were used as inputs into ClustalX (Thompson *et al*., 1997) and the results were viewed in TreeView. Because ClustalX is used to align DNA sequences, a positive tRIP-PCR was referred to as A and a negative tRIP-PCR as T (Table 3.2). However, such approach of classifying the tRIP-PCR profiles would miss the unique and mosaic structure represented by the different GEIs integrated downstream of the same tRNA site. To overcome such limitation, a more robust method would classify the tRIP-PCR results into different profiles using the sequence of the GEIs boundaries (Table 8.3, Appendix) and/or the restriction fragment length polymorphisms (RFLP) (Table 8.4, Appendix) obtained by the SGSP-PCR. But, because the RFLP and the sequence of the GEIs boundaries obtained by the SGSP-PCR were in most cases obtained for only one of the GEI boundaries, such comparison between different GEIs integrated downstream of the same tRNA site was difficult.

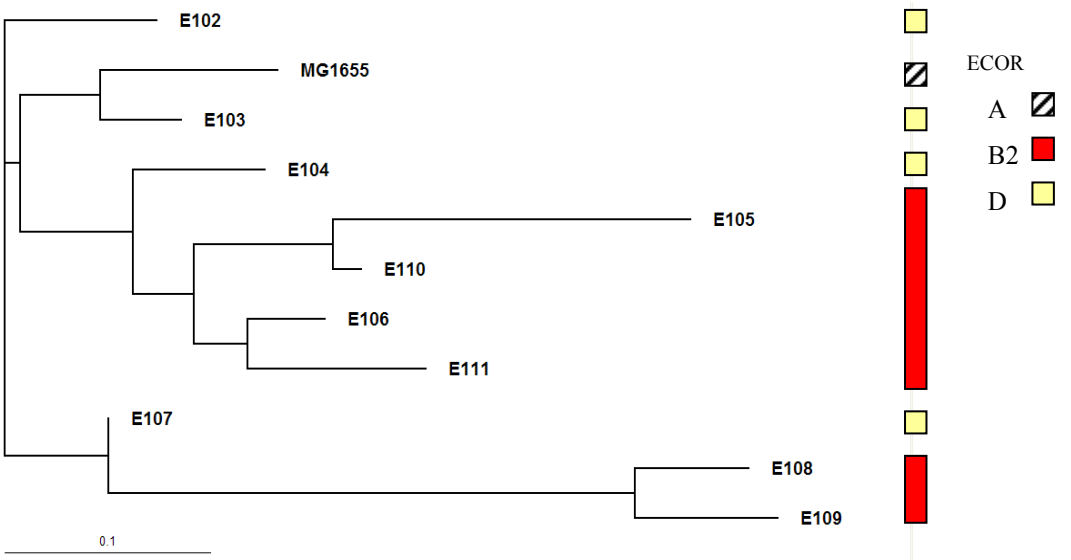


Figure 3.2. Dendrogram of the tRIP-PCR results for the investigated *E. coli* BSI-associated strains. The dendrogram was obtained by introducing the tRIP-PCR results into ClustalX followed by viewing the aligned profiles in TreeView. The distance scale is shown at the bottom of the figure. Six classes could be distinguished at a distance of 0.1, these were indicated in the text above. The phylogenetic (ECOR) group is presented at the right of the dendrogram.

Table 3.2. Alignment of the tRIP-PCR results in ClustalX

Strain	Multiple alignment of tRIP-PCR profiles[a]															
MG1655	A	A	A	A	A	A	A	T	A	A	T	T	T	A	T	T
E102	A	A	A	T	A	A	A	A	A	T	T	T	T	T	T	T
E103	A	A	A	A	A	A	A	T	A	T	T	T	T	T	T	T
E104	A	A	A	T	A	T	A	A	T	T	T	T	T	T	T	T
E105	A	T	T	A	T	T	T	T	T	T	T	T	T	T	T	T
E106	A	A	A	A	A	T	T	A	T	T	T	T	T	T	T	T
E107	A	A	A	T	A	A	A	T	A	A	T	T	T	T	T	T
E108	A	A	A	A	T	T	T	A	T	A	A	A	A	T	T	T
E109	A	A	T	A	T	A	T	A	T	A	A	A	A	T	T	T
E110	A	A	T	A	A	A	T	T	T	T	T	T	T	T	T	T
E111	A	A	A	A	T	T	A	T	T	T	A	T	T	T	T	T

[a]Because the aligned profiles refer to results of tRIP-PCR (+/–), no "complete alignment" by ClustalX was done on the aligned results as this was thought to change the position of the aligned profiles and would not allow a comparison for the results of each tRNA site. On the other hand, changing the position of tRIP-PCR results for an entire column (tRNA) had produce the same profiles obtained in (Figure 3.2) with small changes.

One particular BSI-associated strain (E105) showed negative tRIP-PCR results for 14 tRNA sites (Table 3.1). This BSI and five more clinical strains (E104, E106, E108, E109, E110) were retested by API 20E for being *E. coli* and were confirmed to be *E. coli* with ID numbers above 98% for all of them except for E109 which showed 89.6%

for being *E.coli* and 10.6% for *Kluyvera*. Both *strains* E105 and E109 were sorbitol (API 20E) negative suggesting that they might be *E. coli* O157. These two were further investigated by the latex test for the O antigen of *E. coli* O157and found to be negative.

For some of the examined strain-tRNA sites, the PCR products were either larger or smaller than the expected products of the positive control (MG1655). Table 3.1 highlights these PCR products and their PCR amplicon sizes. Representatives of these larger or smaller amplicons were sent for sequencing using either the up or down-stream primers used in the tRIP-PCR. Out of 7 representative amplicons sent for sequencing, two small GEIs (~1.5kb and ~3.2kb) were revealed after sequence analysis for the tRNA sites E104_*selC* and E106_*selC* respectively. The other 5 amplicons represented the following tRNA sites (E102_*serU*, E108_*argW*, E109_*argW*, E107_*serX*, E104_*thrW*). Analysis of these 5 sequences revealed that they were part of the conserved flanks of the specific tRNA site. An interesting notice about these larger or smaller amplicons is that some of them resemble the size of amplicons produced by tRIP-PCR for an empty tRNA site.

3.4. The association between tRIP-PCR profiles and the phylogenetic groups for the BSI strains

According to phylogenetic grouping, *E. coli* strains are classified into five main phylogenetic groups (A, B1, B2, D and E) (Selander *et al*., 1987; Herzer *et al*., 1990). Classifying *E. coli* strains into different phylogenetic groups can either be done by the multilocus enzyme electrophoresis (Selander *et al*., 1986; Herzer *et al*., 1990) or by ribotyping (Bingen *et al*., 1994; Bingen *et al*., 1996; Bingen *et al*., 1998). However, both of these reference methods had proved to be complex and time-consuming. In this study, we used a PCR based method to determine the phylogenetic groups of the *E. coli* strains investigated in this study (Clermont *et al*., 2000; Duriez *et al*., 2001). This technique which is rapid and simple uses a combination of two genes. The first is *chuA*, which is involved in the heme transport in O157:H7 (Mills and Payne., 1995; Whittam., 1996; Torres and Payne., 1997; Bonacorsi *et al*., 2000). The second is *yjaA*, which is a gene with unknown function identified in the complete genome sequence of *E. coli* K12

(Blattner *et al*., 1997) and anonymous DNA fragment designated TSPE4.C2 (Bonacorsi *et al*., 2000).

The method by which the triplex-PCR results were interpreted is represented in Figure 3.3. A positive PCR results for the gene *chuA* indicates that the tested strain belongs to either group B2 or D while a negative PCR result for this gene indicates that the strain belongs to either group A or B1. Groups B2 and D are then discriminated by the PCR result for the gene *yjaA*. A positive PCR result for *yjaA* indicates that the strain belongs to group B2 and a negative result indicates that the strain belongs to group D. In the same manner, groups A and B1 are discriminated by the PCR results for the DNA fragment TSPE4.C2. A positive PCR for TSPE4.C2 indicates that the strain belongs to group B1 and a negative PCR indicates that the strain belongs to group A.

Our results correspond with observations obtained by other groups. For the positive controls MG1655 and EDL933, the phylogenetic triplex-PCR results were as following: MG1655 belongs to group A and EDL933 to group D (Table 3.1, Figure 3.4), the same results were obtained by previous studies (Clermont *et al*., 2000; Duriez *et al*., 2001). The *E. coli* BSI-associated strains were assigned to group B2 and to a lesser extent to group D which also corresponds with observations obtained by previous studies (Picard and Goullet., 1988; Maslow *et al*., 1995; Boyd and Hartl., 1998; Bingen *et al*., 1998; Boyd and Hartl., 1998; Picard *et al*., 1999; Johnson and Stell., 2000) (Table 3.1, Figure 3.4).

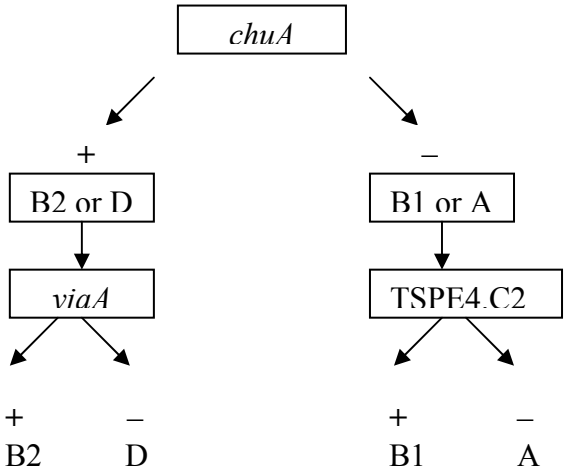


Figure 3.3. Phylogenetic group decision tree. The phylogenetic group determination of an *E. coli* strain by using the results of PCR amplification of the *chuA* and *yjaA* genes and the DNA fragment TSPE4.C2.

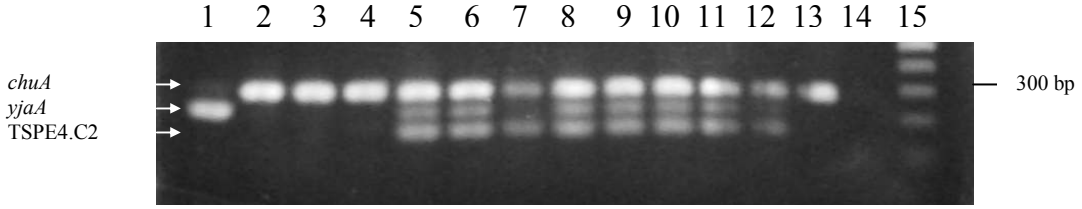


Figure 3.4. Phylogenetic triplex PCR results. Lane 1, MG1655 (group A). Lanes 2-4, 7, 12, 13 (E102, E103, E104, E107, CFT073 and EDL933 respectively) (group D). Lanes 5, 6 and 8-11 (E105, E106, E108, E109, E110, E111 respectively) (group B2). Lane 15, GeneRuler 100 bp DNA ladder (Fermentas). The size of the amplified fragments are as following: *chuA* (279 bp), *yjaA* (211 bp), and TSPE4.C2 (152 bp).

When comparing the results of the phylogenetic grouping with the tRIP profiles (Figure 3.2), an interesting overlap between the tRIP profile clustering and the B2 ECOR (phylogenetic) group was obtained. The strains were separated into six main tRIP profiles (clusters): clusters 1, (E102); 3, (E104); and 5, (E107) contained one strain each, in which all of them belong to group D, cluster 2 contained 2 strains; MG1655 which belongs to group A and E103 which belongs to group D, clusters 4, (E105, E106, E110 and E111); and 6, (E108 and E109) contained only B2 strains (Figure 3.2).

3.5. Analysis of the E104_*selC*

A 1.5 kb PCR product was produced when E104_*selC* was interrogated by tRIP-PCR. The PCR product was smaller than the 2.9 kb PCR product produced by MG1655_*selC* (positive control). The 1.5 kb band was sent for sequencing using the downstream primer (#D) and a 525 bp clipped sequence was obtained after the sequencing. The blastn results of the 525 bp indicated that 489 bp of the sequence hits into a small island in the *selC* site (Figure 3.5), identified in *E. coli* E24377A by the tRNAcc method (Ou *et al*., 2006). The size of the small island in E24377A was 2.9 kb. Additionally the other 33 bp fragment (493-525) matches the 5`-end of the insertion element (IS1X1) with 100% identity. The IS1X1 element is not represented in Figure 3.5 because it hits in different part of the genome of E24377A and the 33 bp matches the 5`-end of the IS1X1 element and not within the IS element.

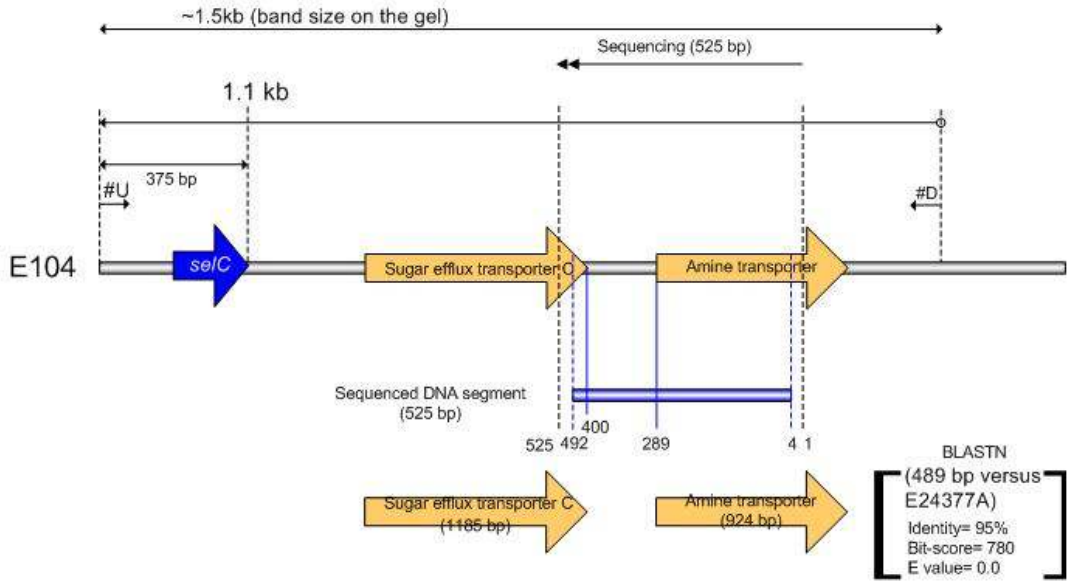


Figure 3.5. The tRIP-PCR GEI of E104_*selC*.

3.6. Analysis of the E106_*selC*

tRIP-PCR analysis of E106_*selC* produced a 3.2 kb product, which is larger than the expected PCR product obtained by the positive control MG1655 (2.9 kb). Sequencing the PCR product using the (#U) primer revealed that the sequence matches the upstream conserved flank and *selC* gene and walk into the sequence of small GEIs identified by the tRNAcc method in the *E. coli* strains APEC O1 and UTI89 at the *selC* site. The size of the *selC* GEI in APEC O1 and UTI89 was 4.95 kb in both strains (Figure 3.6).

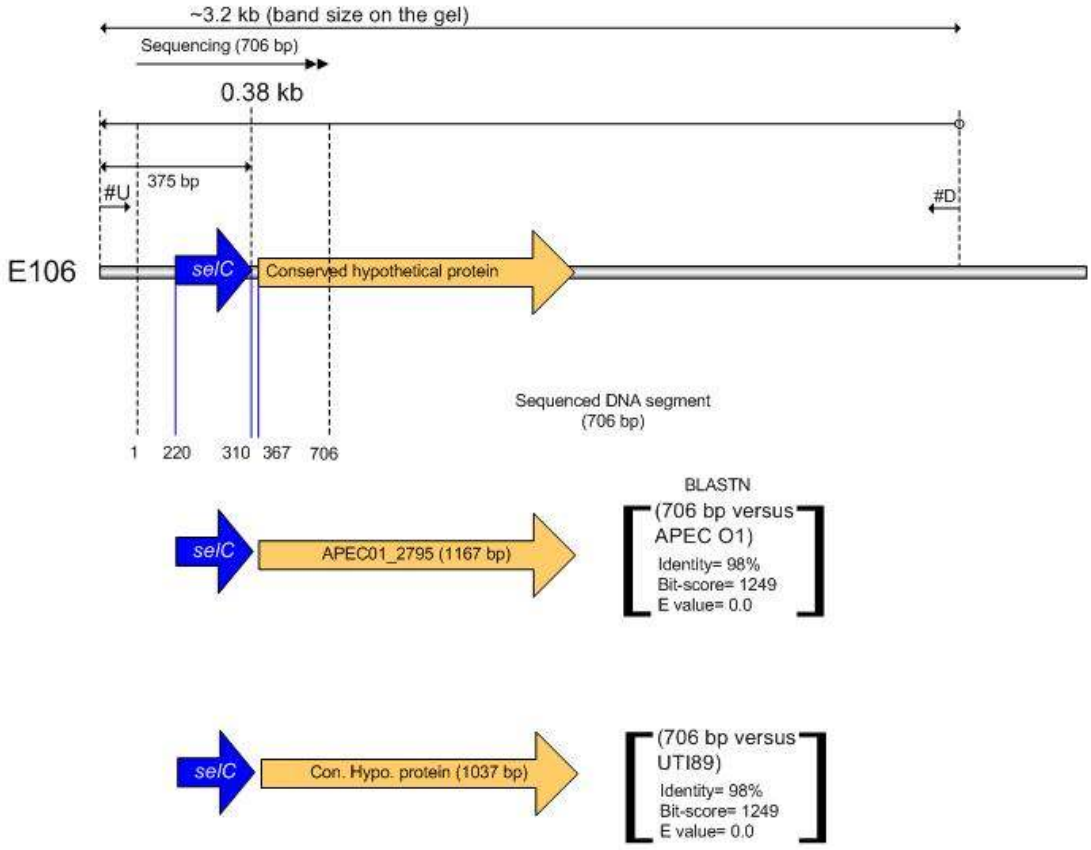


Figure 3.6. The tRIP-PCR GEI of E106_*selC*.

3.7. Single Genome Specific Primer-PCR (SGSP-PCR) results

As stated in the materials and methods, eight different genomic libraries (enzymes used: *Bam*HI, *Eco*RI, *Eco*RV, *Hin*cII, *Hin*dIII, *Kpn*I, *Sal*I, *Pst*I) were constructed for the ten clinical BSI strains. These libraries consisted of genomic and vector DNA (pBluescript II KS^{+}) restricted with the same enzyme and then ligated with T4 ligase. The ligation mixtures were introduced into competent cells by electroporation and cells containing recombinant DNA molecules were identified by the blue/white Lac selection method. The transformation efficiency was 1.1 x 10^5 transformants/μg DNA and about 60 % of the colonies were white. Success of the recombination step was checked by selecting few white colonies for plasmid DNA extraction prior to digestion with the same restriction enzyme used in the construction of the genomic libraries e.g., *Hin*dIII, (Figure 3.7). Digestion of the recombinants with the same restriction enzyme used for constructing the genomic libraries should produce two fragments, the first band 2961 bp is the vector DNA, and a second band belongs to the cloned genomic fragment. The size of the cloned DNA molecules ranged between 1.0 to 2.3 kb. The sizes of these cloned DNA fragments were rechecked by digesting the same recombinants in Figure 3.7 with the unique *Eco*RI site of the vector (Figure 3.8) unless there was another *Eco*RI site in the insert DNA and the results of both gel sizing were summarized in Table 3.3. Only one discordance was noticed when comparing between the two digests; in colony number 4 (Table 3.3), the size of the recombinant molecule was 3.9 and 7.3 for *Hin*dIII and *Eco*RI digests, respectively. This discordance could be due to a duplicate band at 3 kb in Figure 3.7, the duplicate band would be the result of a *Hin*dIII site present in the cloned DNA fragment.

After checking that the recombination and cloning steps were successful, the ligation mixture was used as a template DNA for SGSP-PCR, in which, either U or D primers of the investigated tRNA site was used with one of the vectors primers to amplify the genomic DNA fragment inserted within the multiple cloning site of the vector. Genomic island sequence was revealed by first, sequencing the PCR amplicon using one or two of the primers used to amplify the amplicon. Then, the sequenced SGSP-PCR product was analysed for putative GEI boundaries.

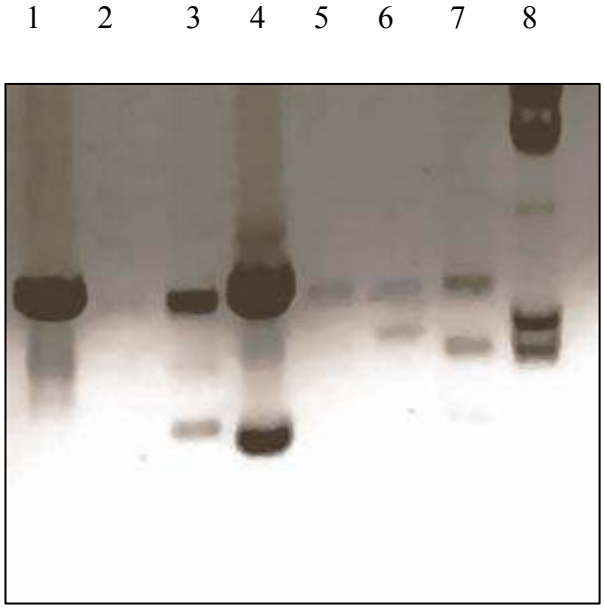


Figure 3.7. Run of recombinant DNA molecules, digested with *Hin*dIII.Lanes: 1, digest of pBluescript II KS^+ plasmid (used as a negative control); 2 and 5, digests of vector DNA extracted from two different blue colonies, both showed weak vector band at 3 kb with no insert of genomic DNA; 3, 4, 6 and 7, digests of recombinant DNA molecules extracted from different white colonies, all showed the vector band at 3 kb and a genomic DNA insert with size range between 1-2.3 kb. Lane 8: Lambda DNA/*Hin*dIII marker (Fermentas).

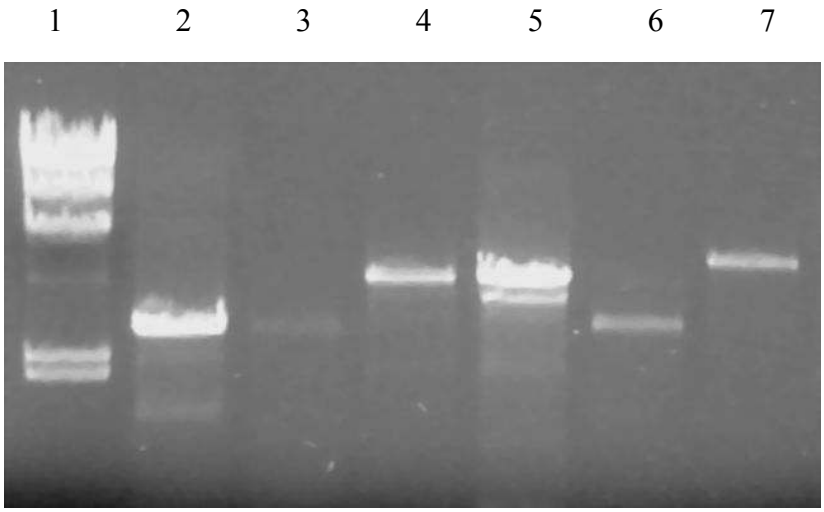


Figure 3.8. Run of recombinant DNA molecules, digested with *Eco*RI. Lanes: 1, Lambda DNA/*Hin*dIII marker (Fermentas), 2, digest of pBluescript II KS^+ plasmid (used as a negative control), lanes 4, 5 and 7: digests of recombinant DNA molecules extracted from different white colonies, all showing a linearized recombinant DNA molecules with size range between 3.5-6 kb. Lanes 3 and 6: digest of vector DNA extracted from two different blue colonies, showing the vector band at 3 kb with no insert of genomic DNA.

Table 3.3. Comparing between sizes of the same recombinant molecules digested with either *Hind* III or *Eco*RI, refer to Figures 3.8 and 3.9 for details about the digested molecules.

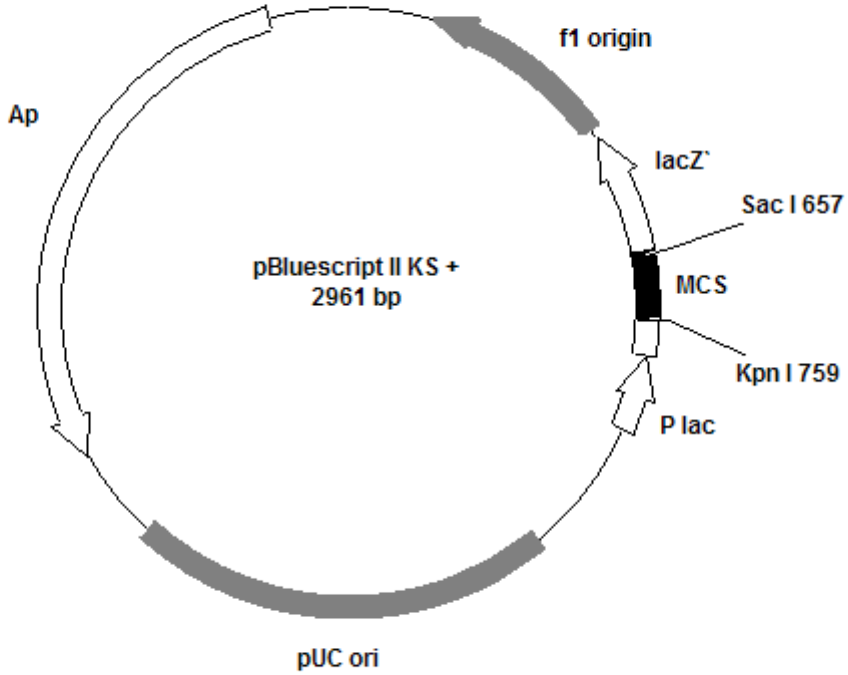
Colony/lane no.[a]	Size of recombinant digests (*Hin*dIII library)[b]			Size of recombinant digests (*Eco*RI library)[b]
	Vector (~kb)	Insert (~kb)	Total size (~kb)	Total size (~kb)
1 (pBluescript II KS^+)	3.0	–	3.0	3.0
2 [blue]	3.0	–	3.0	3.0
3 [white]	3.0	1	4.0	4.0
4 [white]	3.0	0.9	3.9	3.5 and 3.8 [7.3][c]
5 [blue]	3.0	–	3	3
6 [white]	3.0	2.5	5.5	5.3
7 [white]	3.0	2	5	N/D[d]

[a]The color of the colony is indicated between square brackets.
[b]The sizes are shown to the nearest 0.1kb.
[c]The total size of both fragments is indicated between square brackets.
[d]The size was not determined for this recombinant molecule.

3.8. Optimization of the SGSP-PCR

I have optimized the SGSP-PCR protocol several times before it became a robust method. The first SGSP-PCR protocol involved using asymmetric amounts of the tRNA gene-specific and vector-specific primers. A ten fold of the vector primer to tRNA gene primer was used in the PCR to compensate for the low ratio of vector to insert (1:20) used in the ligation step. However, this had resulted in the production of multiple and non-specific bands. The problem was resolved by adding equal amounts of both primers. The second step in the optimization process was to reduce the background smearing produced from the highly dense ligation mixture, to do so; a serial dilution of the ligation mixture was used as a template for the SGSP-PCR. ¼ dilution of ligation mixture was adequate for non-smeared background.

MG1655 genomic libraries were used as a positive control for SGSP-PCR experiments, in which comparing the SGSP-PCR results of MG1655 with its SGSP-PCR *in silico* data identified non-specific products. The non-specific bands were identified and confirmed to be non-specific by band stabbing and reamplifying in a second round PCR using hemi-nested PCR approach. In this approach, a nested primer (e.g., T3) from the vector multiple cloning site was used instead of the one (e.g., M13R) used in the first round of SGSP-PCR (Figure 3.9). Using this strategy of both SGSP-PCR followed by the hemi-nested PCR had proved to be successful in the identification of non-specific bands, as these bands were not amplified in the second round PCR. Figures 3.10-D and E, illustrate examples for the identification of non-specific bands in SGSP-PCR. In Figure 3.10-D, the SGSP-PCR products of the strain-tRNA site MG1655-*serU* were compared to its *in silico* PCR products (Table 8.2, Appendix). The non-specific PCR products were identified to be the SGSP-PCR products for the following genomic libraries: *Eco*RI (1 kb), *Sal*I (1.7 kb), and *Bam*HI (2.3 kb), that is because the *in silico* SGSP-PCR results for the same libraries were: 10.5, 11.4 and 7.5 kb respectively. These non-specific bands were confirmed to be non-specific by a second round PCR using hemi-nested primers (Figure 3.10-E). Only the specific band (1.5 kb) of the genomic library *Hin*dIII was amplified. Two more optimization steps were introduced to secure a robust SGSP-PCR strategy; touchdown and hot start approaches. When these two approaches were applied to SGSP-PCR, only specific bands were amplified making no need to check for band specificity by a second round PCR (Figure 3.10-F). Details of the optimization steps are summarized in (Figure 3.10).



PBluescript II KS $^{+}$ Multiple Cloning Site Regions
(Sequence shown 598-826)

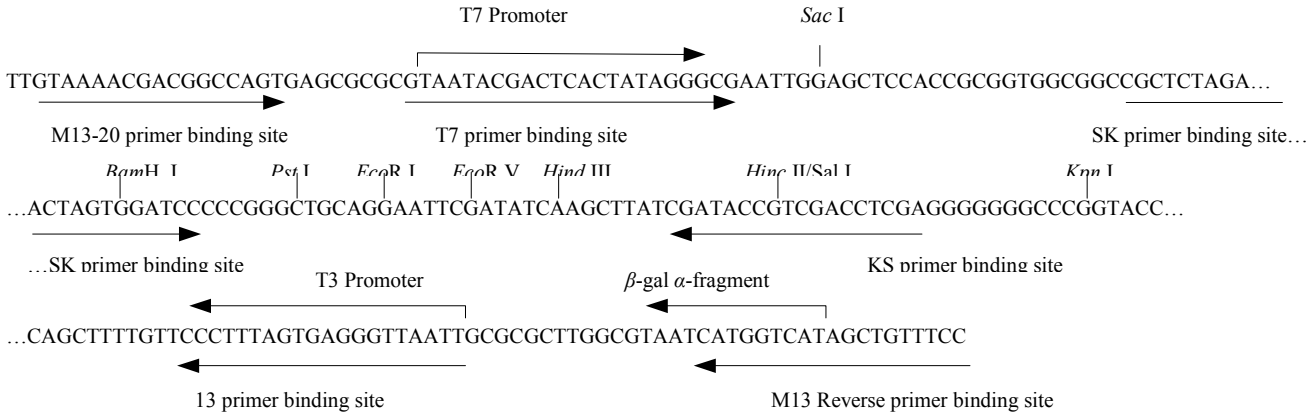


Figure 3.9. pBluescript II KS^{+} (Stratagene).

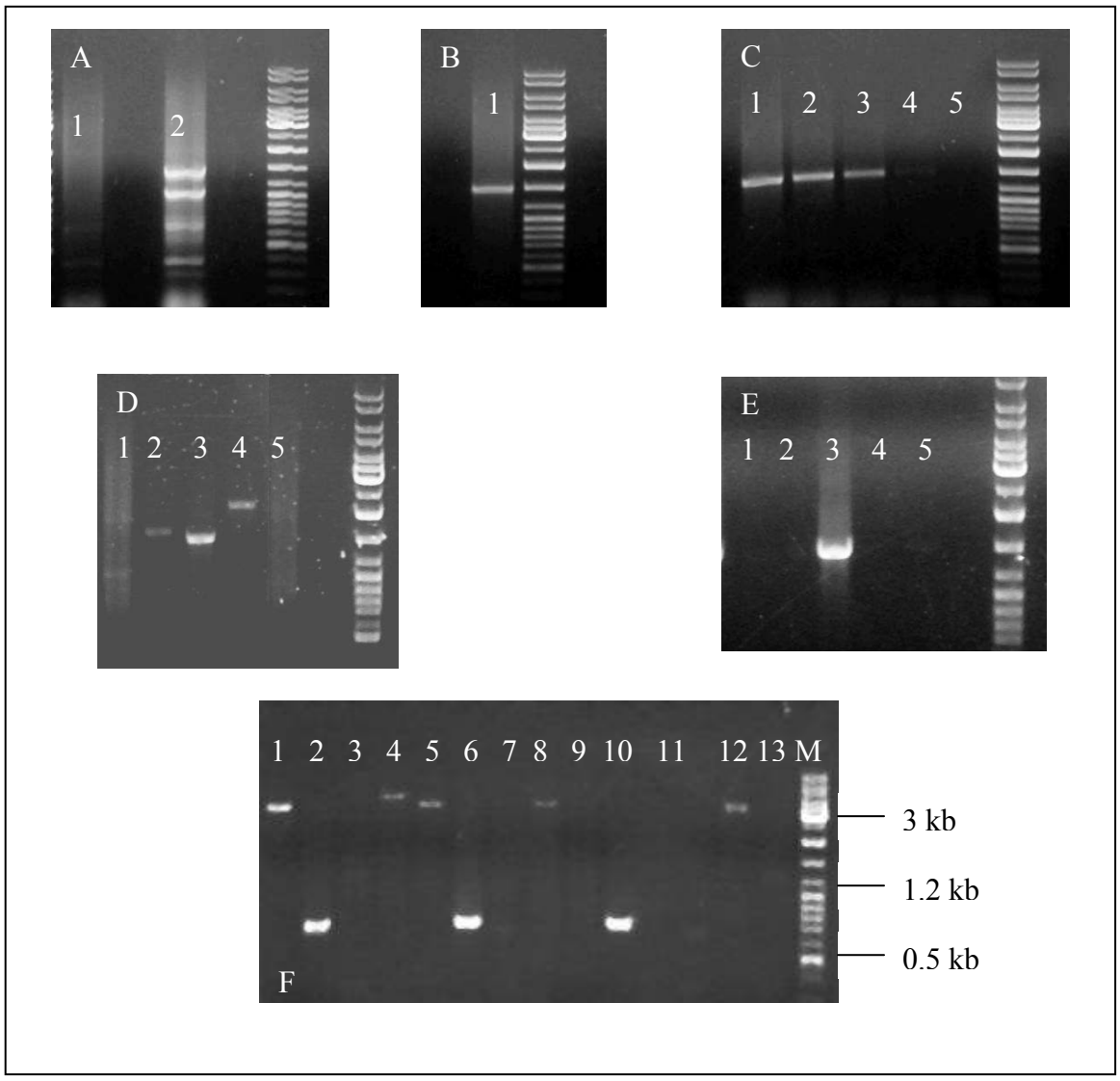


Figure 3.10. SGSP-PCR optimization. A: SGSP-PCR with asymmetric amounts of vector to tRNA gene primers; 200 pmol : 20 pmol. Lanes: 1, negative control (water added instead of DNA); 2, SGSP-PCR result for the strain-tRNA site (MG1655-*serU*) using *Hin*dIII genomic library. B: SGSP-PCR with equal amounts of vector to tRNA gene primers, the same strain-tRNA site and genomic library used in (A). C: Making dilution of the ligation mixture for the same strain-tRNA site and genomic library used in (A). Lanes: 1-5, ½ dilution, ¼ dilution, 10^{-1} dilution, 10^{-2} dilution, negative control, respectively. ¼ dilution was selected for subsequent SGSP-PCR experiments to reduce the amount of background smearing. D: The SGSP-PCR results of four different genomic libraries (lanes: 1, *Eco*RI; 2, *Sal*I; 3, *Hin*dIII; 4, *Bam*HI) for MG1655-*serU* site. Lane 5, negative control. E: Second round PCR (hemi-nested approach) to test for SGSP-PCR specific/non-specific bands. The template DNA samples were stabbed PCR products of the bands in (D). Lanes: 1, stabbed PCR product of lane 1 (*Eco*RI genomic library); 2, stabbed PCR product of lane 2 (*Sal*I genomic library); 3, stabbed PCR product of lane 3 (*Hin*dIII genomic library); 4, stabbed PCR product of lane 4 (*Bam*HI genomic library). F: Applying the touchdown and hot start PCR approaches to the SGSP-PCR had increased the specifity of the SGSP-PCR and excluded the need for a second round PCR. Lanes: 1, SGSP-PCR for MG1655-*asnV* site *Bam*HI genomic library (positive control- 3kb); 2-5, SGSP-PCR for E105-*asnV* site, for the following genomic libraries *Sal*I, *Hin*dIII, *Bam*HI respectively; 6-9, SGSP-PCR for E106-*asnV* site, for the following genomic libraries *Eco*RI, *Sal*I, *Hin*dIII, *Bam*HI respectively; 10-12, SGSP-PCR for E108-*asnV* site, for the following genomic libraries *Eco*RI, *Sal*I, *Hin*dIII, *Bam*HI respectively; 13, negative control. Lane M: GeneRuler^{TM} DNA ladder mix (Fermentas).

3.9. Alternative approaches to the SGSP-PCR

Two more methods were applied to interrogate the genomic islands associated with tRNA genes. These were introduced to solve problematic sites where either no amplicons were amplified with the SGSP-PCR or only short PCR products obtained, meaning that only the conserved flanks were amplified with no sequence of genomic DNA. The two new approaches were termed arbitrary primed PCR and integrase PCR.

3.9.1. Arbitrary primed PCR (AR-PCR)

In this approach, a single primer either U or D of the tRNA gene was used to amplify at random places in the genome of the clinical isolates. The PCR products were either run on gel and specific bands were excised or the PCR tube contents were used as a DNA template for ligation of the PCR products to a thymidine-tailed vector (T-vector); constructed from pBluescript II KS $^{+}$. The ligation mixture was then used as a template in a second round PCR using the same primer used in the former arbitrary PCR and one of the vector primers. The arbitrary primed PCR approach has been used to investigate the negative tRIP-PCR results of the tRNA site *aspV*, however, no PCR products were amplified from the ligation mixture (Figure 3.11). Therefore, despite the fact that the principle behind this strategy was simple and straightforward, however, no GEIs were identified using this strategy.

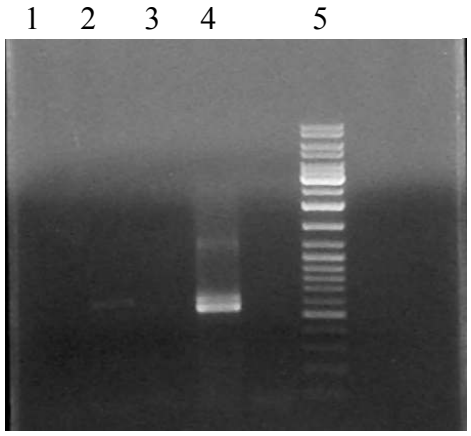


Figure 3.11. Amplifying the Arbitrary PCR products using the *aspV* upstream primer and the vector T7 primer. Samples used in the PCR, Lanes: 1, E102 ligated AR-PCR products; 2, E105 ligated AR-PCR products; 3, E107 ligated AR-PCR products; 4, ligated 542 bp control insert DNA (Promega) used as a positive control for the PCR and to prove that both the T-vector construction and the ligation steps were successful; 5, GeneRuler^{TM} DNA ladder mix (Fermentas).

3.9.2. Integrase PCR

As stated in the introduction, PAIs are often associated with cryptic or even functional mobility genes such as integrases or transposases. In this approach, the integrase genes associated with sequenced genomic islands in six different *E. coli* and *Shigella* strains were grouped into different families according to their sequence similarity and their integration tRNA site. Sequence similarity was determined by sequence alignment in ClustalW and the data obtained was used in the EBI server for Treeview clustering to identify integrase families. After that, a representative sequence for each family was used to design integrase specific primers using the Primaclade application (all of these former steps were done by Dr. Hong-Yu Ou, Shanghai Jiaotong University, P. R. China). tRNA interrogation for PAIs was carried out by amplifying part of the integrase gene using either the tRNA upstream or downstream primer and the integrase primer. Each tRNA site interrogated for PAIs harbouring integrase genes was investigated using four different possible orientations for the integrase genes (Figures 3.12 and 3.13). Presence of genomic island was confirmed after sequencing a positive integrase PCR amplicon in which either the conserved flank or the integrase sequence was identified. 13 different strain-tRNA sites were interrogated with the integrase PCR strategy (Table 3.4, Figures 3.13, 3.14 and 3.15) and five GEIs were identified by this method.

Table 3.4. Results of the integrase PCR.

Strain-tRNA site	Size of integrase PCR amplicon	Integrase family
E102-leuX	1.6 kb	P4
E103-leuX	1.6 kb	P4
E104-leuX	1.6 kb	P4
E111-leuX	1.6 kb	P4
E104-serW	1.5 kb	P4
E105-serW	–	–
E106-serW	–	–
E108-serW	–	–
E111-serW	–	–
E105-ssrA	–	–
E106-ssrA	–	–
E108-ssrA	–	–
E111-ssrA	–	–

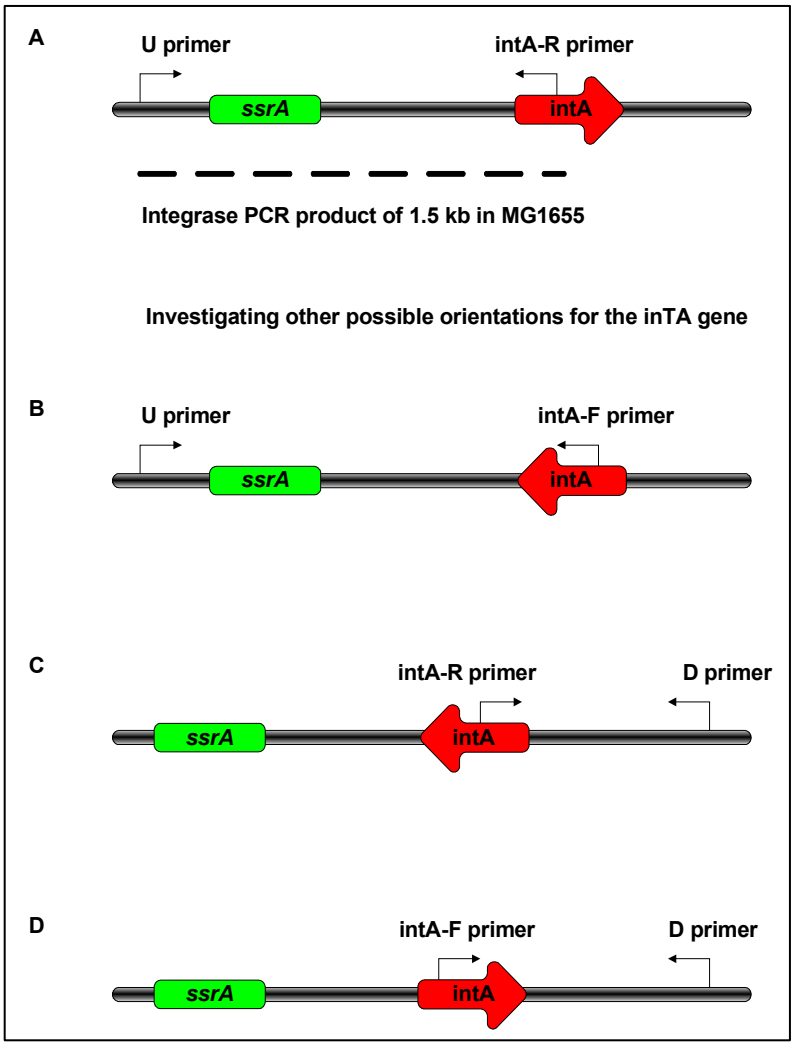


Figure 3.12. Integrase strategy for tRNA interrogation for PAIs harbouring integrase genes. Four possible orientations were investigated depending on the location of the integrase and its orientation in the genome.

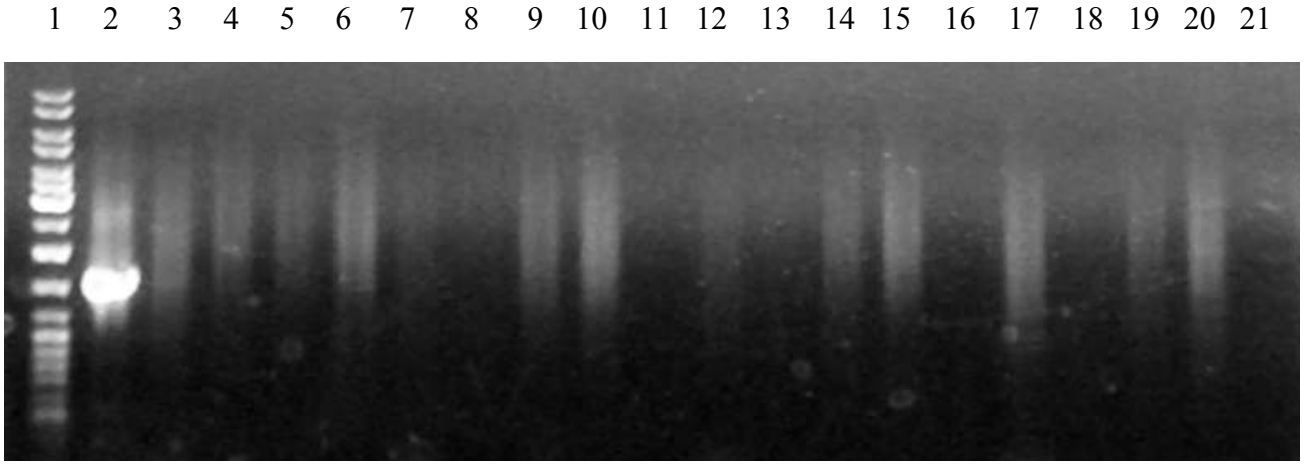


Figure 3.13. Integrase PCR for the tRNA site *ssrA*. Lanes: 2-6, amplification using *ssrA*-tRNA site upstream primer and the integrase specific primer intAR; 2. *E. coli* MG1655 used as a positive control (1.5 kb PCR product); 3-6, integrase PCR results for E105, E106, E108 and E111 respectively; 7-11, amplification using *ssrA*-tRNA site downstream primer and the integrase specific primer intAR for: E105, E106, E108, E111 and negative control respectively; 12-16, amplification using *ssrA*-tRNA site downstream primer and the integrase specific primer intAF for: E105, E106, E108, E111 and negative control respectively; 17-21, amplification using *ssrA*-tRNA site upstream primer and the integrase specific primer intAF for: E105, E106, E108, E111 and negative control respectively; 1, GeneRuler^{TM} DNA ladder mix (Fermentas).

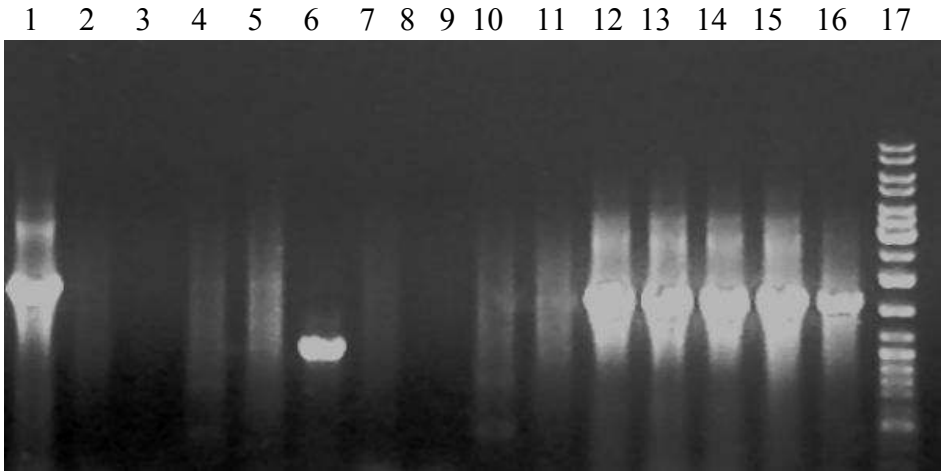


Figure 3.14. Integrase PCR for the tRNA site *leuX*. Lanes: 1-5, amplification using *leuX*-tRNA site upstream primer and the integrase specific primer *leuX*-IR; 1, *S. flexneri* 2a sf301 used as a positive control (1.9 kb PCR product); 2-5, integrase PCR results for E102, E103, E104 and E111 respectivily; 6-11, amplification using *leuX*-tRNA site upstream primer and the integrase specific primer *leuX*-IIR; 6, *E. coli* O157:H7 EDL933 used as a positive control (1.0 kb PCR product); 7-11, integrase PCR results for E102, E103, E104, E111 and negative control respectively; 12-16, amplification using *leuX*-tRNA site upstream primer and the integrase specific primer *leuX*-IIIR; 12, *E. coli* MG1655 used as a positive control (1.6 kb PCR product); 13-16, integrase PCR results for E102, E103, E104 and E111 respectivily; 17, GeneRuler^{TM} DNA ladder mix (Fermentas).

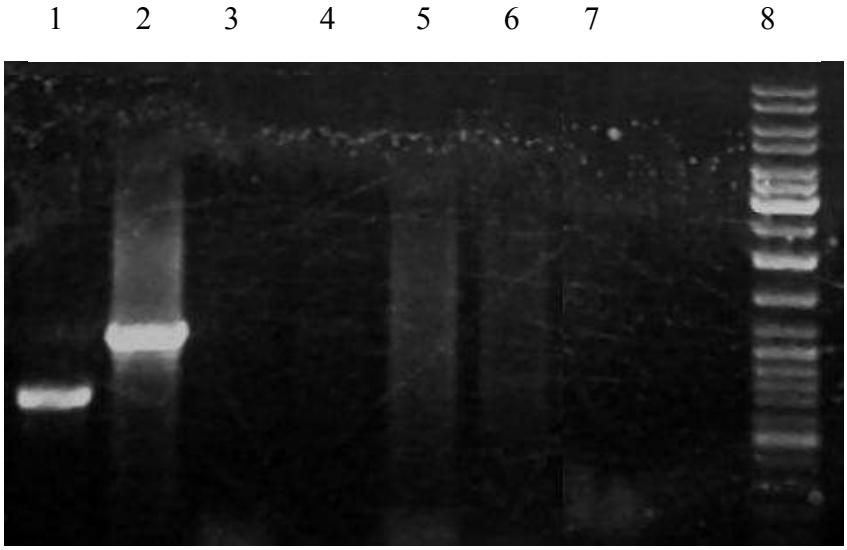


Figure 3.15. Integrase PCR for the tRNA site *serW*. Lanes: 1-7, amplification using *serW*-tRNA site downstream primer and the integrase specific primer P4-IR; 1, O157:H7 EDL933 was used as a positive control (1.0 kb PCR product); 2-7, integrase PCR results for E104, E105, E106, E108, E111 and negative control respectively; 8, GeneRuler^{TM} DNA ladder mix (Fermentas).

3.10. Summary of the investigated GEIs by SGSP-PCR and integrase PCR

Following the screening of the 160 distinct strain-tRNA sites, eighty GEIs were identified and confirmed using the sequential tRIP-PCR and SGSP-PCR strategy. Five additional GEIs were defined by using an integrase gene-based PCR approach. The level of GEI-occupancy varied from 6 to 13 GEIs per strain, with 1 to 10 loci corresponding to each of the 16 tRNA genes shown to be occupied. When the nature of the island present was considered, no two strains possessed an identical repertoire of GEIs. The identified GEIs consisted of: 6 GEIs harbouring novel sequences (Table 3.5), 12 GEIs previously defined by the Islander database (www.indiana.edu/~islander) (Mantri and Williams., 2004), and final set of 62 GEIs done previously using the tRNAcc method, with 9 of these 62 GEIs having an integrase sequence. Collectively, the identified GEIs could be classified into 36 GEI families based on sequence similarities at GEI terminal boundaries. Despite being based on a negative PCR result, the remarkable positive predictive value of the tRIP screen was highlighted by the fact that out of the 86 tRIP-PCR negative strain-tRNA sites for which SGSP-PCR or int-PCR data were available, only one site was shown to be empty. A summary of the identified GEIs is described in Figure 3.16 and Table 3.6 and the blastn results for all sequenced SGSP-PCR amplicons are summarized in Table 8.3 (Appendix). The total number of the SGSP-PCR amplicons produced for all tested tRNA sites are summarized in Table 8.4 (Appendix)

An interesting finding of this study is that 46 of the identified GEIs, falling into 11 different families, resembled uropathogenic *E. coli* CFT073-like entities. However, the ten BSI isolates were derived from patients with no laboratory evidence of immediately preceding or concurrent urinary tract infection. These data are concordant with the current view based on phylogenetic classification and virulence factor repertoires that uropathogenic *E. coli* comprise a subset of extraintestinal pathogenic *E. coli* (ExPEC). ExPEC strains are associated not only with urinary tract infections but also with meningitis and BSIs.

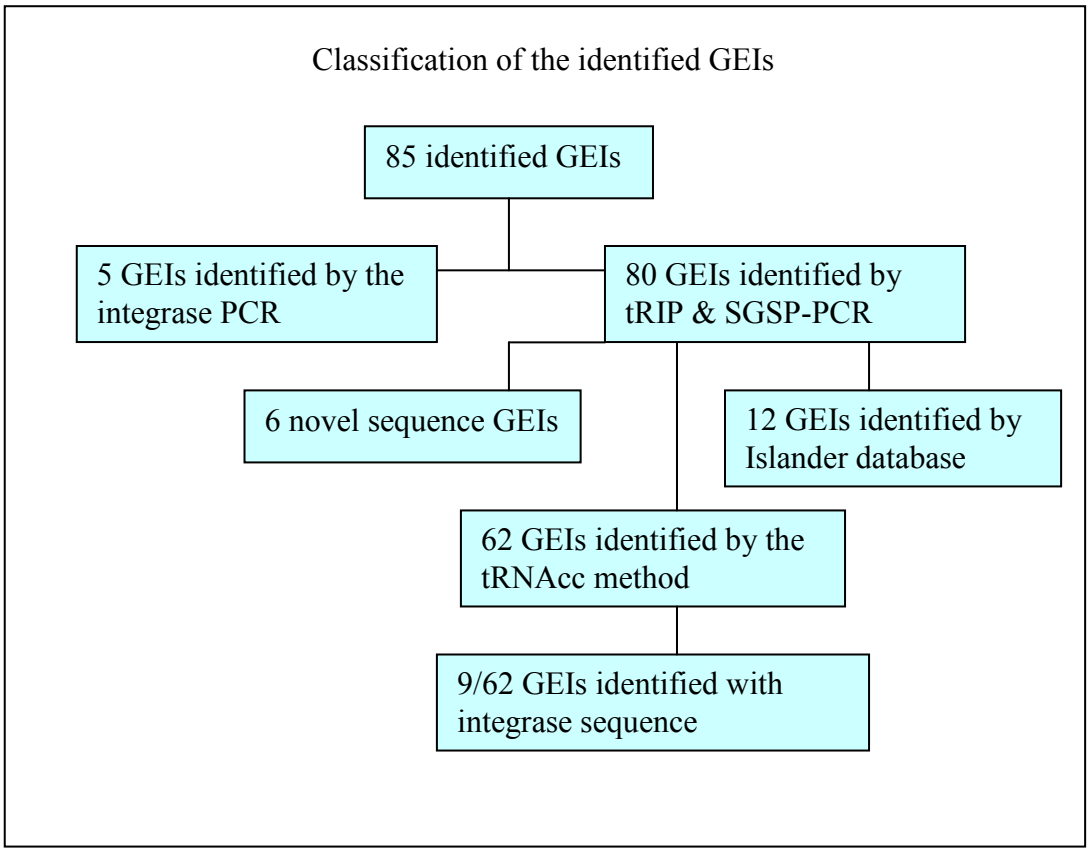


Figure 3.16. Classification of the identified GEIs.

Table 3.5. Summary of available sequence data for GEIs harbouring novel sequences

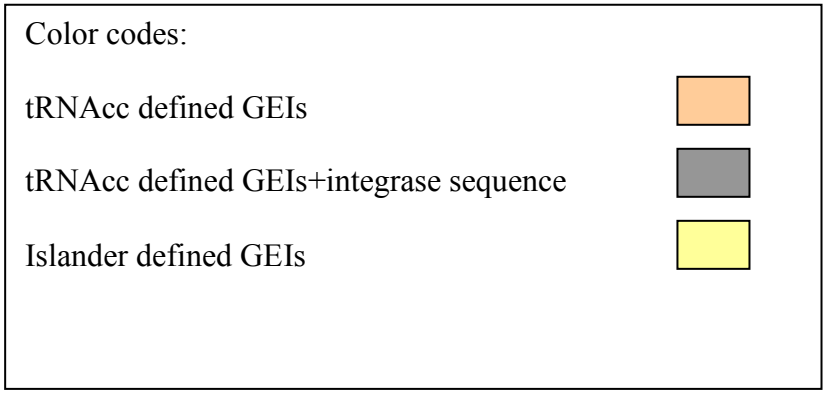
Strain-tRNA gene	SGSP-PCR amplicon[a]	Total sequence length (bp)	Length of GEI-specific sequence (bp)	Blastn hit(s) of GEI-specific sequence (length of match [bp])	Blastn results	NCBI GeneBank accession number
E104-*aspV*	#D-#T3 (0.6 kb) *Hin*dIII (#D)	480	365	*Salmonella enterica* IS*1414* (326 bp)	e-value=7e-81 bit score =309 ID= 83%	AY502962.1
				No significant similarity (37 bp)		
E105-*leuX*	#U-#T3 (1.4 kb) *Pst*I (#KS)	879	200	No significant similarity (200 bp)		
E105-*serT*	#U-#T3 (1.2 kb) *Hin*dIII (#U)	911	789	*E.coli* O157:H7 EDL933 (80 bp)	e-value=5e-90 bit score=340 ID=97%	AE005174.2
				E.coli O157:H7 EDL933 putative ferredoxin (124 bp)	e-value=3e-32 bit score=148 ID=88%	AE005174.2
				Shigella boydii Sb227 (154 bp)	e-value=2e-39 bit score=172 ID=87%	CP000036.1
				Bacteriophage P-EibD (259 bp)	e-value=3e-72 bit score=281 ID=86%	AF151675.1
				No significant similarity (172 bp)		-
E106-*serW*[b]	#U-#T3 (1.9 kb) *Eco*RV (#U)	405	64	No significant similarity (64 bp)		-
E108-*serW*[b]	#D-#T3 (2.3 kb) *Pst*I (#KS)	869	869	*Erwinia carotovora,* hypothetical protein gene (91 bp)	e-value=3e-10 bit score=71.9 ID=84%	BX950851.1
				No significant similarity (778 bp)		
E111-*serW*[b]	#D-#T3 (2.3 kb) *Pst*I (#KS)	820	820	*E.coli strain* BEN2908 PAI EPI-I (265 bp)	e-value=9e-143 bit score=511 ID=99%	AY857617.1
				Psychrobacter arcticus, hypothetical protein gene (41 bp)	e-value=1e-08 bit score=65.9 ID=95%	CP000082.1
				No significant similarity (514 bp)		

[a]This column represents the primers used to amplify the PCR amplicon, the size of amplicon, the restriction enzyme used to construct the genomic library, and the primer used to sequence the amplicon, respectively.
[b]The blast results for these GEIs boundaries were reported before the availability of the full length sequence for UTI89 and APEC O1 (Chen *et al*., 2006; Johnson *et al*., 2007).

Table 3.6. The identified GEIs by the SGSP-PCR and integrase PCR strategies

Strain	tRNA site														
	serT	*selC*	*glyU*	*serU*	*serW*	*asnV*	*thrW*	*metV*	*serX*	*argW*	*leuX*	*asnT*	*aspV*	*pheV*	*ssrA*
E102	+	+	EDL933	+	+	+	+	+	S.b 227	CFT073	MG1655[c]	CFT073	–[e]	CFT073	EAECO42
E103	+	+	0.9	+	+	+	CFT073	+	S.b 227	CFT073	MG1655[c]	CFT073	–[e]	CFT073	EAECO42
E104	+	+	EDL933	+	MG1655[c]	+	+	CFT073	CFT073	EDL933	MG1655[c]	MG1655	novel[a]	CFT073	EAECO42
E105	novel[a]	EAECO42	0.9	CFT073[b]	Empty[d]	CFT073	CFT073	CFT073	CFT073	CFT073	novel[a]	CFT073	EAECO42	MG1655	CFT073
E106	+	+	0.9	+	novel[a]	CFT073	1.3	CFT073	–[e]	EDL933	APECO1[c]	CFT073	–[e]	MG1655	CFT073
E107	+	+	EDL933	+	+	+	MG1655	+	+	EDL933	S.f 301	CFT073	E24377A	Sf301	CFT073
E108	+	+	0.9	CFT073	novel[a]	CFT073	+	CFT073	+	+	+	+	–[e]	MG1655	CFT073
E109	+	EAECO42	0.9	CFT073	+	CFT073	+	CFT073	+	+	+	+	–[e]	CFT073	MG1655
E110	+	CFT073	0.9	+	+	CFT073[b]	CFT073	CFT073	CFT073	CFT073	CFT073	CFT073	–[e]	CFT073	CFT073
E111	+	+	0.9	EAECO42	novel[a]	+	MG1655	CFT073	MG1655	+	CFT073	EDL933	EAECO42	CFT073	CFT073

[a]A GEI with a novel DNA sequence.
[b]These GEIs were identified by comparing their strain-tRNA sites restriction fragment length polymorphism to other strain-tRNA sites with similar restriction fragment polymorphism (Table 8.4 Appendix).
[c]A GEI defined by integrase PCR.
[d]An empty tRNA site identified after sequence analysis of the SGSP-PCR amplicon.
[e]Uncharacterized tRNA site. All interrogation strategies applied in this study had failed to characterize these sites.



3.11. Analysis of the E102_*asnT* (a tRNAcc GEI)

A 1.5 kb PCR product was produced when E102_*asnT* site was interrogated by the SGSP-PCR. The 1.5 kb band was sent for sequencing using the vector primer (#ks) and a 369 bp clipped sequence was obtained after the sequencing. The blastn results of the 369 bp clipped sequence indicated that it hits into a 36.5 and 37.2 kb islands in the *asnT* site, identified in *E. coli* UTI89 and CFT073 respectively by the tRNAcc method (Ou *et al*., 2006) (Figure 3.17).

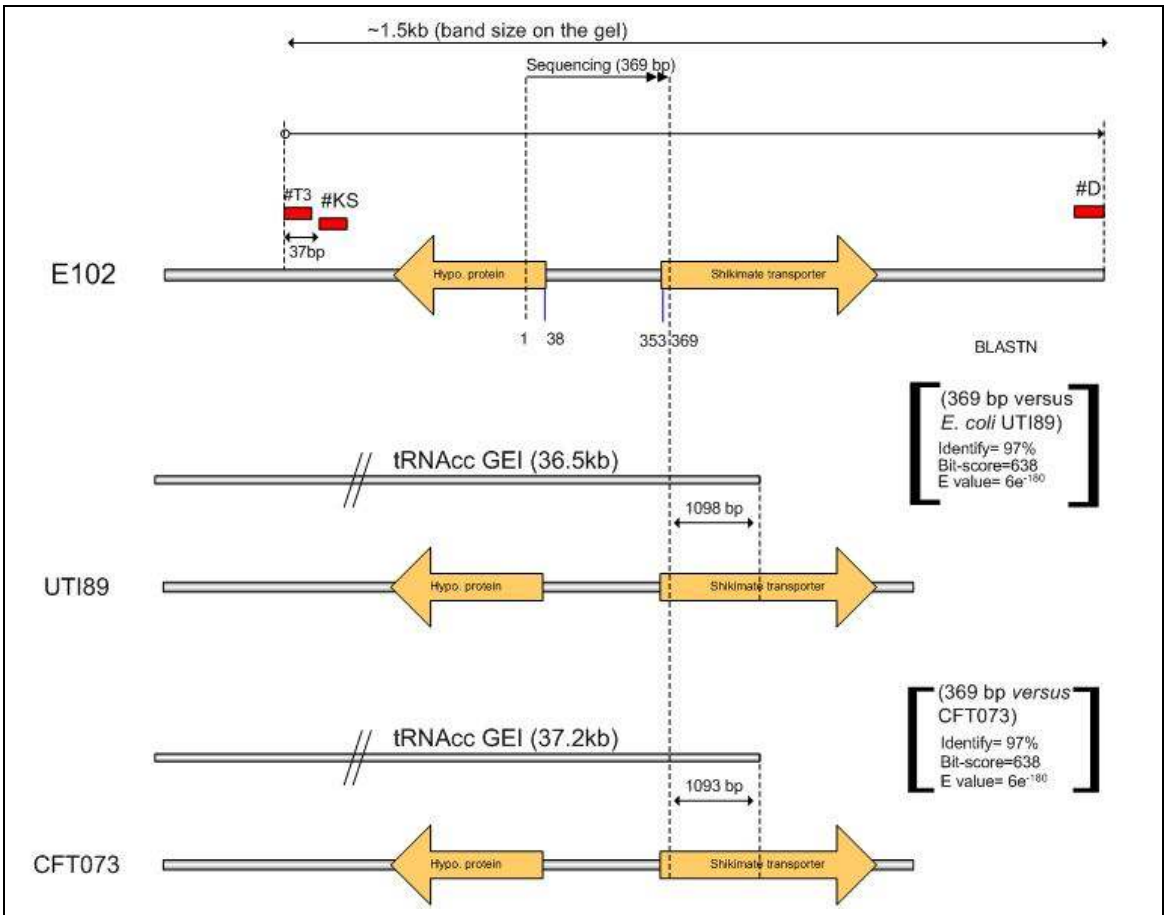


Figure 3.17. The SGSP-PCR GEI of E102_*asnT*.

3.12. Analysis of the E107_*argW* (Islander GEI)

A 1.3 kb PCR product was produced when E107_*argW* site was interrogated by the SGSP-PCR. The 1.3 kb band was sent for sequencing using the vector primer (#T3) and a 750 bp clipped sequence was obtained after the sequencing. The blastn results of the 750 bp indicated that 674 bp hits into 8.6 and 14.1 kb islands in the *argW* site, identified in EDL933 by the islander database and the tRNAcc method respectively. The sequence was found to be located within the common distal termini identified by both methods the islander and the tRNAcc. Additionally the other 77 bp fragment (674-750) matches the 5`-end of a putative transport protein and the 3`-end of a putative prophage integrase with 96% identity. The 77 bp is not represented in Figure 3.18 because it hits in different part of the genome of EDL933.

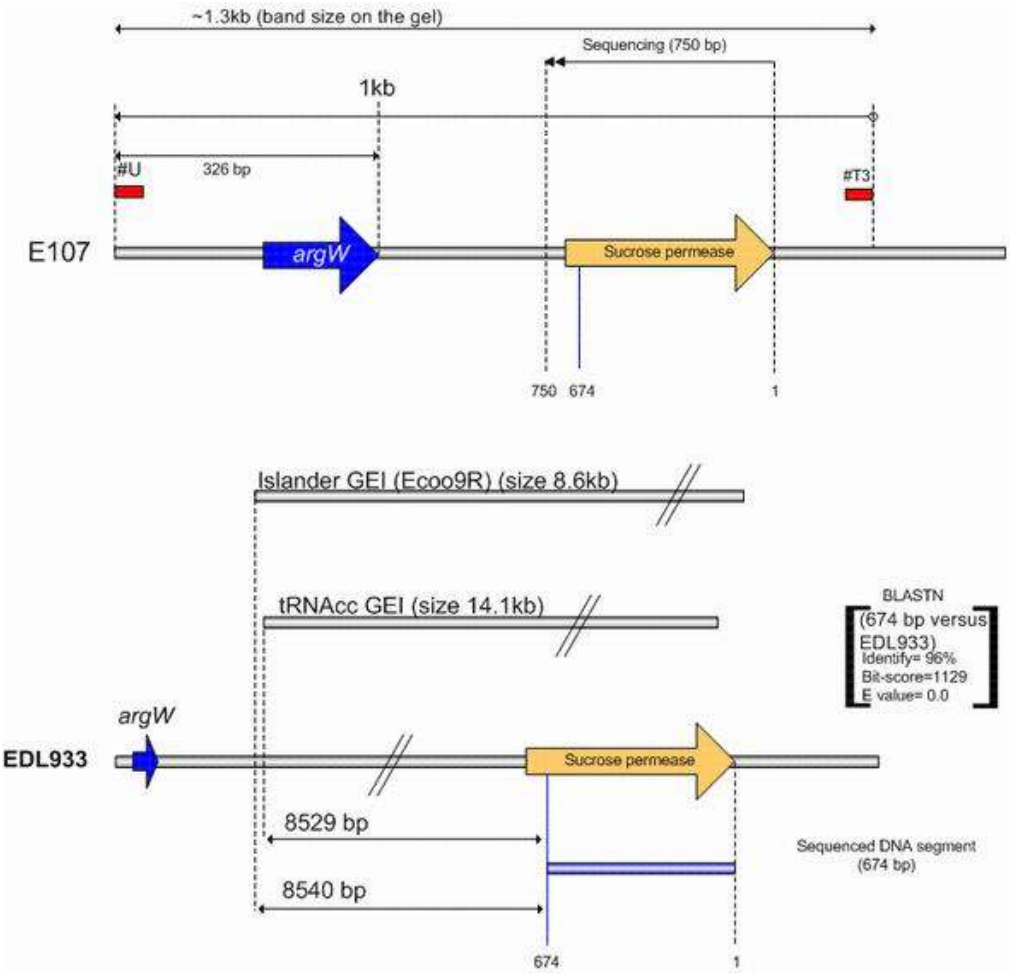


Figure 3.18. The SGSP-PCR GEI of E107_*argW*.

3.13. Analysis of the E105_*thrW* (a GEI identified by integrase sequence)

A 2.5 kb PCR product was produced when E105_*thrW* site was interrogated by the SGSP-PCR. The 2.5 kb band was sent for sequencing using the upstream primer (#U) and a 750 bp clipped sequence was generated after the sequencing. The blastn results of the 750 bp indicated that it hits into an integrase sequence down stream of *thrW* site, identified in CFT073 (Figure 3.19).

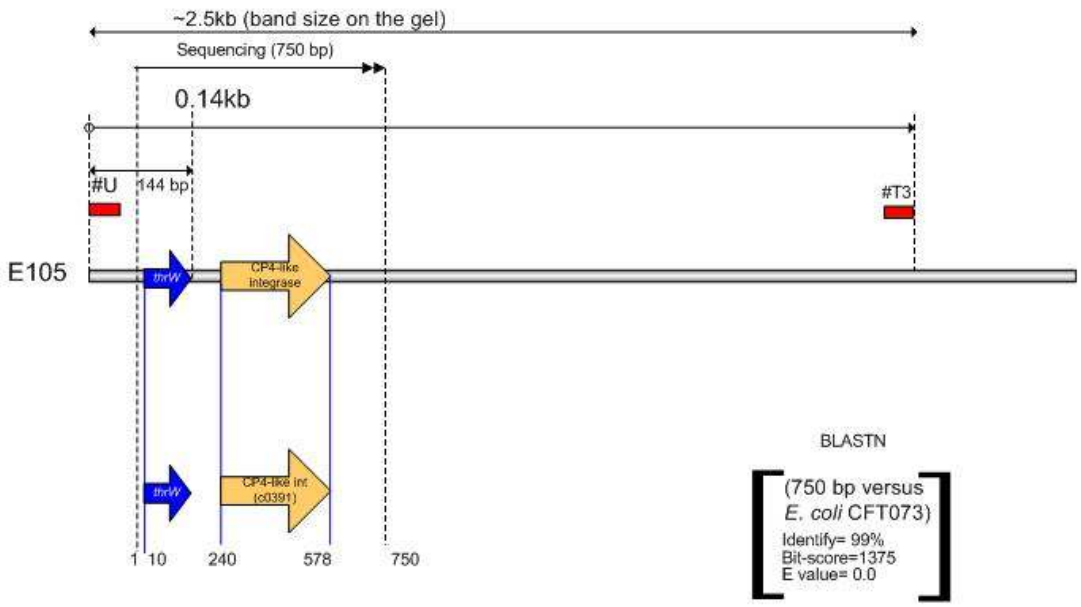


Figure 3.19. The SGSP-PCR GEI of E105_*thrW*.

3.14. Analysis of the E104_*aspV* (GEI with novel sequence and mosaic structure)

A 0.6 kb PCR product was produced when E104_*aspV* site was interrogated by the SGSP-PCR. The 0.6 kb band was sent for sequencing using the down stream primer (#D) and a 516 bp clipped sequence was generated of which, 36 bp of it was part of the vector sequence. The blastn results of the 480 bp (vector clipped sequence) indicated that a 326 bp of the sequence hits into IS element; IS1414 (*tnpA*) identified in *S. enterica* serovar Abortusovis, while a 115 bp of the sequence match to a putative amidase-type enzyme identified in O157:H7 str. Sakai. On the other hand, 37 bp of the sequence has no match in the NCBI database (Figure 3.20).

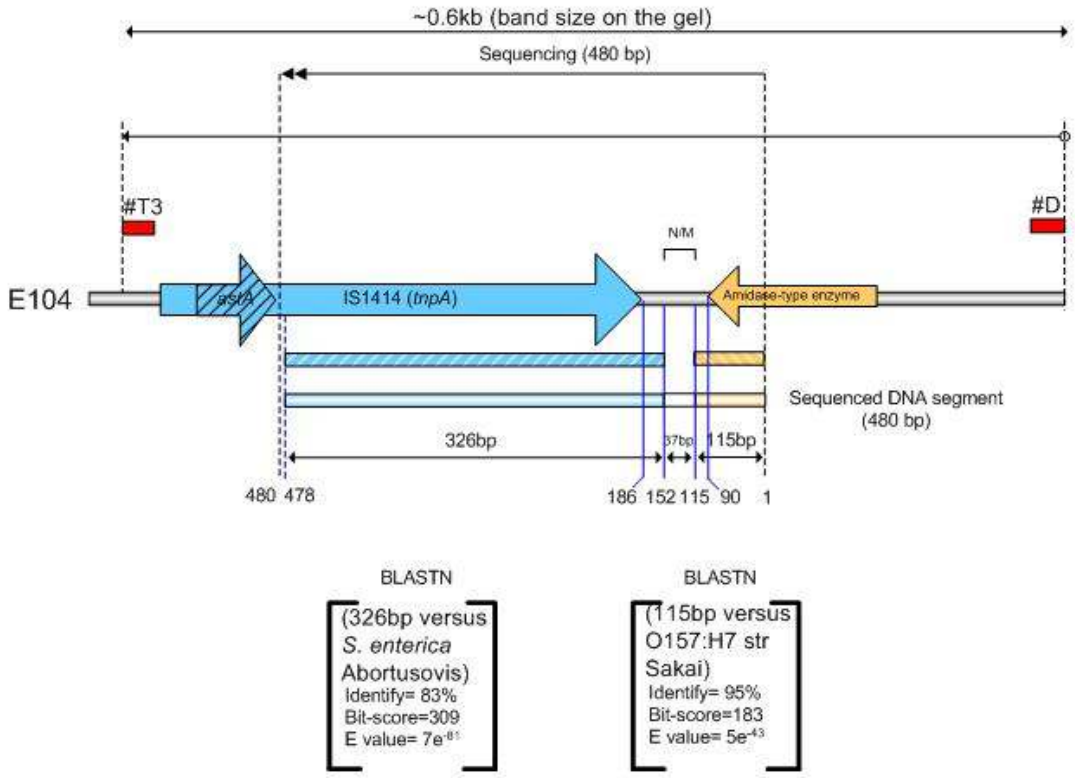


Figure 3.20. The SGSP-PCR GEI of E104_*aspV*. N/M is the 37 bp which has no match in the NCBI database.

3.15. Analysis of the E105_*leuX* (GEI with novel 200 bp that has no match in the database)

A 1.4 kb PCR product was produced when E105_*leuX* site was interrogated by the SGSP-PCR. The 1.4 kb band was sent for sequencing using the vector primer (#ks) and a 903 bp clipped sequence was generated. The blastn results of the 903 bp clipped sequence indicated that a 703 bp of the sequence hits into the conserved flanks of the *leuX* site in UTI89 strain, while a 200 bp of the sequence has no match in the NCBI database (Figure 3.21).

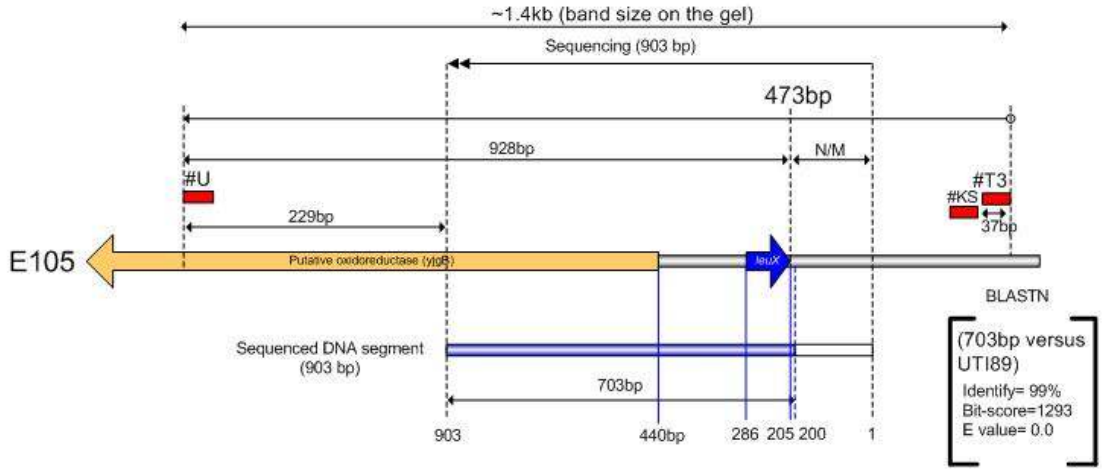


Figure 3.21. The SGSP-PCR GEI of E105_*leuX*. N/M is the 200 bp which has no match in the NCBI database.

3.16. Analysis of the E105_*serT* (GEI with novel 172bp that has no match in the database)

A 1.2 kb PCR product was produced when E105_*serT* site was interrogated by the SGSP-PCR. The 1.2 kb band was sent for sequencing using the upstream primer (#U) and a 911 bp clipped sequence was generated. The blastn results of the 911 bp clipped sequence indicated that a 122 bp of the sequence hits into the conserved flanks of the *serT* site in EDL933 strain, while a 167 bp of the sequence has no match in the NCBI database (Figure 3.22).

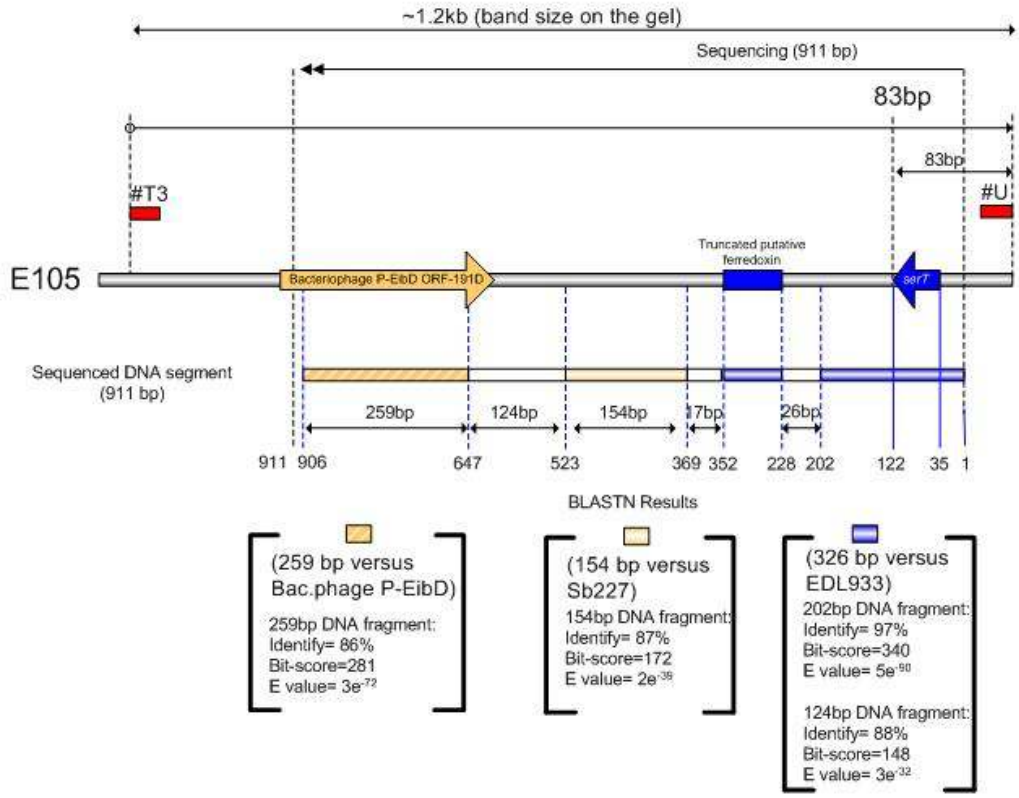


Figure 3.22. The SGSP-PCR GEI of E105_*serT*.

3.17. Analysis of the E106_*serW* (previously identified as a GEI with novel 64bp that has no significant similarity in the database)

A 1.9 kb PCR product was produced when E106_*serW* site was interrogated by the SGSP-PCR. The 1.9 kb band was sent for sequencing using the upstream primer (#U) and a 405 bp clipped sequence was generated. The blastn results of the 405 bp clipped sequence indicated that it hits into the flanks of the *serW* site in APECO1 and UTI89 strains, while previously 64 bp of the sequence has no significant similarity in the NCBI database (Figure 3.23).

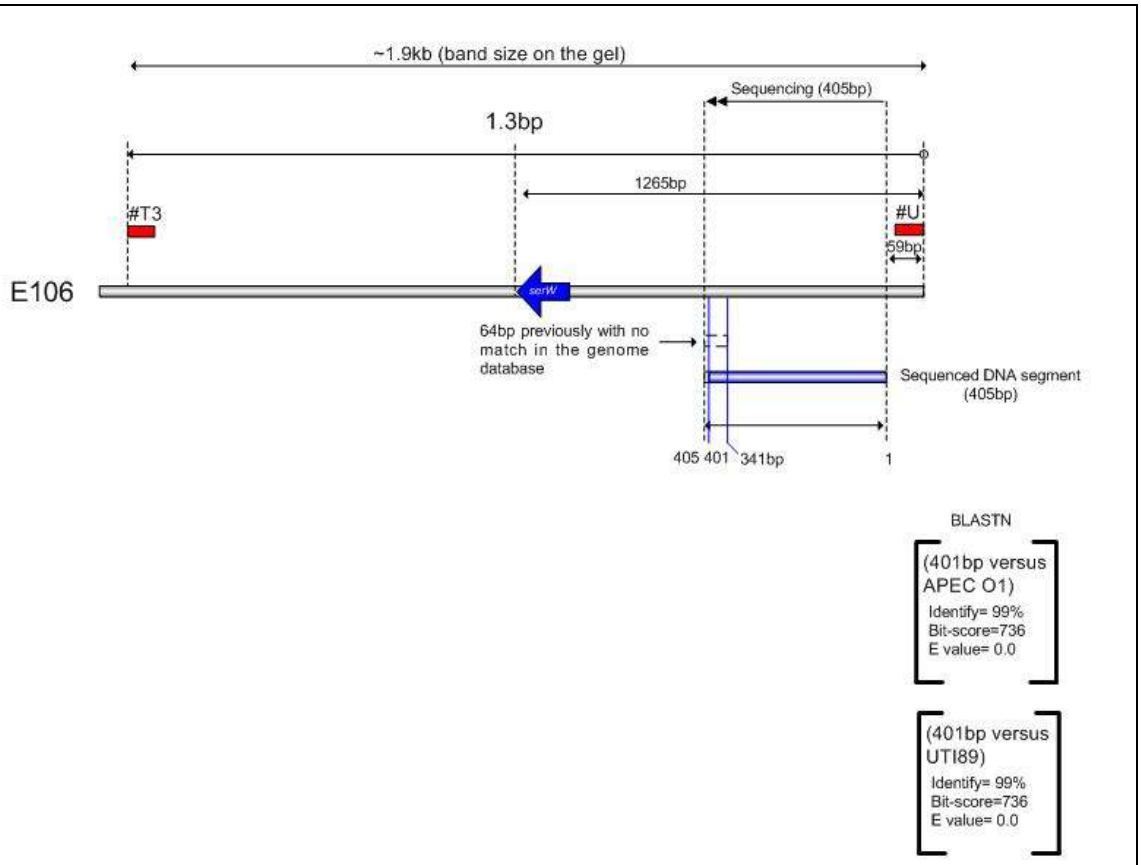


Figure 3.23. The SGSP-PCR GEI of E106_*serW*.

3.18. Analysis of the E108_*serW* (previously identified as a GEI with novel 778bp that has no significant similarity in the database)

A 2.3 kb PCR product was produced when E108_*serW* site was interrogated by the SGSP-PCR. The 2.3 kb band was sent for sequencing using the vector primer (#ks) and an 869 bp clipped sequence was generated. The blastn results of the 869 bp clipped sequence indicated that it hits into the flanks of the *serW* site in APECO1 and UTI89 strains, while previously 778 bp of the sequence has no significant similarity in the NCBI database (Figure 3.24).

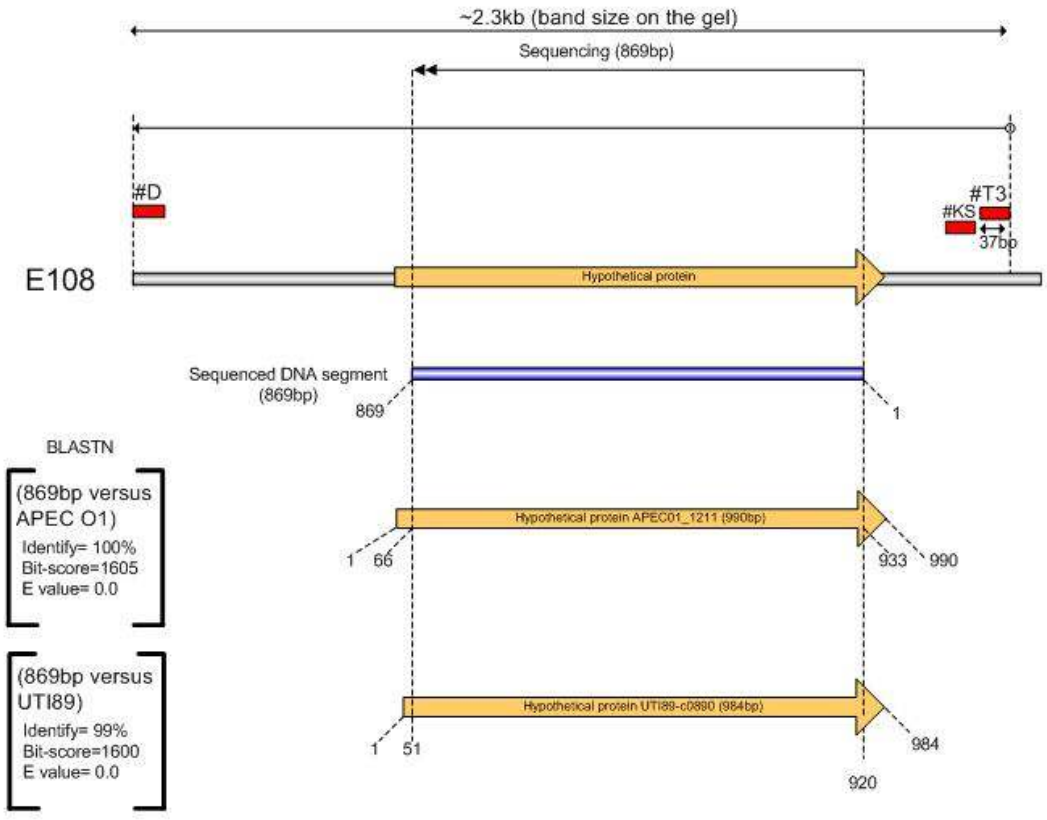


Figure 3.24. The SGSP-PCR GEI of E108_*serW*.

3.19. Analysis of the E111_*serW* (previously identified as a GEI with novel 514bp that has no significant similarity in the database)

A 2.3 kb PCR product was produced when E111_*serW* site was interrogated by the SGSP-PCR. The 2.3 kb band was sent for sequencing using the vector primer (#ks) and an 820 bp clipped sequence was generated. The blastn results of the 820 bp clipped sequence indicated that it hits into the flanks of the *serW* site in APECO1 and UTI89 strains, while previously the 514 bp of the sequence has no match in the NCBI database (Figure 3.25).

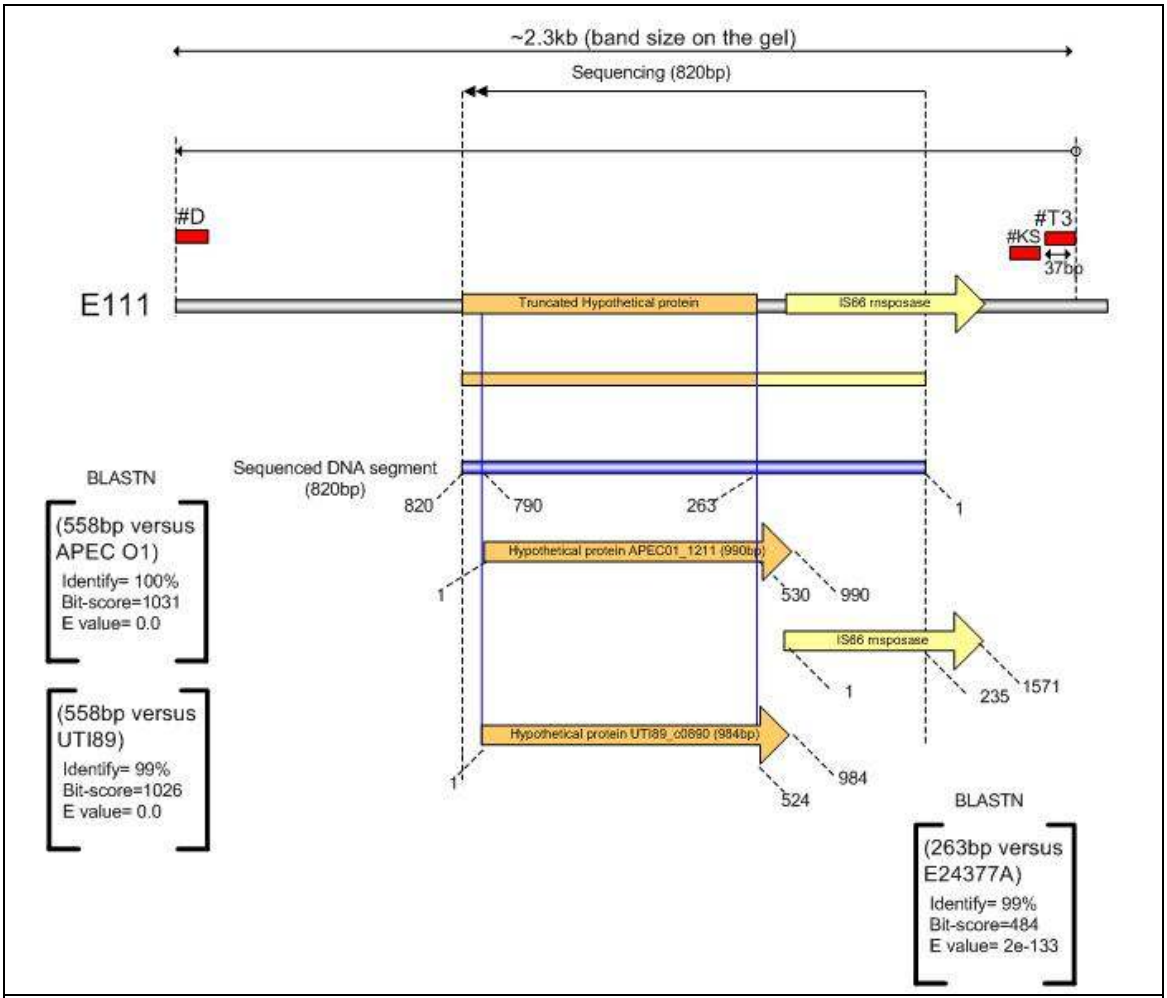


Figure 3.25. The SGSP-PCR GEI of E111_*serW*.

3.20. Discussion

In this study a large-scale interrogation of tRNA sites, known hot spots for the integration of foreign DNA sequences such as genomic islands, was designed to investigate the tRNA occupancy for GEIs and to discover new GEIs with mosaic DNA structures or novel DNA sequences that have no match in the genome database.

A negative tRIP-PCR result was considered as an indication of the presence of a GEI. A negative PCR result due to other reasons such as the deletion of the complementary sequence for one of the primers or to PCR failures were also considered. Therefore, positive control reactions (using genomic DNA of MG1655) were included in the tRIP-PCR to check for PCR failure. Moreover, the obtained results for the tested strains were confirmed by repeating the tRIP-PCR for at least tree times. A deletion of the complementary sequence for one of the primers used in the PCR was considered unlikely as the primers used in the tRIP-PCR were designed using four to five strains. Furthermore, previous studies by our group (Ou *et al*., 2006), had confirmed that the upstream and downstream flanks for 20 tRNA sites were generally found to be conserved with only 2 upstream (*argU* in CFT073 and Sf301) and 5 downstream deletions (*thrW* in Sf301, *serX* in CFT073, *pheV* in EDL933, *selC* in Sf301, and *ssrA* in CFT073). However, a negative tRIP-PCR was only regarded as an indication of a GEI after part of the GEI sequence was revealed by the PCR genome walking method SGSP-PCR.

Different *in silico* approaches (e.g., the tRNAcc method and Islander database) were previously applied to analyse the contents of prokaryotic genomes and search for genomic islands. However, identifying genomic islands using such methods is obviously limited by the number of sequenced genomes available in the database. On the other hand, sequencing genomes of new pathogenic microbes to reveal strain specific sequences associated with virulence traits would be a waste of time and effort as the members of a species share a common core DNA sequence that represents more than 60% of the genome. Therefore, the isolation or identification of such anomalous (atypical and foreign) DNA sequences *in vitro* would represent a shortcut in the genome

sequencing projects. Sequencing of these anomalous DNA fragments could broaden our understanding about the association between the genome content of strains and the potential to give rise to highly virulent, highly transmissible and/or environmentally persistent strains that could pose a threat of severe disease and/or exhibit outbreak potential. One important issue that could come out of such understanding is in the treatment of bacterial infections using anti-virulence traits regardless of the microbial species that harbour these pathogenic markers.

Few *in vitro* methods have been applied to isolate or interrogate anomalous DNA from un-sequenced genomes. These include subtractive hybridization strategies, in which two different strains from the same species are compared (Lisitsyn *et al*., 1993; Bart *et al*., 2000; Malloff *et al*., 2001). However, such studies would identify anomalous DNA by searching for chromosomal differences between the two strains and would probably miss small variations between similar GEIs present in both strains. Van Passel *et al* (2004) had tried to identify anomalous DNA sequences by clustering of restriction enzyme recognition sites. The method proposed that frequencies of restriction sites obtained from genomes of different microbial species are different. However, a horizontal gene transfer event that introduced the same anomalous DNA into different microbial genomes would produce a cluster of certain restriction sites that could be used as a marker for the presence of such foreign DNA in different microbial strains. However, it is known that horizontal gene transfer (HGT) events are involved in a mutational process that affects the entire genome and that foreign DNA is thought to be adjusted to the host's DNA sequence over time in a process called amelioration. Consequently, only evolutionary recent HGT events with DNA sequence that differ from the host genome sequence can be adequately identified (Van Passel *et al*., 2004). Moreover, restriction sites methylated by modification systems would render the genome resistant to the actions of restriction enzymes (Van Passel *et al*., 2004). Screening the whole microbial genome for GEIs would require a comparative approach as mentioned above, however only a limited number of strains could be interrogated for their GEI content by these methods. To the contrast, the screening method (tRIP-PCR) we applied would enable the interrogation of large number of strains for their GEIs content as it is designed to investigate specific and known hot spots of the genome

(tRNA genes) for integration of GEIs. Germon *et al* (2007) had used the same strategy to interrogate 36 tRNA sites in 54 different *E. coli* strains. However, in the study of Germon *et al* the negative screening PCR results were considered as an indication of the presence of foreign chromosomal DNA, without further investigation for the boundaries of these foreign chromosomal DNAs to prove that such HGT events had occurred.

The results of our study showed that the following tRNA sites (*argW*, *metV*, *serX*, *asnT*, *leuX*, *aspV*, *pheV* and the tmRNA site *ssrA*) had negative tRIP-PCR results for at least 7/10 strains. These 8 genes are among 16 tRNA sites previously defined as being hot spots for GEIs integration (Ou *et al*., 2006) in different bacterial strains. The occupancy of these 8 tRNA sites could indicate an association between the genes encoded by GEIs downstream of these tRNA sites and the pathogenicity of BSI-associated *E. coli* strains.

The clustering of tRIP-PCR results into different groups had showed a strong correlation between the tRIP-PCR profiles obtained and the phylogenetic group B2. However, no association could be drawn for the ECOR group D when compared to the tRIP-PCR profiles. Similar observations were made by Germon *et al* (2007), associating the clustering of extra-intestinal pathogenic *E. coli* (ExPEC) to the clonal group B2. This association was further investigated by Picard *et al* (1999) and Escobar-Paramo *et al* (2004), in which they proposed that specific genetic backgrounds (ECOR groups) play an important role in the integration, retention, and expression of virulence factors. Their investigations showed that the B2 ECOR group might work as the right background for acquisition and expression of ExPEC virulence factors.

Two small GEIs (islets) were obtained by the tRIP-PCR. Both were *selC* GEIs, the first E104_*selC* was 1.5 kb. Sequencing the PCR product for E104_*selC* revealed that the sequence matches to a small GEI (2.9 kb) defined by the tRNAcc method in strain E24377A. The second GEI E106_*selC* was 3.2 kb and sequencing the PCR product using the upstream primer had revealed that the sequence of this GEI is similar to the sequence of a small GEI defined by the tRNAcc method and found downstream of the *selC* site in both strains APEC O1 and UTI89. These two GEIs could be the remnants left behind following partial excision of a larger GEI. The *selC* site has been found to

be an integration site for known GEIs like the LEE island found in EPEC E2348/69 and EHEC EDL933 (Jores *et al*., 2004), PAI I_{536} (Middendorf *et al*., 2004) and the CFT073 *selC* GEI (Welch *et al*., 2002). Integration of different GEIs to the *selC* site indicates that this locus works like a hot spot for integration of different anomalous DNA molecules and no attribution to a specific GEI can be made as most of the above GEIs comes from different strains and encode for different virulence factors.

Investigating the negative tRIP-PCR results with the SGSP-PCR strategy has revealed some of the virulence genes associated with the identified GEIs. For example, the truncated putative ferredoxin gene identified in E105_*serT* (Figure 3.22) has been associated with the production of enzymes involved in the fermentation process by *Giardia lamblia* and *Entamoeba histolytica* (Nixon *et al*., 2002). These enzymes are required to survive the anaerobic conditions within the intestinal lumen. However, the ferrodoxin is also involved in the activation of metronidazole, a drug used in the treatment of infections caused by giardia, amoebae and gram-negative bacteria. This could explain why metronidazole-resistant amoebae show decreased expression of ferredoxin (Nixon *et al*., 2002). The presence of a truncated ferredoxin in strain E105 could also indicate that this strain is resistant to metronidazole. Other examples of virulence genes associated with the identified GEIs by the SGSP-PCR strategy included the putative oxidoreductase, found in E105_*leuX* (Figure 3.21). The oxidoreductase is thought to assess in the folding, stability and activity of many proteins secreted by the gram-negative bacteria (Miki *et al*., 2004). The mechanism involves introducing a disulfide bond into proteins exported from the cytoplasm to the periplasm. In pathogenic strains the folding and assembly of virulence proteins is required for the expression of these determinants (Miki *et al*., 2004). In E110_*thrW* and E104_*asnT* (Table 8.3. Appendix) the genes for a haemoglobin protease and glycosyltransferase were identified, respectively. Haemoglobin protease is considered an important iron acquisition system in pathogenic *E. coli* (Scott *et al*., 2002), while, glycosyl transferase has been found to be an important protein expressed by *E. coli* K-12 and is involved in the adhesion of bacteria to polystyrene plates as well as for lipopolysaccharide synthesis (Narimatsu *et al*., 2004).

As stated above an interesting finding of this study is that 46 of the identified GEIs, falling into 11 different families, resembled uropathogenic *E. coli* CFT073-like entities. CFT073 was isolated from the blood and urine of a woman with acute pyelonephritis (Kao *et al*., 1997). The source of urinary tract infections is presumed to be fecal after contamination of the periurethral area (Lloyd *et al*., 2007). The bacteria ascend through the urethra to the bladder. Colonization of the bladder results in cystitis. The spread of the infection up the ureters to the kidneys would result in pyelonephritis. The life-threatening complication of pyelonephritis could result in histological damage of the uroepithelium, and transfer of bacteria into the blood stream producing systemic infection (Guyer *et al*., 1998; Lloyd *et al*., 2007).

Another interesting finding about the CFT073-like entities is that they are clustered and associated with mainly 5 different tRNA sites: in *metV* 7/10 strains had a CFT073-like entity, similarly in *asnT*, *pheV*, and *ssrA* 6/10 strains had a CFT073-like entity and finally in *asnV* 5/10 strains had a CFT073-like entity. From previous studies (Welch *et al*., 2002 and Lloyd *et al*., 2007), 13 GEIs were identified in CFT073 and found to be associated with the tRNA sites (*aspV*, *thrW*, *serX*, *serU*, *asnT*, *asnV*, *argW*, *ssrA*, *metV*, *pheV*, *selC*, *pheU* and *leuX*). Comparing the results of Welch *et al* and Lloyd *et al* with our results, only 2 tRNA sites (*aspV* and *pheU*) previously identified as hot spots for the integration of CFT073-like entities are not occupied with CFT073-like entities in our data. Characterization of the following *aspV* sites is still not completed E102_*aspV*, E103_*aspV*, E106_*aspV*, E108_*aspV*, E109_*aspV* and E110_*aspV* and so no confirmation can be drawn about the occupancy of the *aspV* site with CFT073-like entities. However, the *pheU* site had a positive tRIP-PCR results for all the tested strains in this study, indicating that no GEIs are inserted at this locus.

The fact that the 10 *E. coli* strains investigated in this study and the CFT073 strain were all isolates of blood culture, might explain the high number of the CFT073-like entities identified in our study. Further characterization of these entities could reveal their importance for the survival of these pathogens in the blood of infected patients. On the other hand, different reasons could be obtained from previous studies to explain why the *pheU* gene was the only site not occupied with CFT073-like entities between 11

other tRNA sites identified as hot spots for CFT073-like entities in the tested strains. PAI II_{CFT073}, which is associated with the *pheU* gene has been characterized in previous studies (Rasko *et al*., 2001 and Parham *et al*., 2005), and was found to be genetically unstable as most of its open reading frames (ORFs) were insertion sequences and transposases (28%). It also contains the iron acquisition system FbpA-D (for "ferric binding protein") as part of its adaptation to the iron-limiting urinary tract environment (Parham *et al*., 2005). It is known that UPEC strains possess more than one iron siderophore system to adapt to the limited iron environment of the UTI, e.g., $HPI_{CFT073,}$ which is associated with the *asnT* tRNA site and encodes for the yersiniabactin iron acquisition system (Welch *et al*., 2002). The change of the bacterial environment from the urinary tract to the blood, which is iron rich environment and the instability of the PAI II_{CFT073} due to the many insertion sequences and the transposases could result in the subsequent deletion of a PAI II_{CFT073}-like entity downstream of *pheU* site in the tested BSI-associated *E. coli* strains. Moreover, the TosA (for "type one secretion") adhesion protein, which is encoded by PAI II_{CFT073} was proposed to inhibit the *E. coli* strains from ascending the urinary tract and/or invading the blood stream (Parham *et al*., 2005). This is because TosA has been found more frequently among UTI strains causing cystitis than in strains causing pyelonephritis and septicemia (Parham *et al*., 2005). These finding about PAI II_{CFT073}, could explain the excision of a GEI "with a similar characteristics to PAI II_{CFT073}" from the *pheU* site in the *E. coli* BSI-associated strains investigated in this study.

The findings of Germon *et al* (2007) had supported the hypothesis that an integration of anomalous DNA at the 3` end of a tRNA site would be regarded as a disadvantageous. This is because some tRNA sites have high transcription rate or are encoding tRNA molecules that recognize frequently used codons. Therefore, integration of anomalous DNA at these sites would have severe effects on bacterial growth. On the other hand, Germon *et al* (2007) suggest that integration of such foreign DNA is more common downstream of tRNA sites with rare transcription rate or when tRNA genes encode tRNA molecules with less frequently used codons. Furthermore, Germon *et al* {2007) study has emphasized on the importance of the co-transcribed downstream genes of a tRNA gene as the disruption of these genes could have severe effects on the bacteria

(Bosl and Kersten., 1991; Fournier and Ozeki., 1985; Li and Deutscher., 2002; Lee *et al*., 1981; Schnell *et al*., 2003).

However, because of the lack of information concerning the affinity of tRNA for different anticodons which is used to calculate the tDNA usage and because Germon *et al* (2007) had used data from Adrell and Kirsebom (2005) which are only estimated transcription rates of tDNAs, no such conclusions could be estimated from our results as more precise data are required to draw such conclusions. On the other hand, our data support the hypothesis that the conditions in which a tDNA is expressed can influence its role as an insertion site for anomalous DNA. For instance, the expression of 4 $tRNA^{Leu}$ genes was found to be dependent on growth conditions except for *leuX* tDNA (Dobrindt and Hacker., 2001; Rowley *et al*., 1993), which is concurrent with our results for the *leuX* site. The *leuX* site had 8/10 strains occupied with anomalous DNA entities. Further studies will be required to elucidate whether the integration of other DNA loci with anomalous DNA is dependent on growth conditions or not.

As the sequence information of the *E. coli* genome increases, the *E. coli* pan-genome ("defined as the total nonredundant DNA set occurring among all *E. coli* genomes") has also increased (Medini *et al*., 2005; Johnson *et al*., 2007). However, the number of the novel DNA sequence contributed by each newly sequenced strain has declined as well as the number of ORFs being considered strain specific e.g., 20% of the CFT073 ORFs were thought to be strain-specific at the time of its genome sequence publication (Welch *et al*., 2002; Johnson *et al*., 2007). However, subsequent publication of other ExPEC DNA sequences showed that most of these unique CFT073 sequences were shared by other genomes (Brzuszkiewicz *et al*., 2006; Chen *et al*., 2006). More recently, Johnson *et al* (2007) had shown that 4.5% of the APEC O1's chromosome is absent from other sequenced *E. coli* genomes. However, three of the novel sequences previously identified in this study (E106_*serW*, E108_*serW* and E111_*serW*) were found to be part of GEIs sequences present in APEC O1 and UTI89 (These GEIs were identified by the tRNAcc method; Island screen online application). Therefore, the sequence of these three anomalous DNA sequences associated with the *serW* site should be further characterized to elucidate the full sequence of these GEIs and compare it to

the newly sequenced genome of APEC O1. Such comparative analysis between strains of ExPEC could define a common backbone for these strains containing sequences not found in other pathogenic strains of *E. coli*.

To conclude, the tRIP strategy has proved to be a very efficient strategy to identify GEIs associated with tRNA genes and further characterization of the many identified islands could help elucidate the role that these elements play in pathogenesis and the evolution of *E. coli*. Remarkably, despite only studying 10 *E. coli* strains, associated with a single disease type, we have identified at least 3 GEIs that contain novel sequences and one particular strain E105 had 13 tRNA sites occupied with GEIs. The full-length sequences of these 3 GEIs with novel DNA sequence are currently being investigated by using a capture vector containing two conserved targeting sequences flanking each end of the region of interest (GEI) within the genome. The method was developed by Wolfgang *et al* (2003) and has been used successfully in capturing an 80 kb genomic island from *Pseudomonas aeruginosa*. As a continuation to this project our group has used the same capture vector system and we identified that the full-length sequence of E105_*leuX* GEI was found to be ~9.6 kb. Preliminary analysis of the sequence showed that most of the ORFs were derived from prophages. The sequencing and characterization of full-length sequence of another GEI E104_*aspV* (~ 20kb) is in progress and would be available within the few coming months.

Chapter 4 Physical sizing of bacterial genomes in BSI-associated *E. coli* strains using rare cutting enzymes

4.1. Introduction

Beginning in the early 1980s, several experimental approaches have been made to obtain physical maps of megabase DNA fragments: pulsed-field gel electrophoresis (PFGE)-related methods, genome encyclopedia "ordered sets of overlapping clones containing the entire genome", and total DNA sequencing (Fonstein and Haselkorn., 1998). The ideal targets for these physical approaches are bacterial genomes, as these range from 600 kb for *Mycoplasma genitalium* to 9.5 Mb for *Myxococcus xanthus* (Fonstein *et al.*, 1995). The experimental approach chosen to be used to characterise a prokaryotic genome depends on the future use of the map and the degree of resolution required. Both gene encyclopedia and total DNA sequencing are used when complete characterisation of a microorganism is required. On the other hand, PFGE is more associated with comparative and epidemiological studies as these require a lower degree of resolution (Cole *et al.*, 1994).

In this chapter the PFGE has been used to obtain accurate sizing of the genomes for the 10 clinical bloodstream *E. coli* strains studied in this work. This has been achieved by dividing the megabase DNA fragments into different groups according to their size. These were then sized using different pulsed-field conditions. Physical mapping of microbial genomes using PFGE requires that these genomes being restricted and digested before being separated, as sizing of large circular molecules on agarose gels is more complicated than sizing linear DNA molecules e.g., large circular DNA molecules are generally fainter than restricted and digested DNA. Moreover, they tend to move slowly on agarose gels compared to linear DNA molecules (Jumas-Bilak *et al.*, 1995). Three different approaches could be used when trying to digest the genomic DNA for PFGE separation using rare cutting restriction enzymes:

4.1.1. Enzymatic methods to alter restriction enzyme specificity

The restriction enzyme *Dpn*I (5`-G$^{m}$ATC-3`), has been used to cut double-stranded DNA molecules when both strands are N^{6}-methylated at adenine residues (McClelland *et al.*, 1984). Using *Dpn*I in combination with various methyltransferases that recognize a sequence overlapping *Dpn*I restriction site by 2 or 3 bp has proved effective in producing recognition sequences of eight to twelve bp (McClelland *et al.*, 1984; Hanish and McClelland., 1991). In a second approach, 'cross-protection', large DNA fragments can be generated by the fact that restriction enzymes are sensitive to methylation at most bases within the target site (Nelson *et al.*, 1984; Nelson *et al.*, 1993). In this approach a defined subset of restriction target sites are blocked from cleavage at partly overlapping methylase/restriction endonuclease sites by prior methylation. For example, modification by M.*Fnu*DII or M.*Bep*I (5'-^{m}CGCG-3') blocks *Not*I cleavage at overlapping sites (5'-CGCGGCCGC-3', which is equivalent to 5'-GCGGCCGC-3') and increases the apparent specificity of *Not*I digestion about two-fold (Qiang *et al.*, 1990). Other approaches such as the competition reaction between methylases and endonucleases have proved to be useful in physical genome sizing as the latter strategies do not always work well if the DNA substrate is embedded in agarose blocks. Because both enzymes have the same target site specificity only partial cleavage of DNA with restriction endonuclease is achieved in this method (Hanish and McClelland., 1990).

4.1.2. Transposons carrying rare cleavage sites

One way of introducing rare restriction sites into bacterial genomes is by carrying these sites on transposons. Transposons often differ in their DNA (G+C) content from the rest of the genome, reflecting their recent origin in phylogenetically distant species. One example involved the rare *Not*I site that occurs in Tn*5* and which resulted in a new site in the *E. coli* genome wherever it was integrated (McClelland *et al.*, 1987; Heath *et al.*, 1992). Tn*5* pulse field mapping (pfm) constructs were used as valuable tools in pulsed-field mapping of gram-negative bacterial genomes. Wong and McClelland (1992) had used the *Bln*I sites in both Tn*5* (pfm) and Tn*10* transposons to make restriction map of *S. typhimurium* LT2. Moreover, different combinations of Tn*5* (pfm) insertions and

Tn*10* insertions were transduced by selecting for the corresponding Tc and Km markers. For example, Suwanto and Kaplan (1992) have constructed a Tn*5* derivative that contains an *Ase*I, *Dra*I and *Sna*BI recognition site [rare in (G+C)-rich genomes] and Jumas Bilak *et al* (1995) had developed a Tn*5* derivative containing the 18 bp I-*Sce*I site, delivered from a RP4-mobilizable, RK6-derived suicide vector.

Other strategies use rare restriction sites within vectors bearing cloned sequences that undergo homologous recombination with the genome of the species from which they were derived. For example, Le Bourgeois *et al* (1992) used this strategy to map genes in *Lactococcus lactis* using a vector carrying sites for *Apa*I, *Not*I and *Sma*I.

4.1.3. Intron-encoded endonucleases

The homing endonuclease I-*Ceu*I is encoded by a class I mobile intron. The enzyme which is specific for a 19 bp sequence is involved in the homing of the intron by inserting it into the *rrl* gene of the large subunit ribosomal RNA (23S rRNA) in the chloroplast DNA of intronless strains (Colleaux *et al.*, 1986; Dujon *et al.*, 1989; Gauthier *et al.*, 1991; Marshall *et al.*, 1994; Dujon *et al.*, 1989; Gauthier *et al.*, 1991; Marshall *et al.*, 1994). The sequence of the *rrl* genes has been found to be highly conserved among the genomes of many bacteria as well as in chloroplast and mitochondria of eukaryotes. Due to the large number of bases in the recognition site, it is predicted not to occur at other sites in the bacterial genome. I-*Ceu*I sites are present in all seven *rrl* genes in enteric bacteria, but at no other locations (Liu *et al.*, 1993; Honeycutt *et al.*, 1993). The 19 bp sequence for I-*Ceu*I recognition and cleavage is found in the *rrn* operons of a large number of prokaryotes, but not in eukaryotes (Liu *et al.*, 1993). Thus, the use of I-*Ceu*I in PFGE physical mapping allows the determination of the number, location and orientation of *rrn* loci on the physical map of a eubacterium (Honeycutt *et al.*, 1993).

Because of the specificity of I-*Ceu*I in cleaving within the highly conserved *rrl* genes in enteric bacteria (Krawiec and Riley., 1990), the fingerprint produced from restriction digests with I-*Ceu*I between closely related bacteria is expected to be very similar. This

would provide valuable epidemiological information regarding the genome structure and can be used to draw phylogenetic relationships between related bacteria. For example, related wild type strains of *S. typhimurium* produced similar I-*Ceu*I fingerprints. However, partial I-*Ceu*I analysis of three independent strains of *S. typhi* revealed substantial genomic rearrangements rather than the stability seen in *S. typhimurium*. Therefore, it can be presumed that genomic rearrangements are rare in *S. typhimurium* but common in *S. typhi* (Liu and Sanderson., 1995). In contrast, the cleavage sites for class II restriction endonucleases are usually only 6 bp or 8 bp and are present everywhere in the genome, including less well conserved genes. These sites are frequently gained or lost during evolution due to nucleotide mutation, producing different fingerprints even for related species. Fingerprints generated by digestion with *Xba*I, a type II restriction endonuclease with a 6 bp site, are much more variable, providing little direct genomic information. This explains why data from Type II restriction endonucleases such as *Xba*I, *Bln*I, *Not*I and others cannot be used for genomic comparisons in the same direct ways as for I-*Ceu*I.

4.2. Results

4.2.1. Physical mapping of the bacterial genomes using transposon carrying rare cleavage sites

The aim of this experiment was to introduce a unique restriction site to the chromosomes of ten *E. coli* BSI strains and to use the linearized genomes to obtain the precise size of the bacterial chromosome. The pGF2 plasmid (11239 bp – Figure 4.1) a derivative of ColE1::Tn*5*Map and pGP704 (Jumas-Bilak *et al*., 1995) was used to transfer the Tn*5*Map containing the unique restriction site for I-*Sce*I to the *E. coli* isolates. This endonuclease recognizes a non-symmetrical double-stranded 18 bp sequence 5` TAGGGATAACAGGGTAAT 3`. The enzyme is encoded by a group I intron (r1 intron) of the *Saccharomyces cerevisiae* mitochondrial 21S rRNA gene. This intron propagates itself during crosses between intron plus-strains (omega^{+}) and intron-minus ones (omega^{-}) (Colleaux *et al.*, 1988). The enzyme was overexpressed in *E. coli* as a protein identical to the expected genuine mitochondrial protein. Because the yeast mitochondrial genetic code differs from the universal code, a new protein was

constructed using a universal code equivalent of the r1 ORF of *S. cerevisiae*, by oligonucleotide-directed mutagenesis. This universal code equivalent efficiently directs the synthesis in *E. coli* of a full length protein which bears all characteristics predicted for the native mitochondrial protein. This protein generates a specific double strand cut at the omega$^-$ site (Colleaux *et al.*, 1986), but has no cleavage sites found for the I-*Sce*I in the entire yeast genomes and in a range of bacterial genomes (Jumas-Bilak *et al*., 1995; Monteilhet *et al*., 1990 and Perrin *et al*., 1993). To determine whether I-*Sce*I could be used on naturally occurring genomes, Thierry *et al* (1991) have analyzed a number of genomes under optimal conditions and no sites were found in the genomes of the bacteriophages λ, ΦX174, SSP 1, T4 or T5, for the Adenovirus Ad 2 or for the bacteria *Bacillus anthracis, Borrelia burgdorferi, Leptospira biflexa* and *Leptospira interrogans.* In contrast, one site is present on the genome of phage T7. Cleavage at this site, however, does not go to completion in the conditions used by Thierry *et al* (1991), consistent with the fact that no sequence corresponding to the 18 bp wild-type recognition sites occurs in the entire T7 DNA sequence.

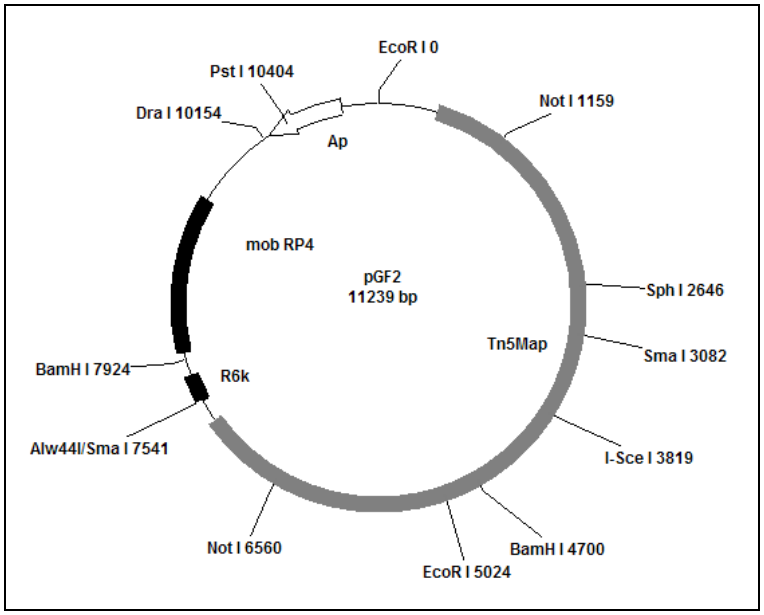


Figure 4.1. Structure of plasmid pGF2. The gray shaded area refers to the Tn*5*Map part, the large black shaded area refers to the mob RP4 fragment and the small black shaded area refers to the R6K fragment. The white arrow refers to the ampicillin gene. Redrawn from Jumas-Bilak *et al* (1995).

pGF2 plasmid has the R6K origin of replication and the *ori*T sequence from RP4. The plasmid was transferred to *E. coli* SM10*λpir* (Simon *et al.*, 1983; Jumas-Bilak *et al*., 1995), which provides the RP4 *tra* functions needed for the conjugation experiments and the *pir* gene encoding the R6Kπ protein needed for the plasmid replication. Plasmid R6K is a 38 kb resistance transfer plasmid, encoding resistance to ampicillin (Ap) and streptomycin (Sm), that occurs naturally in *E. coli* in many copies per cell. The construction of low molecular weight derivatives of this plasmid has defined the essential replication region of R6K within no more than a 2.1 kb segment of the DNA (Kolter and Inuzuka., 1978). Inuzuka and Helinski (1978) have developed an *in vitro* system for R6K replication and have shown that R6K replication requires a plasmid-encoded protein, designated π for initiation of DNA replication. Kolter and Inuzuka (1978) have described the isolation of replication mutants of R6K derivatives and the construction of a trans-complementation system in which the gene coding for the π protein, inserted in the host chromosome or cloned on a ColE1 plasmid derivative, trans-complements the replication of a small segment of a 420 bp of R6K DNA carrying a functional origin of R6K DNA replication.

This small fragment of 420 bp containing the origin of replication of plasmid R6K was cloned into pGP704, a derivative of pBR322 that has a deletion of the pBR322 origin of replication (*ori*El) (Kolter and Inuzuka., 1978). pGP704 also contains a 1.9-kb *Bam*HI fragment encoding the mob region of RP4. The pGP704 mobilization by the RP4 was trans-complemented by the transfer functions derived from RP4 and integrated in the chromosome of SM10*λpir* (Kolter *et al.*, 1978; Inuzuka and Helinski., 1978; Miller and Mekalanos., 1988). pGP704 was then ligated to 7541 bp fragment of the plasmid ColE1::Tn*5*Map to construct the pGF2 plasmid (Jumas-Bilak *et al*., 1995).

4.2.2. pGF2 plasmid extraction and restriction map

The plasmid was extracted using a midi-preparation (10ml cultures) of the alkaline lysis method described in materials and methods. In order to confirm that the plasmid extracted was pGF2, four different restriction enzymes (*Eco*RI, *Bam*HI, *Hin*dIII, and *Sal*I) were used to confirm the size and restriction map of the plasmid. The size of the

DNA fragments produced by the restriction digests was determined by agarose electrophoresis on 0.8% agarose gels. The results of the restriction digests were tabulated and are presented in (Table 4.1 and Figures 4.2 and 4.3). When the DNA fragments for each restriction digest were summed the total size was between 9.3 and 11.2 kb. Such contradiction in the total size of the same plasmid produced using different restriction digests could be attributed to limitations associated with gel electrophoresis. For example, it is considered difficult to resolve DNA fragments with nearly identical size or to resolve small DNA fragments < 60 bp on 0.8% agarose gel. However, the number of cleavage sites and the sizes of the DNA fragments produced by each restriction enzyme do not correspond with the plasmid structure map provided by the original author (Figure 4.1) (Jumas-Bilak *et al.,* 1995). Tracing back the construction of pGF2 revealed that there were many restriction sites not represented in the map by Jumas-Bilak *et al.*, and consequently making a real comparison between the restriction digests of the plasmid and the map of pGF2 a hard task.

Table 4.1. Frequencies of the restriction enzyme cleavage sites and the size of DNA fragments produced for pGF2 plasmid digests.

Restriction enzyme	No. of observed sites	Size of DNA fragments (kb)	Sum of DNA fragments (kb)
*Eco*RI	1	11.2	11.2
*Bam*HI	4	4.2, 3.5, 1.7, 1.4	10.8
*Hin*dIII	6	3.4, 1.9, 1.6, 1.5, 0.5, 0.4	9.3
*Sal*I	4	6.5, 3.1, 0.5, 0.4	10.5

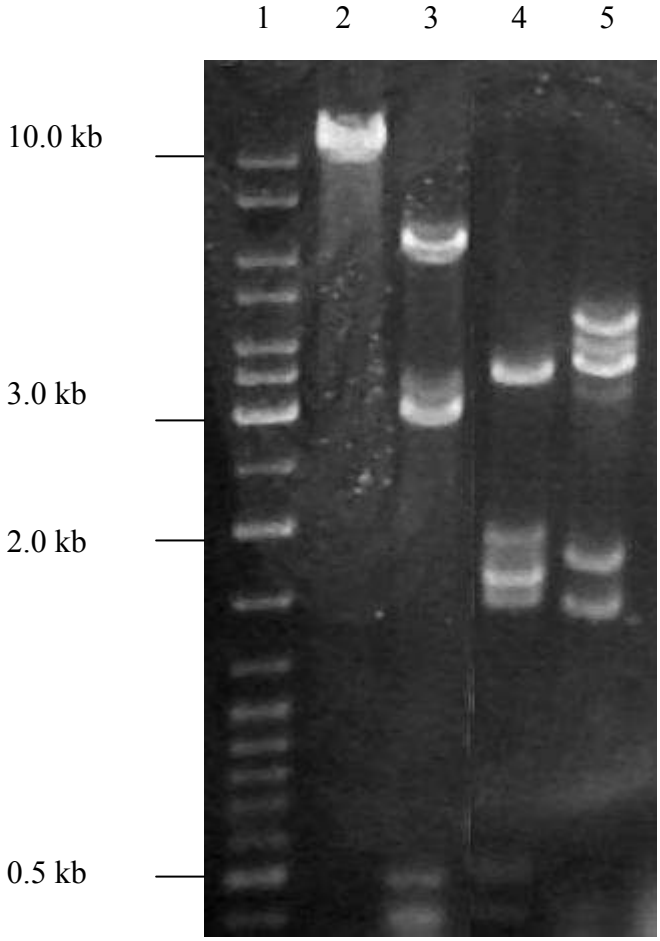


Figure 4.2. Restriction digests of pGF2. Lanes: 1, 0.5µg of GeneRuler™ DNA ladder mix; 2, *EcoR*I digest of pGF2; 3, *Sal*I digest of pGF2; 4, *Hin*dIII digest of pGF2; 5, *Bam*HI digest of pGF2.

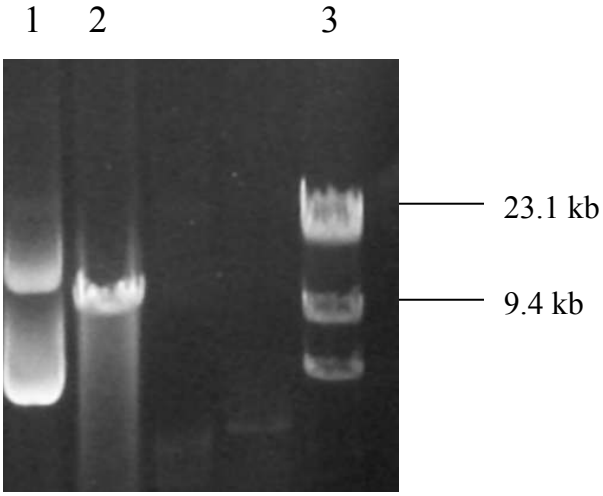


Figure 4.3. Restriction digests of pGF2. Lanes: 1, pGF2 uncut plasmid; 2, *EcoR*I digest of pGF2; 3, 0.5µg of Lambda DNA/*Hin*dIII marker.

In order to confirm the restriction map of the plasmid, the plasmid was reconstructed using precursor plasmid DNA sequences of pGF2 obtained from plasmid specialized websites and from NCBI (Table 4.2). The restriction fragments obtained by *in silico* digestion of the reconstructed plasmid resembled the restriction digests obtained by experiment for *Eco*RI, *Bam*HI, and *Sal*I (Table 4.3). The *Hin*dIII digest produced more restriction sites than the *in silico* digest. This difference could be due to two missing DNA fragments of 400 bp (cos region) and the 347 bp containing the I-*Sce*I site, which could not be retrieved from their original sources (Figure 4.4).

Table 4.2. Websites used to reconstruct the pGF2 plasmid sequence.

Precursor plasmids/ DNA fragments	Accession no.	Sequence length (bp)	Reference	Website
ColE1	J01566.1	6646	Chan *et al*., 1985	http://www.ncbi.nlm.nih.gov
Tn*5*	U00004.1	5818	Berg., 1977	http://www.ncbi.nlm.nih.gov
ColE1::Tn*5*cos	-	4598	Zuber and Schumann	http://www.addgene.org
pBR322	J01749.1	4361	Kolter *et al*., 1978	http://www.ncbi.nlm.nih.gov
pSCM201	AY443099.2	3463	Hiller *et al*., 1994	http://www.ncbi.nlm.nih.gov
pGP704	-	3703	Miller and Mekalanos., 1988	http://www.addgene.org
pWP14	-	3068	Shinder and Gold., 1987	http://www.addgene.org

Table 4.3. Frequencies of the restriction enzyme cleavage sites and the size of DNA fragments produced for pGF2 plasmid *in silico* digests.

Restriction site	No. of observed sites	Size of DNA fragments (kb)	Sum of DNA fragments (kb)
*Eco*RI	1	11.2	11.2
*Bam*HI	4	4.8, 3.2, 1.7, 1.5	11.2
*Hin*dIII	5	3.5, 2.3, 1.9, 1.8, 1.7	11.2
*Sal*I	4	6.9, 3.6, 0.4, 0.3	11.2

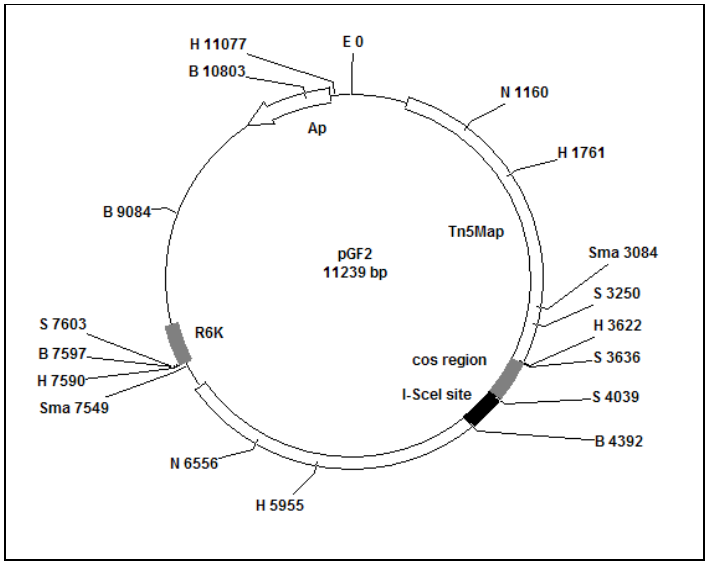


Figure 4.4. Map of pGF2. Reconstruction of the pGF2 plasmid map from *in silico* data. Two missing parts in the map (sequence data not available), the 400 bp (cos region) and the 347 bp containing the I-*Sce*I site. Abbreviations: B: *Bam*HI, E: *Eco*RI, H: *Hin*dIII, N: *Not*I, S: *Sal*I, Sma: *Sma*I.

4.2.3. Results of the conjugation experiment

For the conjugation step to take place the donor and the recipient cell were incubated overnight on tryptone soya agar plates without the addition of antibiotics. Then, transconjugants were selected by overnight incubation on M9 minimal medium supplemented with 25 μg ml^{-1} kanamycin and 15 μg ml^{-1} nalidixic acid. The SM10*λpir* (donor strain) is kanamycin resistant but sensitive to nalidixic acid; however the recipient strains varied when tested for their sensitivity to both antibiotics (Table 4.4). The transconjugants were resistant to both antibiotics and they were confirmed to be true transconjugants as they tested sensitive for ampicillin except for the strains that were found to be ampicillin resistant before the conjugation step (E102, E103 and E104, Table 4.4).

When the pGF2 plasmid is introduced to a recipient cell that lacks the π protein required by the plasmid R6K origin of replication, the plasmid will remain unreplicated and will eventually be lost. The only way the composite transposon (Tn*5*Map) can survive is by hopping and integrating into another plasmid or the chromosome of the recipient cell. When the Tn*5*Map is introduced into the bacterial chromosome two scenarios are proposed (Figure 4.5a and 4.5b). In the first (Outside-end transposition), the transposase encoded by the Tn*5*Map acts on the inverted repeats at the farthest ends of the composite transposon and so the Tn*5*Map DNA fragment will be introduced into the recipient chromosome. In the second strategy, the transposases will act on the inverted repeats at the inside ends of both IS*50* elements and in this case the pGP704 part of the pGF2 will be introduced into the chromosome (Snyder and Champness., 2003). Therefore, the putative transposon mutants are tested for being Tn*5*Map mutants by checking for kanamycin resistance (Tn*5*Map-encoded) and ampicillin sensitivity (Ap^R pGP704-encoded). However, because some of the original clinical isolates were kanamycin and ampicillin resistant (Table 4.4), the outcome of the conjugation and the transposition steps had to be confirmed by PCR analysis.

Table 4.4. Resistance profile of the ten clinical isolates to the supplemented antibiotics added to M9 medium for transconjugants selection.

Bacterial strain[a]	M9+glu[b]	M9+glu+thiam[c]	Nalidixic acid[d]	Kanamycin[e]	Ampicilin[f]
SM10*λpir*	-	-	-	+	+
DH5α	-	-	+	-	-
K12 (MG1655)	+	+	-	-	-
E102	+	+	+	+	+
E103	+	+	+	+	+
E104	+	+	+	+	+
E105	+	+	-	+	-
E106	-	+	-	+	+
E107	+	+	-	+	+
E108	+	+	-	+	-
E109	+	+	-	+	-
E110	+	+	-	+	+
E111	+	+	+	+	-

[a]'+', resistant; '-', sensitive.
[b]M9 medium supplemented with glucose.
[c]M9 medium supplemented with glucose and thiamine.
[d]The concentration of nalidixic acid was 15 μg ml^{-1}.
[e]The concentration of kanamycin was 25 μg ml^{-1}.
[f]The concentration of ampicillin was 100 μg ml^{-1}. Only three strains were resistance to the three antibiotics: *E. coli* E102, E103 and E104.

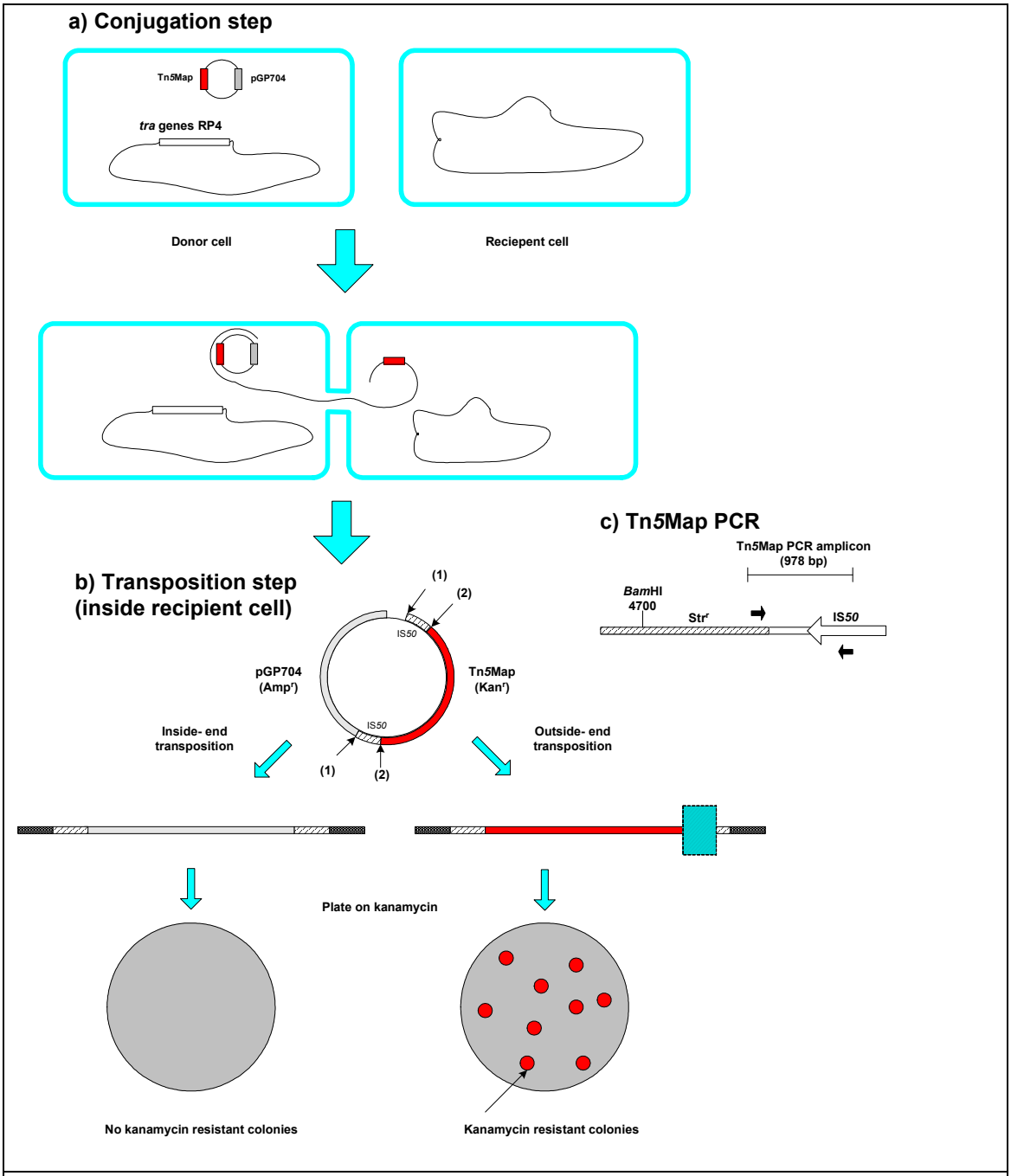


Figure 4.5. Transfer of Tn5Map into the *E. coli* BSI isolate chromosome and confirmation of the transposition step by PCR.

(a) Transfer of pGF2 plamid by conjugation from the donor cell (SM10*λpir*) into the recipient cell (*E. coli* BSI isolates).

(b) Transposition of Tn*5*Map composite transposon. The outside-end (1) transposition will transpose Tn5Map (red), including the gene encoding kanamycin, whereas inside-end (2) transposition will transpose the pGP704 DNA frgment (light blue) into the chromomsome of the recipient cell. Hashed bars indicate the IS*50* elements of the Tn*5*Map. Black dotted bars refer to the recipient chromosome.

(c) Tn*5*Map PCR using primers that amplify (978 bp) within the Tn5Map to confirm the presence of the Tn*5*Map and consequently the I-*Sce*I site within the recipient chromosome. The black arrows refer to the primers used to amplify the PCR DNA fragement.

4.2.4. Confirming the transposition step by PCR

The aim of this experiment was to confirm that the transconjugants had a PCR profile similar to that of the original recipient cell and not of the SM10*λpir* donor cell. The second step is to check that the transposition step had occurred by amplifying a region within the integrated Tn*5*Map site.

tRIP-PCR for the *pheU* and *serU* tRNA sites was done to check the PCR profile of SM10*λpir* (donor cell), recipient and the transconjugants: Both *pheU* and *serU* tRNA sites had been selected to perform the tRIP-PCR, because in both sites the up and down stream primers were designed using K12 as a template, therefore, K12 (MG1655) could be included in the PCR as a positive control. Furthermore, from previous experience with tRIP-PCRs with the 16 different tRNA sites investigated, both *pheU* and *serU* gave specific and single sharp band with all of the samples tested.

From the results of the *pheU* tRIP-PCR no difference can be observed between the PCR profiles as all of them had the same PCR amplicon size product (0.7 Kb) (Figure 4.6). However, with the tRIP-PCR for the *serU* tRNA site different profiles were observed (Figure 4.7). According to the gel picture both E104 and E222 (E104 harboring Tn*5*Map) had a PCR product of 1 Kb, and for E105 and E223 (E105 harboring Tn*5*Map) both had no PCR products, while for the donor a 2.3 kb PCR product was amplified. Therefore, it can be concluded that after the conjugation experiment only Tn*5*Map mutants had been selected.

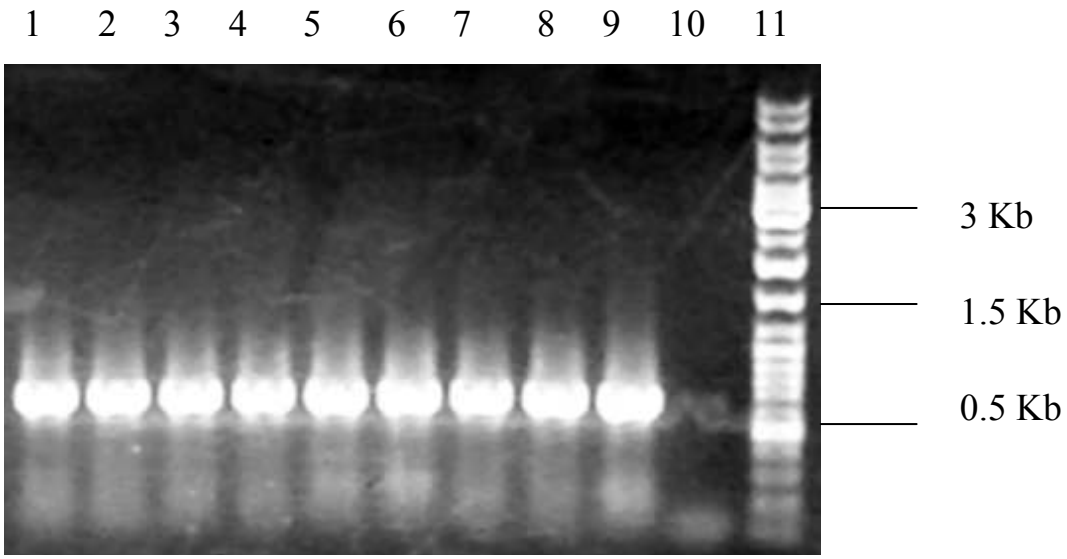


Figure 4.6. tRIP-PCR for *pheU* tRNA site. Lanes: 1, K12 (MG1655); 2, E104; 3 and 4, E105; 5 and 6, E222; 7 and 8, E223; 9, *SM10λpir*; 10, PCR negative control (distilled water added instead of DNA); 11, 0.5 µg of GeneRuler™ DNA ladder mix.

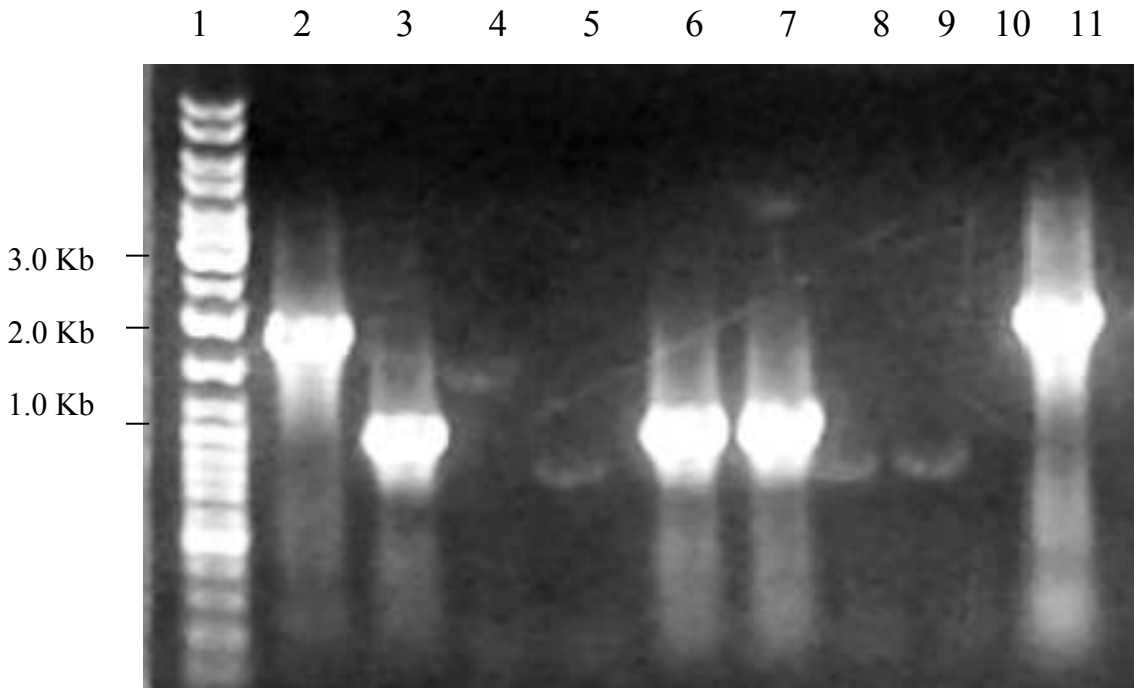


Figure 4.7. tRIP-PCR for *serU* tRNA site. Lanes: 1, 0.5 µg of GeneRuler™ DNA ladder mix; 2, MG1655; 3, E104; 4 and 5, E105; 6 and 7, E222; 8 and 9, E223; 10, PCR negative control; 11, *SM10λpir*.

Amplifying within the Tn*5*Map: A 978 bp DNA fragment located between the streptomycin resistant gene and the IS*50* in Tn*5*Map was amplified to prove the integration of Tn*5*Map into the chromosome of the recipient cell (Figure 4.5c). The corresponding DNA fragment of 978 bp was obtained with the positive control pGF2 plasmid DNA and the transconjugant strain E223 but not the original recipient cell E105 (Figure 4.8).

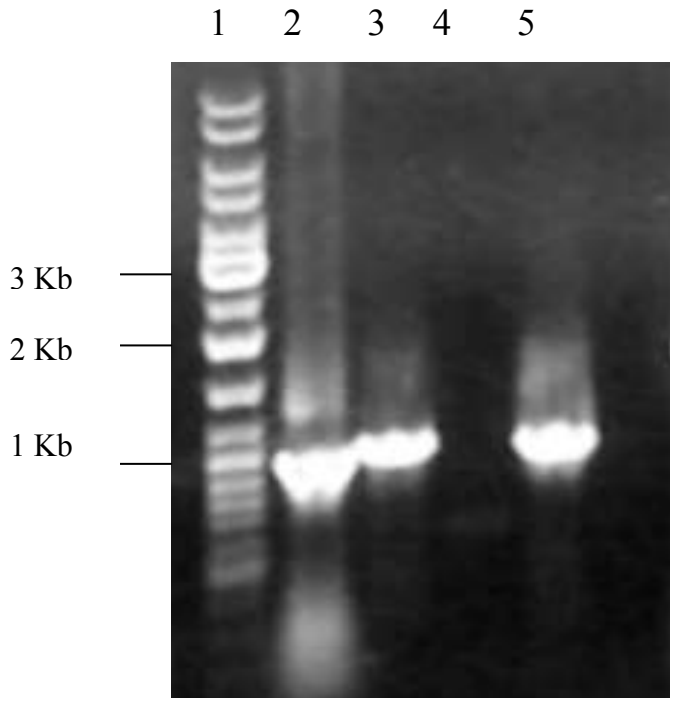


Figure 4.8. Tn5Map PCR. Lanes: 1, 0.5 µg of GeneRuler^{TM} DNA ladder mix; 2, Tn*5*Map PCR product using pGF2 plasmid as a template DNA; 3 and 5, Tn*5*Map PCR product using E223; 4, PCR negative control (E105 template DNA).

4.2.5. Genome sizing of *E. coli* BSI-associated chromosomes using I-*Sce*I

The aim of this experiment was to size the bacterial genomes of the 10 clinical BSI *E. coli* strains using the introduced and unique restriction site I-*Sce*I. Sizing the bacterial genomes with such approach has more advantages over other methods, which use restriction endonucleases that cut more frequently in the genome. Digesting the bacterial genome with the unique I-*Sce*I is presumed to enable accurate relative sizing of the genome as only one band representing the size of the whole chromosome is being

measured. Moreover, because the whole genome would run on the gel as a one band, no minor bands would be missed from the gel. Finally, using this method would avoid problems usually associated with similar bands that form doublets.

Jumas-Bilak *et al* (1995) used the following run conditions to separate bands in the size range of 1-2 Mb for I-*Sce*I digested genomes of *E. coli* strains: switch time of 60-130 s at 6 volts/cm in a 0.8% agarose gel for 24 hours. The same parameters were applied to check the reproducibility of the method (Figure 4.9), however, under these conditions the linearized genomes moved at the same rate regardless of genome differences between different strains, this is because these run conditions (switch time, and the length of run) are not suitable for the separation of DNA fragments in the size range of 3-5 Mb. Therefore, the run conditions were optimized in order to obtain accurate sizing of the genomes.

The following parameters were obtained from the BioRad CHEF-DR® II pulsed-field gel electrophoresis instruction manual and were used to separate DNA fragments in the size range of 2.4-3.0 or 3.5-5.7 Mb: switch time of 900-1200 or 1800-3600 s at 3 or 2 volts/cm in a 0.8% agarose gel for 72 or 144 hours respectively. However, the DNA of all samples was either degraded or partially digested under these conditions (Figure 4.10). The same results were obtained when the experiment was repeated with the addition of 50 μM thiourea to the running buffer (Tris-HCl) or the use of HEPES as a running buffer instead of the Tris-HCl (both chemicals were reported to inhibit the Tris-HCl DNA degradation effects during electrophoresis (Ray *et al*, 1992).

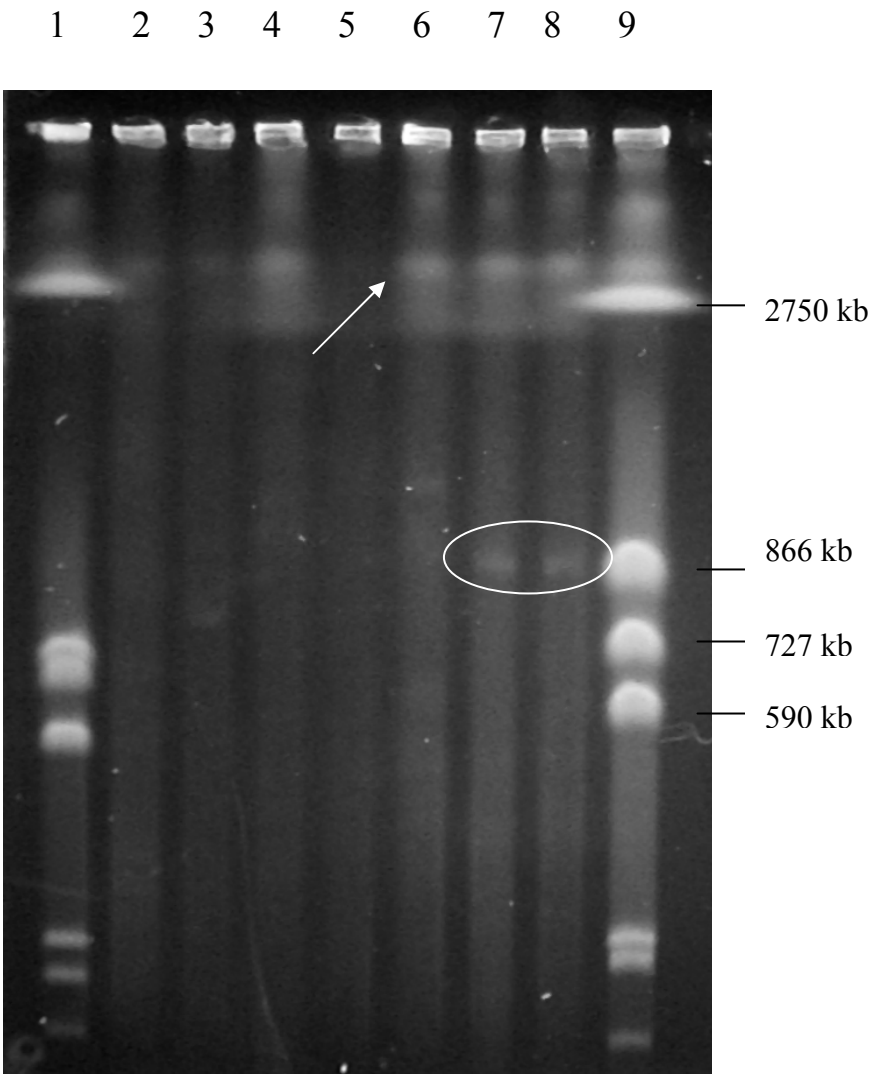


Figure 4.9. Run of HMW DNA of *E. coli* strains digested with I-*Sce*I. Under the run conditions reported by Jumas-Bilak *et al* (1995) all linearized genomes move at the same rate and no differentiation can be made between the different genomes as seen from the gel picture. Lanes: 1, I-*Ceu*I digest of MG1655 DNA used as a DNA marker; 2, I-*Sce*I digest of E217; 3, I-*Sce*I digest of E222; 4, I-*Sce*I digest of E223; 5, I-*Sce*I digest of E224; 6, I-*Sce*I digest of E225; 7, I-*Sce*I digest of E226; 8, I-*Sce*I digest of E227; and lane 9, I-*Ceu*I digest of CFT073 DNA used as a DNA marker. White arrow points to the genomes linearized with I-*Sce*I. The faint bands observed with E226 and E227 are highlighted with white circle, however, these bands were not reproducible.

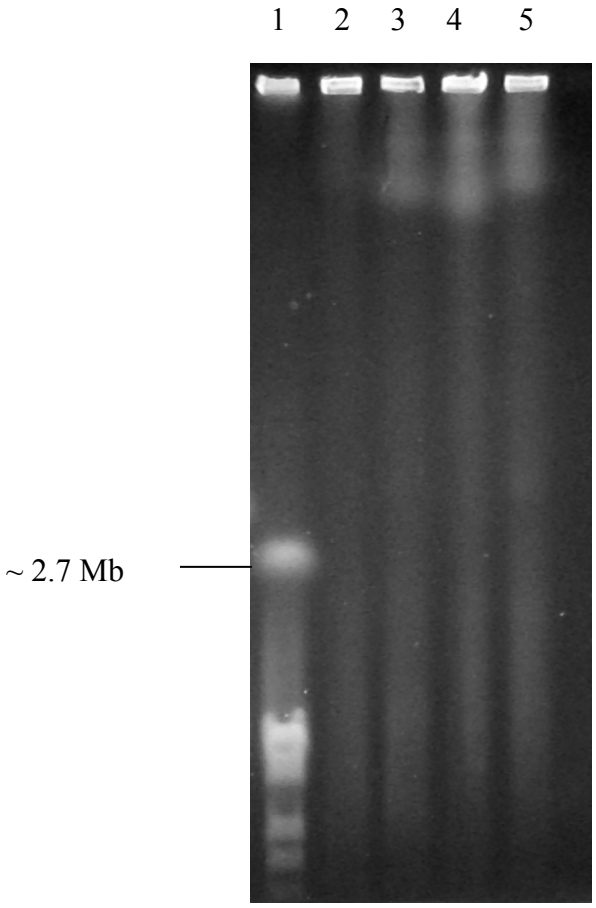


Figure 4.10. Run of HMW DNA of *E. coli* strains digested with I-*Sce*I. Run conditions: switch time of 900-1200 s at 3 volts/cm in a 0.8% agarose gel for 72 hours Lanes: 1, I-*Ceu*I digest of E224 used as a DNA marker; 2, Undigested DNA of CFT073 used as a negative control (non-linearized DNA); 3, I-*Sce*I digest of E223; 4, I-*Sce*I digest of E225; and 5, I-*Sce*I digest of E227.

4.2.6. Genome sizing of *E. coli* BSI-associated chromosomes using intron-encoded endonuclease I-*Ceu*I:

As an alternative approach for the I-*Sce*I method, another homing endonuclease I-*Ceu*I was applied. I-*Ceu*I specifically digests the 23S rDNA sequence in *rrn* operons. In most Gram-negative facultative anaerobic bacteria, *rrn* operons have a copy number of five to eight with almost identical sequences (Shu *et al,* 2000); and with such a low copy number, sizing the bacterial genome using I-*Ceu*I will produce more reliable genomic sizing and will reduce the percentage error compared to other endonucleases that cut more frequently in the genomic DNA.

To achieve such accurate sizing of the bacterial genomes, three different size ranges were first identified after running trial experiments for the HMW DNA of the 10 clinical isolates digested with I-*Ceu*I. The three different size ranges were: the first genomic DNA size range was 40-400 kb (Figure 4.11), the second size range was between 400-1000 kb (Figure 4.12) and the third was between 1-3.1 Mb (Figure 4.13). The produced sizes are summarized in Table 4.5. Two positive controls (MG1655 and CFT073) were used in the process to estimate the percentage error in each experiment Table 4.6. The tabulated results in Table 4.5 are the mean of the measurements of two replicate experiments that were run using the same conditions e.g., switch time, volts, agarose concentration and the duration of the run and the run buffer used.

Out of the ten *E. coli* clinical isolates two strains E102 and E103 showed DNA degradation (Figures 4.12 and 4.13). The addition of the 50 μM thiourea to the Tris-HCl running buffer enhanced the sizing of DNA genomic bands in the size range of 40-400 kb (Figure 4.11) but not for larger DNA bands. On the other hand, the use of the HEPES as a running buffer instead of Tris-HCl improved the sizing of genomic bands in the size range of 400 to 1000 kb (Figure 4.14) but not in the size range of 1-3.1 Mb.

The PFGE results for strain E107 showed two faint bands (Figure 4.12) sized 0.696 and 0.728 Mb. The band intensities and the size range of these two faint bands were concordant with circular plasmids bands from similar and previous studies (Kinashi *et*

al., 1987; Lennon and DeCicco., 1991; Buchrieser *et al*., 1994). Circular plasmids move in PFGE according to their apparent mobility. Therefore, the size of the two plasmids in E107 could be attributed to the apparent mobility of the two plasmids. An estimated actual size of these two plasmids elucidated from the literature (Kinashi *et al*., 1987; Lennon and DeCicco., 1991; Buchrieser *et al*., 1994) is thought to be between 90-130 kb.

Attempts to extract the two large plasmids with the alkaline method (Kado and Liu., 1981) had failed and several modifications were applied e.g., the addition of phenol-chloroform to extract the plasmid but no consistent results could be confirmed.

Table 4.5. Summary Table for the ten clinical strains and two positive controls (CFT073, MG1655) genome sizes.

Strain	*E.coli*	*E.coli*	*E.coli* E102	*E.coli* E103	*E.coli* E104	*E coli* E105	*E coli* E106	*E.coli* E107	*E.coli* E108[b]	*E.coli* E109	*E.coli* E110	*E.coli* E111[b]
DNA band no.	(MG1655) [a]	CFT073										
1	2.7	2.9	2.8[c]	2.8[c]	3	2.9	2.8	2.7	2.7	2.8	2.9	2.8
2	0.703	0.776	0.806[d]	0.804[d]	0.798	0.736	0.749	0.86	0.782	0.817	0.794	0.786
3	0.651	0.724	0.738 [d]	0.74 [d]	0.693	0.686	0.683	0.728[e]	0.559	0.796	0.686	0.552
4	0.53	0.589	0.562	0.561	0.564	0.56	0.542	0.696[e]	0.14	0.551	0.575	0.14
5	0.131	0.13	0.134	0.133	0.143	0.139	0.137	0.575	0.122	0.141	0.14	0.123
6	0.095	0.11	0.104	0.102	0.096	0.113	0.119	0.537	0.047	0.12	0.121	0.049
7	0.042	0.02	0.044	0.044	0.042	0.043	0.045	0.169		0.047	0.046	
8								0.105				
9								0.048				
Total genome size (Mb)	4.852	5.249	5.188	5.184	5.336	5.177	5.075	6.418[4.994][f]	4.350[5.132]	5.272	5.262	4.450[5.236]

[a]The measured bands are the mean of two replicates run under the same conditions. The SD for K12 (MG1655) bands was as following: band no. 1, 0.007; 2, 0.003; 3, 3.8×10^{-5}; 4 to 7, 0. Total SD was 0.003. The bands were measured in Mb.

[b]Because of the doubled band intensities for the 782 and 786 kb of E108 and E111 respectively, (Figure 4.12), the total size of the genomes were recalculated to account for the doublet bands. The total genome sizes after adding the duplicate bands were indicated between square brackets.

[c]Band number 1 of strains E102 and E103 had experience severe DNA degradation under the electrophoresis conditions specified for the size range of 1-3.1 Mb and therefore the sizes were roughly estimated using gels that run to separate bands in the size range of 400-1000 kb.

[d]Bands number 2 and 3 for strains E102 and E103 were measured in a separate gel run with HEPES buffer (Figure 4.14) (the size of the bands measured for MG1655 and CFT073 under the same buffer were as following; MG1655: 0.670 Mb (3.9% error) and 0.618 Mb (6.0% error). For CFT073: 0.719 Mb, (1.2% error) and 0.548 Mb (7.2% error).

[e] Bands number 3 and 4 of strain E107 are presumed to be large plasmid DNA (both observed as faint bands Figure 4.12).

[f]The total genome size of strain E107 after excluding bands number 3 and 4 which are presumed to refer to plasmid DNA fragments was indicated between square brackets.

Table 4.6. Calculations of the percentage error associated with each genomic band for K12 MG1655 and CFT073.

Strain[a]	E. coli MG1655			E. coli CFT073		
DNA band no.	*in silico* DNA size	Experimental DNA size	% Error	*in silico* DNA size	Experimental DNA size	% Error
1	2.498	2.70	-8.1	2.751	2.9	-5.4
2	0.698	0.703	-0.7	0.866	0.776	10.5
3	0.657	0.651	1	0.727	0.724	0.5
4	0.521	0.53	-1.8	0.59	0.589	0.3
5	0.131	0.131	0.4	0.138	0.13	5.4
6	0.094	0.095	-0.9	0.118	0.11	7.3
7	0.041	0.042	-1.2	0.04	0.02	51.1
Total size (Mb)	4.64	4.852	-4.6	5.23	5.249	-0.36

[a]The bands are measured in Mb.The total genome sizes for MG1655 K12 (RefSeq accession no. NC_000913) and CFT073 (RefSeq accession no. NC_004431) were obtained from the NCBI web site (http://www.ncbi.nlm.nih.gov).

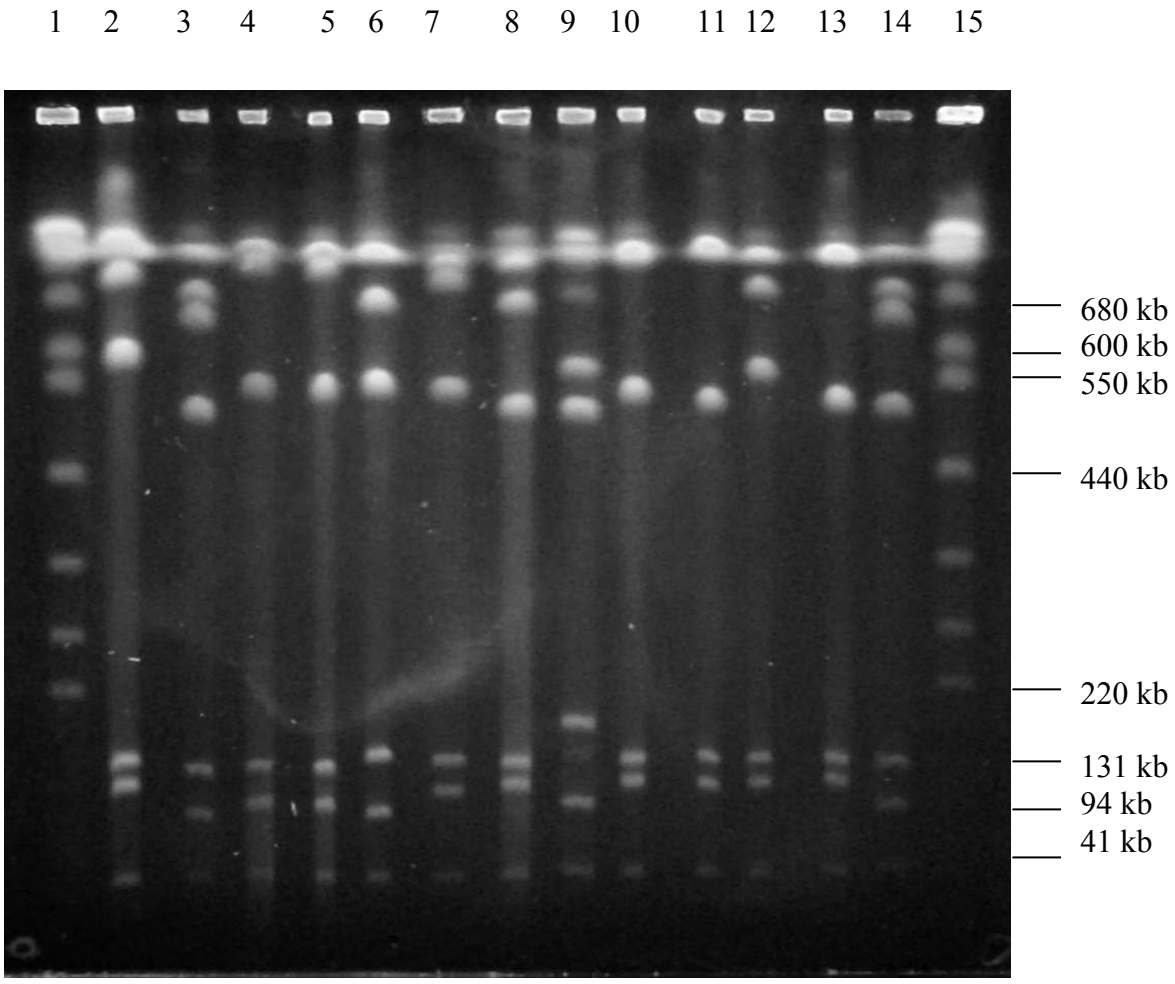


Figure 4.11. Run of HMW DNA (size range 40-400 kb) of *E. coli* strains digested with I-*Ceu*I. Run conditions for the size range of 40-400 kb: switch time of 20-60 s at 6 volts/cm in a 1.0 % agarose gel for 24 hours, with addition of 50 μM thiourea to the Tris-HCl running buffer. Lanes: 1 and 15, *Saccharomyces cerevisiae* strain YPH80 megabase DNA marker (CAMBREX); 2, I-*Ceu*I digest of *E. coli* CFT073 used as a positive control and a DNA marker; 3 and 14, I-*Ceu*I digest of *E. coli* K12 (MG1655) used as a positive control and a DNA marker; 4, I-*Ceu*I digest of strain E102; 5, I-*Ceu*I digest of strain E103; 6, I-*Ceu*I digest of strain E104; 7, I-*Ceu*I digest of strain E105; 8, I-*Ceu*I digest of strain E106; 9, I-*Ceu*I digest of strain E107; 10, I-*Ceu*I digest of strain E108; 11, I-*Ceu*I digest of strain E109; 12, I-*Ceu*I digest of strain E110; 13, I-*Ceu*I digest of strain E111.

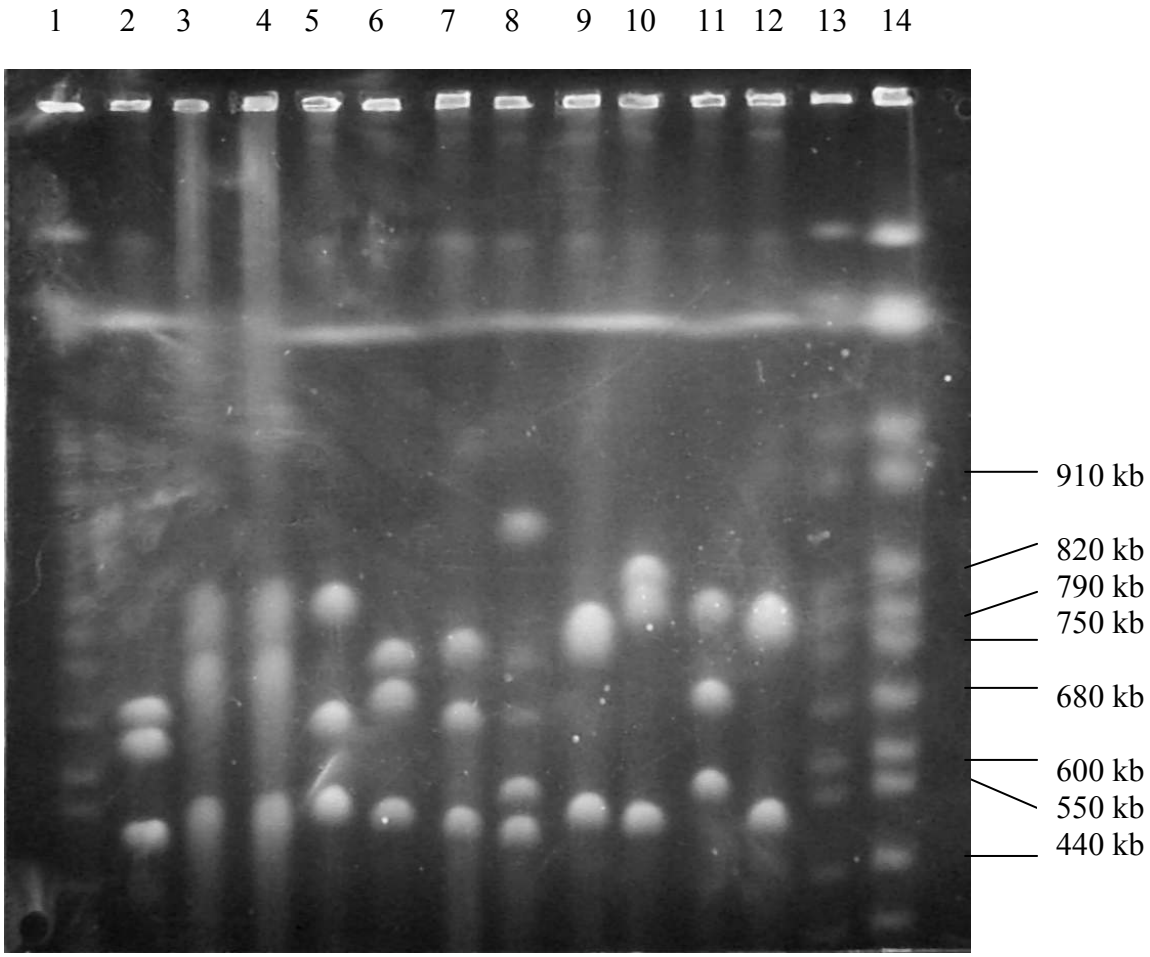


Figure 4.12. Run of HMW DNA (size range 400-1000 kb) of *E. coli* strains digested with I-*Ceu*I. Run conditions for the size range of 400-1000 kb: switch time of 70-100 s at 6 volts/cm in a 1.2 % agarose gel for 44 hours, with addition of 50 μM thiourea to the Tris-HCl running buffer. Lanes: 1 and 13, *S. cerevisiae* strain YNN295 megabase DNA marker (BIO-RAD); 2, I-*Ceu*I digest of of *E. coli* K12 (MG1655) used as a positive control; 3, I-*Ceu*I digest of strain E102; 4, I-*Ceu*I digest of strain E103; 5, I-*Ceu*I digest of strain E104; 6, I-*Ceu*I digest of strain E105; 7, I-*Ceu*I digest of strain E106; 8, I-*Ceu*I digest of strain E107; 9, I-*Ceu*I digest of strain E108; 10, I-*Ceu*I digest of strain E109; 11, I-*Ceu*I digest of strain E110; 12, I-*Ceu*I digest of strain E111; 14, *S. cerevisiae* strain YPH80 megabase DNA marker (CAMBREX).

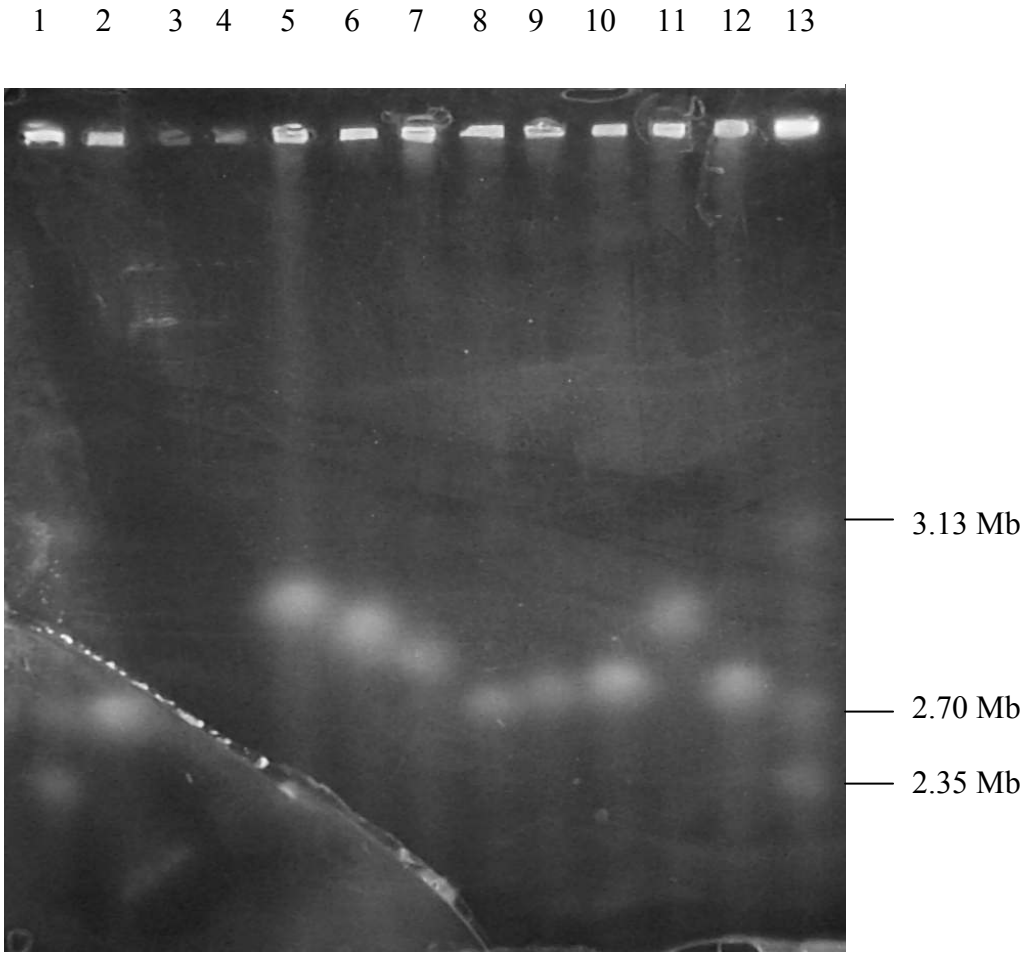


Figure 4.13. Run of HMW DNA (size range 1-3.1 Mb) of *E. coli* strains digested with I-*Ceu*I. Run conditions for the size range of 1-3.1 Mb: switch time of 250-900 s at 3 volts/cm in a 0.8 % agarose gel for 144 hours, with addition of 50 μM thiourea to the Tris-HCl running buffer. Lanes: 1 and 13: *Hansenula wingei* megabase DNA marker (BIO-RAD); 2, I-*Ceu*I digest of of *E. coli* K12 (MG1655) used as a positive control; lane 3, I-*Ceu*I digest of E102; 4, I-*Ceu*I digest of E103; 5, I-*Ceu*I digest of E104; 6, I-*Ceu*I digest of E105; 7, I-*Ceu*I digest of E106; 8, I-*Ceu*I digest of E107; 9, I-*Ceu*I digest of E108; 10, I-*Ceu*I digest of E109; 11, I-*Ceu*I digest of E110; 12, I-*Ceu*I digest of E111. The left bottom corner of the gel was broken during visualization of the gel, however the sizing of the bands was assisted using two DNA markers and no significant difference was found between the measurements using either the marker at lane 1 or 13.

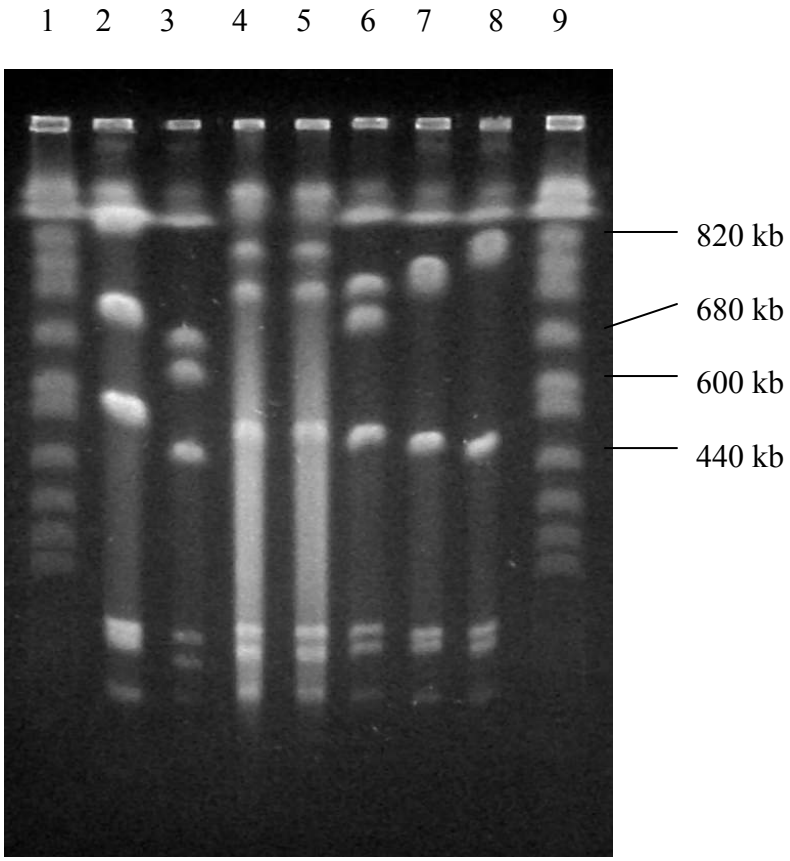


Figure 4.14. Run of HMW DNA (size range 400-1000 kb) of *E. coli* strains digested with I-*Ceu*I. Run conditions for the size range of 400-1000 kb: switch time of 70-100 s at 4 volts/cm in a 1.2 % agarose gel for 30 hours, with HEPES as a running buffer instead of the Tris-HCl. Lanes: 1 and 9, *S. cerevisiae* strain YPH80 megabase DNA marker (CAMBREX); 2, I-*Ceu*I digest of *E. coli* CFT073 used as a positive control; 3, I-*Ceu*I digest of of *E. coli* K12 (MG1655) used as a positive control; 4, I-*Ceu*I digest of E102; 5, I-*Ceu*I digest of E103; 6, I-*Ceu*I digest of E105; 7, I-*Ceu*I digest of E108; 8, I-*Ceu*I digest of E109.

4.3 Discussion

4.3.1. Physical mapping using transposon-carrying I-*Sce*I

The introduction of rare cutter restriction endonucleases in molecular biology had facilitated large scale genome mapping, cloning and sequencing projects. Restriction endonucleases of bacterial origin have recognition sites of up to 8 bp long, and even when their target sites have been methylated the cleavage specificities could only be increased up to 12 bp (Patel *et al.*, 1990). Artificial endonucleases, made by chemical modifications of either DNA binding proteins or synthetic oligodeoxynucleotides have been used to introduce rare cutting sites; however cleavage by such endonucleases occurs at low efficiency (Thierry *et al.*, 1991)

In this work I have described the construction of strains of *E. coli* carrying chromosomal insertion mutations. These insertions introduce a unique restriction site (I-*Sce*I) into the chromosome of these strains and it is proposed to enable a very accurate relative sizing of the bacterial chromosome, as the linearized genome will be gel visualized and measured as a single band. This has the advantage of decreasing the percentage error obtained with endonucleases that cut the genome at different sites. Moreover, using such approach will avoid problems usually associated with sizing multiple bands e.g., formation of doublets and run of minor bands out of the gel.

The DNA partial digestion observed in this study with some of the I-*Sce*I digests has been reported by Jumas-Bilak *et al* (1995). Colleaux *et al* (1986) interpreted this to be due to a protein produced by *E. coli* and that this protein may not be as active as the native mitochondrial protein due to the presence of two leucines instead of two threonines (at positions 123 and 156). Another possible explanation could be that the double strand break observed represents an intermediate in the reaction catalyzed by the protein instead of a final product. On the other hand the DNA degradation observed with some I-*Sce*I digests could be due to the instability of the enzyme in the absence of substrate (I-*Sce*I site) under conditions of activity (Monteilhet *et al.*, 1990). This would happen if an inside-end transposition had occurred and no I-*Sce*I site was introduced into the bacterial chromosome (Figure 4.5b). Jumas-Bilak *et al* (1995) had explained

the DNA degradation associated with I-*Sce*I digests to be the results of longer digestion times (more than 15 minutes) as the endonuclease exhibits non-specific cleavage under such conditions. However, in our experiments the digestion incubation time was kept at about 15 minutes. An interesting point that has been investigated by Colleaux *et al* (1986) is that no additional protein(s) or co-factor of either mitochondrial or nuclear origin is required or involved in determining the specificity of this enzyme.

4.3.2 Genome sizing of *E. coli* BSI-associated chromosomes using intron-encoded endonuclease I-*Ceu*I

Gauthier *et al* (1991) noticed that the growth of *E. coli* was found to stop after transcription induction of the Chloroplast intron CeLSU · 5 ORF encoding the I-*Ceu*I in *E. coli*. This deleterious effect and toxicity was later determined to correspond to the endonuclease activity of I-*Ceu*I *in vivo* and that this protein had actually restricted the *E. coli* chromosome. Further investigation had revealed that restriction sites for the I-*Ceu*I were found to occur at each of the seven copies of the ribosomal operons in the *E. coli* genome (Marshall and Lemieux., 1992). It was concluded (Marshall and Lemieux., 1992) that I-*Ceu*I homing endonucleases could be useful for the analysis of very large genomes and that the endonuclease would be particularly useful to analyse bacterial genomes because it recognizes a sequence that is highly conserved among *rrl* genes making the number of I-*Ceu*I homing sites encountered in bacterial genomes dependent on the copy number of ribosomal DNA operons.

I-*Ceu*I physical mapping by PFGE had more advantages over restriction mapping with other more frequent restriction cutters: Liu *et al* (1993) had found that (i) the I-*Ceu*I sites are not protected from digestion by modification or other means, (ii) it is not found in other parts of the chromosome, and (iii) the I-*Ceu*I site is present in all *rrl* genes in *E. coli*. These last two features had proved very useful for comparing genomes of bacteria as it cuts at no sites other than *rrn* genes. The patterns of fragments from digestions of *E. coli* with other endonucleases are very different; however, after I-*Ceu*I digestion homologous fragments can usually be recognized by inspection of fragment size. This conservation of fragment pattern facilitates comparisons of bacterial genomes.

Moreover, even although the *rrn* operons have several sites for rare-cutting endonucleases; e.g., the *rrnB* gene of *E. coli* has two sites for *Bln*I and one site for each *Xba*I, *Not*I, and I-*Ceu*I cleavage by *Bln*I, *Xba*I, and *Not*I does not occur in all *rrn* genes, because of either mutation or modification of the site (Liu *et al.*, 1993).

Analysis of our I-*Ceu*I fragments revealed that the chromosome sizes of all 10 BSI isolates were in a similar size range as of CFT073 (5.2 Mb) making them larger than MG1655 (4.8 Mb) and that all of them carry seven copies of *rrn* operon as well. The difference in size observed with E107 compared to the other BSI isolates correspond to two extra bands 700 and 730 kb, which are presumed to be two large plasmid DNA fragments (both were observed as faint bands) (Figure 4.15). Data from the literature support that these could be the circular forms of two large plasmids (Kinashi *et al*., 1987; Lennon and DeCicco., 1991; Buchrieser *et al*., 1994). The chromosome size of E107 excluding the two bands related to the presumed plasmid DNA was 4.99 Mb. It is of interest that all 10 BSI isolates resemble CFT073 strain in chromosome size as this might indicate that they have evolved from a common pathogenic ancestor (Hobman *et al*., 2007). Moreover, comparing the genome sizes of these BSI strains with the genome of K-12 (a laboratory attenuated and avirulent *E .coli* strain) had revealed large DNA fragments (~400 kb) associated with the genomes of the 10 BSI isolates. These large DNA fragments might be involved in the pathogenicity of these BSI isolates.

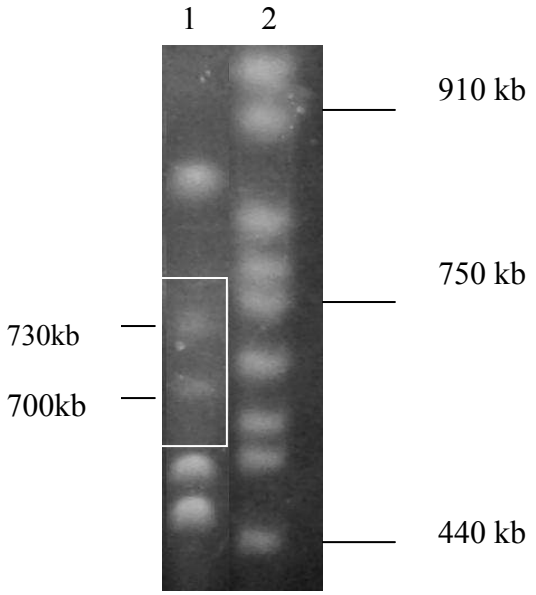


Figure 4.15. PFGE of the plasmid DNA present in E107. The two faint DNA fragments are presumed to refer to two large circular plasmids. The white square highlights the two DNA fragments referring to the plasmids.

Most of the differences observed between strains in band sizes are mainly restricted to the size ranges between 400-1000 kb and the 1-3.1 Mb. These size differences could be due to chromosomal rearrangement and homologous recombination between *rrn* operons. Such events had been reported in *S. enterica* spp. (Kothapalli *et al.*, 2005). The following four classes of chromosome rearrangements might be formed due to recombination between *rrn* operons: deletions, duplications, translocations, and inversions. Deletions of entire I-*Ceu*I fragments would be rare since all the fragments have essential genes. Duplications would be detected by doubled intensity of the duplicated I-*Ceu*I fragments, both E108 and E111 (780 and 790 kb respectively) doublets could be due to a duplication rearrangement event (Kothapalli *et al.*, 2005).

In the study done by Kothapalli *et al* (2005) the inversion and translocation rearrangement events had been reported to occur more frequently than the other two classes. As a requirement to report such events the orientation of these I-*Ceu*I fragments was determined by PCR. They found that all strains tested had at least one translocation or inversion compared with the standard type (the genome order normally found in *Salmonella* and *E. coli*). To investigate the phylogenetic relatedness between the BSI isolates covered in this study a further investigation will require the determination of

the I-*Ceu*I fragments orientations as a priority to obtain more information regarding the inversion and translocation events.

According to the genome balance the lengths of the replichores between the *oriC* and Ter sites on a circular bacterial chromosome must be maintained for balanced bidirectional replication (Figures 4.16 and 4.17) (Kothapalli *et al.*, 2005). Such presumption was discovered by previous studies related to the replication of the chromosome and the locations of *oriC* and Ter; Hill and Gray (1988) had observed that moving *oriC* relative to Ter reduced the growth rate of *E. coli* K-12. Other reports showed that rearrangements between fragments D to F do not change the genome balance because they are all in the same replichore (Figures 4.16 and 4.17) (Kothapalli *et al.*, 2005). Moreover, strains with inversions of I-*Ceu*I-C, which would be highly unbalanced because it contains the origin of replication, were not detected (Kothapalli *et al.*, 2005), and finally the observation of Eisen *et al* (2000) that chromosomal inversions around the origin and termination of replication are symmetrical, thus retaining chromosome balance.

The positions of the origin of chromosomal replication (*oriC*) and the termination of replication were determined from the *E. coli* K-12 *oriC* sequence. The size of replichore 1 in MG1655 (Table 4.7), from *oriC* in I-*Ceu*I-C through fragments D, E, F, G, and A to Ter, is 2,178 kb; the size of replichore 2, from fragment C through fragments B and A to Ter, is 2,462 kb. Genome balance was calculated by dividing the size of each replichore by the total genome size; the off-balance value, which was one-half the difference between the replichore sizes, was 142 kb. To check whether our data support the genome balance hypothesis in the *E. coli* BSI-associated strains, the I-*Ceu*I DNA fragments for the tested strains were reordered according to their sizes and their presumed location and orientation in the genome (Figures 4.16, 4.17 and Table 4.7).

The results from (Table 4.7), suggest that our results support the genome balance hypothesis, however, further investigation is required to confirm the orientation and location of the I-*Ceu*I DNA fragments in the genomes of the tested strains.

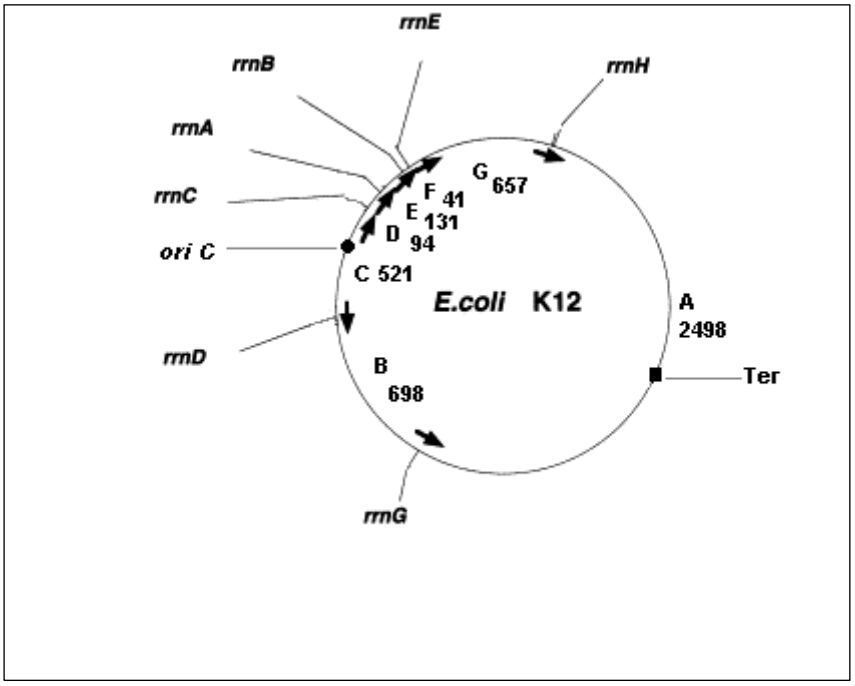


Figure 4.16. Position and direction of rRNA operons on *E. coli* strain K-12 chromosome. Arrows indicate transcriptional direction of rDNA. Reproduced from Shu *et al* (2000).

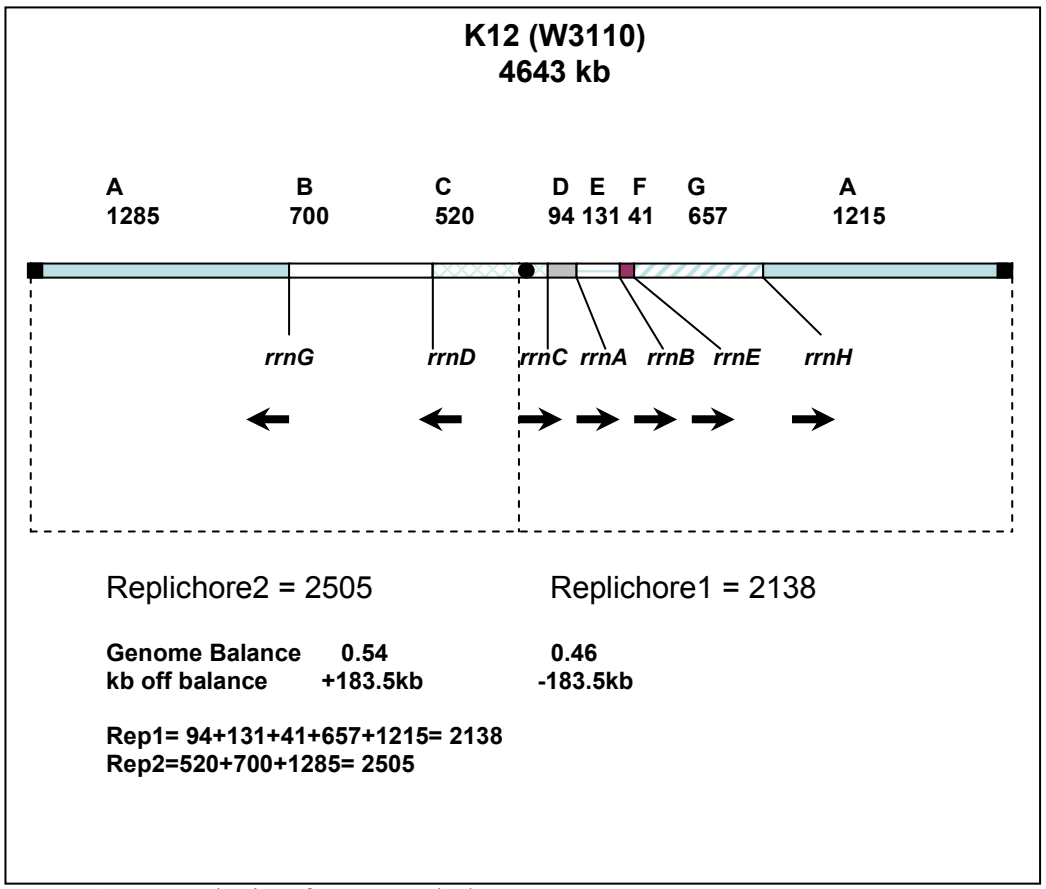


Figure 4.17. Analysis of genome balance. The chromosome is bidirectionally replicated from *ori*C, therefore two replichores are shown. The dot in fragment C represents *oriC*. The black squares in fragment A represents Ter. The cross-hatched line represents calculation of the total fragment sizes to determine the length of replichore 1 and replichore 2. Reproduced from Kothapalli *et al* (2005).

Table 4.7. Genome balance between the two replichores in BSI *E. coli* isolates

Strain[a]	I-*Ceu* I DNA fragments									
	Replichore2			Total Mb [%]	Replichore1					Total Mb [%]
	A	B	C		D	E	F	G	A	
MG1655	1284	657	521	2.462[53]	94	131	41	698	1214	2.178[47]
CFT073	1414	727	590	2.731[52]	118	138	40	866	1337	2.499[48]
E102	1439	738	562	2.739[53]	104	134	44	806	1361	2.449[47]
E103	1439	740	561	2.740[53]	102	133	44	804	1361	2.444[47]
E104	1542	693	564	2.799[52]	96	143	42	798	1458	2.537[48]
E105	1491	686	560	2.737[53]	113	139	43	736	1409	2.440[47]
E106	1439	683	542	2.664[52]	119	137	45	749	1361	2.411[48]
E107	1388	575	537	2.500[50]	105	169	48	860	1312	2.494[50]
E108	1388	782	559	2.729[53]	122	140	47	782	1312	2.403[47]
E109	1439	796	551	2.786[53]	120	141	47	817	1361	2.486[47]
E110	1491	686	575	2.752[52]	121	140	46	794	1409	2.510[48]
E111	1439	786	552	2.777[53]	123	140	49	786	1361	2.459[47]

[a]Sizes in % are indicated between square brackets. For both MG1655 and CFT073, the replichore sizes were obtained from *in silico* data.

Applying different pulse conditions to separate DNA bands for different size ranges had proved useful when sizing DNA fragments between 40-400 kb and 400-1000 kb, as could be interpreted from the small standard deviation (SD) and the percentage error obtained for most of the bands at these size ranges. However, the method still needs to be improved to size bands and to high precision in the size range of 1-3.1 Mb. Furthermore, applying different pulse conditions was used to resolve close doublets bands. Examples include E105 (fragments 700 and 740 kb) and E109 (fragments 800 and 820 kb), these fragments were separated after applying the pulse condition used to separate fragments in the size range of 400-1000 kb. However, separation of the doublet bands of both E108 and E111 at 780 and 790 kb, respectively (observed as large band with doubled intensity compared to other bands on the gel) was unachievable even after applying different pulse conditions to separate them. An alternative approach to resolve the doublets of E108 and E111 would be to gel excise the doublet bands and then to redigest them with one of the less frequently endonuclease cutters (*Bln*I, *Xba*I, *Not*I). If the excised band corresponds to two DNA fragments with similar sizes then the digestion with the less frequently endonucleases (*Bln*I, *Xba*I, *Not*I) should produce bands that sum up to a size approximately equal to the size of the one doublet.

PFGE and rare cutting restriction endonucleases have become very powerful techniques for the study of bacterial genomes. Characterization of bacterial strains by PFGE had allowed for (i) the analysis of the genome organization and gene locus relationships, (Bloch and Rode., 1996; Bloch *et al*., 1996; Blackwood *et al*., 1997; Maurelli *et al*., 1998) (ii) production of fingerprints for population analysis and epidemiologic studies, (Rode *et al*., 1995) and (iii) comparative studies within and between species (Bloch *et al*., 1994; Rode *et al*., 1999). One of the advantages that could come from such studies is to reduce the cost of sequencing projects of various prototypic genomes. Indeed, comparative macro-restriction mapping guided by the genetic map shared by different strains may provide a significant cost saving by identifying the strain-specific DNAs to be sequenced while avoiding shared DNA sequences between these strains. (Melkerson-Watson *et al*., 2000).

4.3.3. DNA degradation associated with E102 and E103

The DNA degradation observed with E102 and E103 is presumed to be similar to a previously characterized phenomenon in *Streptomyces lividans* and *Streptomyces avermitilis*, this is because in both species the DNA degradation can be resolved by HEPES or thiourea treatment. *S. lividans* and *S. avermitilis* have the ability to site specifically modify their DNA *in vivo*. As a consequence of the modification, however, the DNA undergoes site-specific double-strand cleavage during conventional gel electrophoresis and PFGE, producing random fragments most commonly ranging from 40 to 150 kb. This electrophoretic instability was dependent on the activation of Tris buffer (strand cleavage was observed with minimum concentration of 10 mM Tris) at the anode to generate a nucleolytic peracid derivative (oxidising inorganic acid) species. The addition of 5 μM thiourea (radical scavenger and reducing agent) resulted in complete inhibition of DNA strand cleavage during gel electrophoresis, indicating that thiourea had neutralized the active component. The use of an alternative buffer such as HEPES had also produced a non-degradative electrophoresis (Ray *et al*., 1992).

Since then, thiourea and HEPES buffer had been reported as being useful in PFGE typing of other degradation-sensitive species: *Pseudomonas aeruginosa*, *Clostridium*

difficile, Salmonella, Mycobacterium abscessus and enterohemorrhagic *Escherichia coli* non-O157:H7 strains (Corkill and Stubbs., 2000; Römling and Tümmler., 2000; Koort *et al.*, 2002; Zhang *et al.*, 2004).

The post-replicative DNA modification, which produces the DNA degradation, acts site-specifically on closely opposed guanines on either strand. In contrast to the highly reactive $^{\bullet}$OH radicals, which cause DNA damage at every nucleotide, the nucleolytic peracid derivative causes DNA damage specifically at guanines. Guanine has been found to be the most easily oxidized among the four DNA bases because its oxidation potential is lower than that of the other DNA bases (e.g., guanine, 1.29 V; adenine, 1.42 V; cytosine, 1.6 V and thymine, 1.7 V versus normal hydrogen electrode, NHE). Analysis of modified sites revealed that flanking sequences of direct and inverted repeat structures are involved in the process of the *in vivo* modification, and that it could be even influenced by the local DNA topology e.g., DNA supercoiling (Boybek *et al.*, 1998; Habib and Tabata., 2004). Zhou *et al* (2005) had identified the gene cluster (*dnd*) involved in this modification and found it to be localized on an 8 kb DNA fragment. The phenotype was named Dnd (for DNA degradation). Experiments involving the disruption of the *dnd* locus abolish the Dnd phenotype, and gain of the locus conferred it respectively. Extensive analysis of the *dnd* gene cluster revealed five open reading frames, whose hypothetic functions suggested an incorporation of sulphur or a sulphur-containing substance into the genome. The precise chemical nature of the DNA S modification associated with the Dnd phenotype has recently been revealed (Wang *et al.*, 2007). The study by Wang *et al* (2007) showed that the sulphur-containing species is a 5`-d($G_{ps}$A)-3` dinucleotide with a chiral R_p configuration of the phosphorothioate (phosphorus-sulphur) bond.

Similar modification systems were observed in many unrelated bacteria (He *et al*., 2007), one of these was discovered in ETEC strain B7A as *dnd*-bearing fragments located on a GEI downstream of the *leuX* tRNA site (He *et al*.,2007; Ou *et al*.,2007). An interesting finding by our study is that an integrase gene was discovered downstream of *leuX* in strains E102 and E103 by the tRIP-PCR strategy described earlier in chapter 3. A further characterization of the E102_*leuX* and E103_ *leuX* GEIs

might reveal a similar *dnd* clusters to that found in strain B7A. This and the findings of Zhou *et al* (2004) and He *et al* (2007) that the *dnd* gene cluster is carried on a genomic island (GEI size 93kb) in *S. lividans* 66, strongly supports that this system could be mobile among diverse organisms. As a continuation to this work, the Dnd phenotype observed in both E102 and E103 is being further investigated by Xinyi He from Bio-X Life Science Research Centre and School of Life Science and Biotechnology, Shanghai Jiaotong University, China. The investigation aims to verify if a similar modification system to that of *S. lividans* 66 exists in these two strains. This is done by amplifying segments of the *dnd* gene cluster. Preliminary results indicated that the *dnd* gene exists in both strains.

Finally, in this study we have validated the use of two strategies. The first is sizing the bacterial genomes using I-*Ceu*I, which restricts the bacterial chromosomes at the *rrn* operons producing 5-8 DNA fragments. We speculated that sizing such few DNA fragments would be more reliable than sizing multiple bands produced by other endonucleases that cut more frequently in the genome. The second strategy is to use different PFGE run conditions to measure DNA fragments from different size ranges. Three different size ranges (40-400 kb), (400-1000 kb) and (1-3.1 Mb) were selected. Both strateges are proposed to reduce the standard deviation (SD) and the percentage error of the measured DNA fragments and to achieve more accurate genome sizing.

The obtained results indicated high accuracy of measurements for the I-*Ceu*I DNA fragments as presented by the low percentage error which was less than –5% and the total SD (0.003). To our knowledge, no previous studies have used such strategies to achieve accurate sizing of the bacterial genomes. For example, Thong *et al* (1997) had used PFGE to assess the extent of genome variation in bacterial genome sizes by applying restriction endonucleases that cut relatively more frequently than I-*Ceu*I. Comparing our data to that presented by Thong *et al* (1997) highlights the significance of our strategy in sizing bacterial chromosomes using PFGE. For example, the SD obtained by our approach (0.003) is less than that obtained by Thong *et al* (1997) SD 6. Moreover, in Thong *et al* (1997) study no control samples were used to calculate the percentage error.

Chapter 5 Estimating the sizes of microarray-visualized genomes of BSI-associated *E. coli* strains

5.1. Introduction

Although whole-genome sequencing is a powerful method for genome analysis, it is still laborious and expensive. Recently, comparative genomic hybridization (CGH) has been used to facilitate comparisons of unsequenced bacterial genomes in order to look for characteristic genes or chromosomal regions related to unique phenotypes (Salama *et al*., 2000; Dobrindt *et al*., 2003; Fukiya *et al.*, 2004; Rajashekara *et al*., 2004; Wick *et al*., 2005). Unlike other typing methods such as multi-locus enzyme electrophoresis, random amplification of polymorphic DNA and pulsed-field gel electrophoresis, which do not allow direct comparison between the gene and / or DNA sequence content of bacterial genomes, comparative genomic indexing (CGI) (Anjum *et al.*, 2003) was very useful in such studies. Anjum *et al* had expected that CGI would allow the definition of the core genes common to pathogenic strains and the commensal MG1655 and also identify regions of differences between these strains. In this study, *ShE.coli* meta-array probe sequences (supplied by the Molecular Microbiology Group, Institute of Food Research, Norwich Research Park, UK) were used as the baseline for determining the genomic contents of the *E. coli* BSI isolates in a CGI microarray. The microarray represented 6239 CDS from *E. coli* K-12 MG1655, *E. coli* EDL933, *S. flexneri* 2a Sf301 and virulence-associated *E. coli* CDS derived from strains representative of different pathotypes of *E. coli* (Ou *et al.*, 2005). The use of multiple strain-specific sequences in the CGI microarray has the advantage of increasing the gene reporters it contains and making it more representative of the species (Witney *et al.*, 2005).

CGI studies are limited by the composition of the microarray, meaning that it is unable to detect any genes specific to the test strain that are not present in the reference strain and therefore not represented on the array. However, the use of multiple strain-specific sequences provides extensive coverage of a particular species and would enable a two-way comparison to the reference strain that not only identifies deleted/divergent genes in the test strain but also detects the presence of genes specific to the test strain that are

absent in the particular reference strain but present on other genomes also represented on the array (Witney *et al.*, 2005). The use of *ShE.coli* meta-array and other similar approaches increases the scope of the genes represented on the array, making it more informative and comprehensive.

Other similar approaches have included the use of arrays representative of five *E. coli* strains, two K12 strains and three ECOR collection (Ochman *et al.*, 2000), and seven sequenced (MG1655, W3110, O157:H7 EDL933, O157:H7, CFT073, 042 and E2348/69) *E. coli* genomes in addition to several sequenced *E. coli* plasmids, bacteriophages, pathogenicity islands and virulence genes (Willenbrock *et al.*, 2006).

Following normalization of the microarray data, the output files were extracted. Then, different methods for threshold estimation were applied to infer the presence/absence of genes. The output analysis of these steps was used as an input to ArrayOme, which is a program that estimates the total size of the portions of genomes represented by microarray-based probes. The program uses protein-coding sequences (CDS) that are contiguous on annotated reference templates and classified as 'Present' in the test strain by hybridization to microarrays. CDS, are merged into inferred contigs (ICs), which are then extended to account for flanking intergenic sequences. Finally, the lengths of all extended ICs are summated to yield the 'microarray-visualized genome (MVG)' size. ArrayOme permits rapid recognition of discordances between PFGE-measured genome and MVG sizes, thereby enabling high-throughput identification of strains rich in novel genes. This approach was termed Microarray-Assisted mobilome Prospecting (MAmP) (Ou *et al.*, 2005). The MAmP approach combines CGI, ArrayOme and pulsed-field gel electrophoresis (PFGE) to predict the size of the novel, non-microarray-represented mobilome in a test strain (Ou *et al.*, 2005). Ochman and Jones (2000) had estimated that the amount of unique DNA present in four *E. coli* strains investigated ranged from 65 to 1183 kb based on differences in chromosome length and gene content relative to *E. coli* K-12 MG1655 (Ou *et al.*, 2005).

5.2. Results

5.2.1. Microarray data analysis

The processed slides were scanned with a GenePix 4200A scanner (Axon Instruments, Inc) and the data were quantified and analysed by the Bluefuse software version 3.2 (BlueGnome, Cambridge, UK). Spots were excluded when the probability presence of a biological signal pON in both channels was less than 0.5 to eliminate the bias generated by the inclusion of unhybridized spots in the statistical interpretation of the data (Overton *et al.*, 2006; Snyder and Saunders., 2006; Whitehead *et al.*, 2007) or when the standard deviation of the replicates was > 0.5 as this represented poor reproducibility between replicates (default parameter of Bluefuse). When points with high measurement errors were removed, the residuals of the log-log plots for replicates appeared to follow a normal distribution. This suggested that the other component of uncertainty in the ratio is a proportional error contribution (Figure 5.1) (Ideker *et al.*, 2000; Rocke and Durbin., 2001; Goryachev *et al.*, 2001).

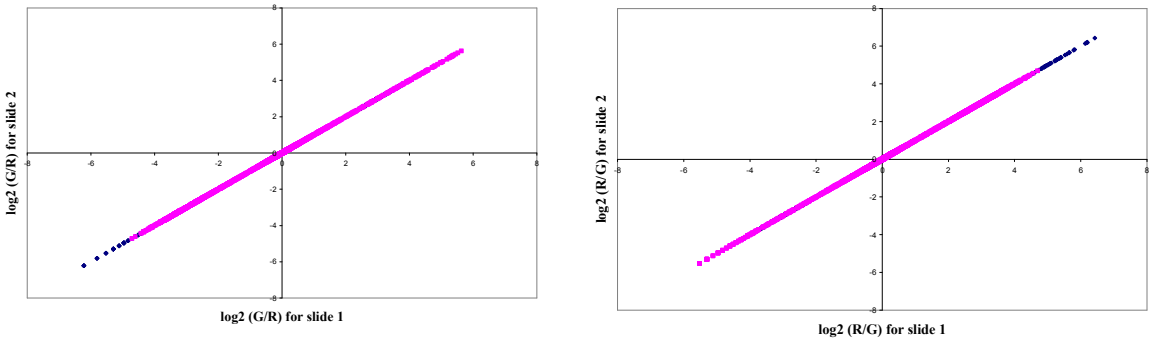


Figure 5.1. Log-ratio vs. log-ratio plots of replicates in E107. Log-ratio vs. log-ratio plots of replicates should be uniformly distributed around the regression line (the errors should exhibit a uniform distribution). Same data distribution for log-ratio vs log-ratio plots of replicates were observed with the rest of the tested strains.

The pON is generated during the quantification process to estimate whether the probe-signal in each channel is genuine (Overton *et al.*, 2006 and BlueFuse manual, 2006). The pON score evaluated the data from each microarray spot and reported a hybridization signal for it. Unlike other ratio-metric methods, this score is not dependent upon the signal in the other channel. A pON score of zero indicated no

hybridization to the spot, while a score of one indicated strong evidence of spot hybridization.

The quantitation of pON score is calculated as an outcome of two features: the signal intensity above background, and the circularity and uniformity of a microarray spot. The hybridization signal is determined to be the proportion of pixels within it that are statistically inconsistent with the Bayesian calculated background noise. Characteristics of the Spot features such as being round and the uniformity of the signal are considered when quantifying the signal intensity as the signal score could be influenced by inconsistent background and other features that are not hybridized microarray spots. This allowed for discrimination between signals that were large and with high intensity noise (Snyder and Saunders., 2006).

The data were normalized by Global LOWESS (locally weighted linear regression) normalization and then replicates were fused. In Figures 5.2 and 5.3 the effects of the normalization steps were observed by the transformation of both the ratio 50^{th} percentile and the mode of the ratio natural log [In] for the data were around 1 and 0 respectively. This is because most of the genes are part of the core genome and therefore are presumed to have equal intensity values in both channels; the reference or the comparator channel, which is composed of an equal genome mixtures of two strains EDL933 and MG1655 (see materials and methods for details) and the other channel is for the test strains. Consequently, a log of (Ch1/Ch2) which equals to log (Ch1) – log (Ch2) for the core genes would be expected to be around or equal to zero (the same will apply for log (Ch2/Ch1)).

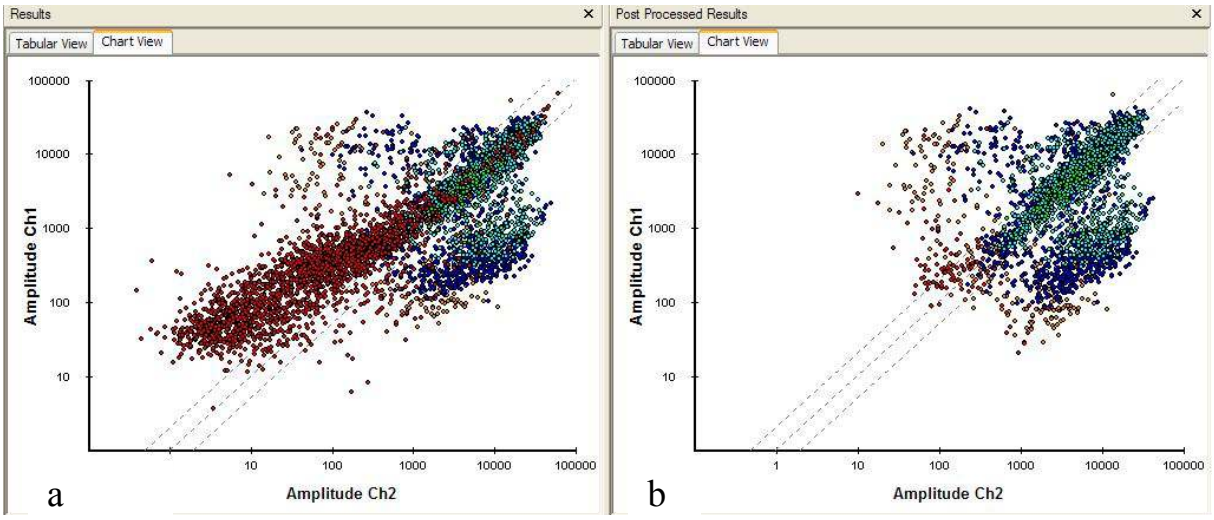


Figure 5.2. Effects of normalization. The effects of normalization are observed by the disappearance of the banana shape. (a) The banana shape (skew of data) obtained when the amplitude of channel one is plotted against channel two before normalization. (b) The disappearance of the banana shape after data normalization. Genes shifted towards Ch1 are present in channel one only and genes shifted towards Ch2 are present in channel two only. While genes present in both channels are represented between the two channels.

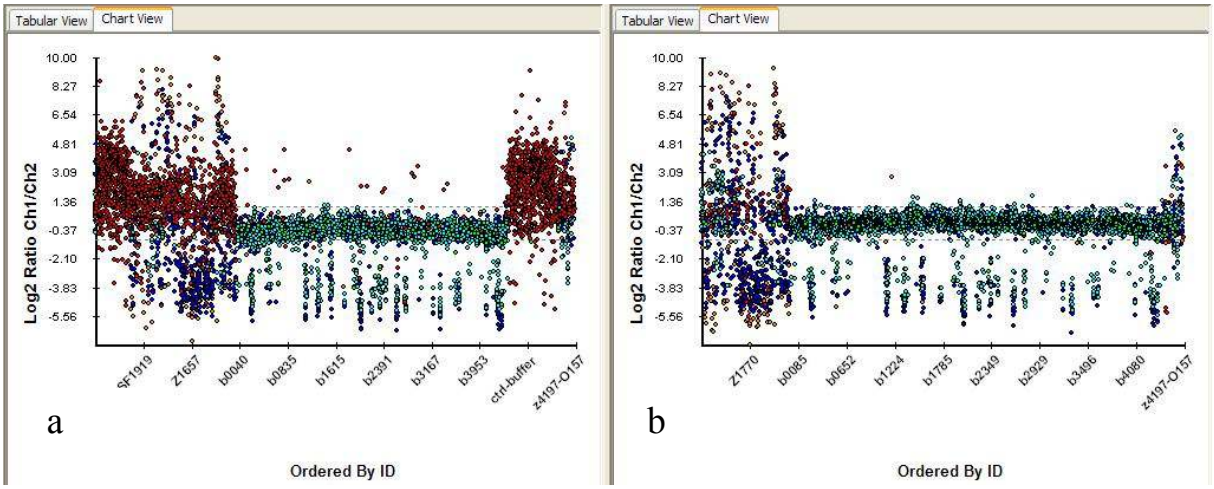


Figure 5.3. Effects of normalization. (a) When Log2 ratio was plotted versus the microarray CDS the data was either shifted above or below zero (before normalization). (b) after the data being normalized the log2 ratio was tend to be around zero for core genome genes which are present in both strains, while genes specific to Ch1 have log2 values above zero and genes specific to Ch2 have log2 values below zero.

5.2.2. Validation of the *ShE.coli* microarray

Two methods were previously described to validate the microarray results. In the first method described by Witney *et al* (2005) an arbitrary cutoff of twofold was used to identify genes that were specific to one of the strains. The upper cutoff was set at a ratio of 2 and the lower cutoff at a ratio of 0.5. Genes with an intensity ratio (test strain / reference strain) greater than the upper cutoff were deemed to be specific to the test strain, genes with a ratio less than the lower cutoff were deemed to be specific to the reference strain, and genes with ratios between 0.5 and 2 were deemed to be present in both strains. In the second method, Anjum *et al* (2003) used the natural log [In] of the intensity ratio (reference strain / test strain) with a cutoff value equal to or greater than 2 to define the presence of this CDS in the reference strain but its absence in the test strain. Accordingly, any CDS with [In] (reference strain / test strain) value of less than 2 would indicate the presence of this CDS in the test strain, regardless of which dye had been used for labelling. Both methods were used in this study to define the presence / absence of CDS in the test strains.

The two methods shared common features related to their specificity and sensitivity that made them useful when comparing microarray data. Both had been shown to be highly specific in reducing the number of false positives and false negatives (both the positive predictive value (PPV) and the negative predictive value (NPV) are high) (Anjum *et al.*, 2003; Witney *et al.*, 2005). Witney *et al* (2005) had defined PPV to be "the proportion of genes unique by microarray that are truly unique by sequence prediction" and it is calculated as the number of unique genes by both sequence prediction and microarray divided by the unique genes by microarray. Thus, subtracting the PPV from 100 would indicate the level of false positive. On the other hand, NPV was defined as "the proportion of genes called not unique by microarray that are truly not unique by sequence prediction" (Witney *et al.*, 2005). It is calculated as the number of genes not unique by both sequence prediction and microarray divided by the genes not unique by microarray. Subtracting the NPV from 100 would give the level of false negatives (Witney *et al.*, 2005). However, both methods were unable to detect divergent genes, which may have been present but did not hybridize and therefore are regarded as absent in the test strain. Other methods such as GACK software which uses microarray data to

classify genes into present or absent/divergent and the 3SD method, which determines variable cutoffs for each strain, gave greater sensitivity at the risk of a higher number of false positives (Witney *et al.*, 2005).

5.2.3. Estimating the microarray-visualized genome (MVG) size using the *She.coli* meta-array data

The *ShE.coli* (PCR amplicon-based) metagenome microarray slides used in this study were supplied by the Molecular Microbiology Group, Institute of Food Research (IFR), Norwich Research Park, UK. The construction of the microarray slides were performed by the Molecular Microbiology group in the IFR: The CDS were amplified with specific primer pairs (Sigma-Genosys). Then, DNA from the PCRs was resuspended in a spotting solution containing 50% dimethyl sulfoxide and 0.3× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate). PCR products were spotted onto gamma amino propylsilane-coated GAPS II slides (Corning) by using a Stanford MarkII arrayer. The DNA was UV cross-linked to the slides by using a Stratalinker (Stratagene) at 300 mJ. Slides were washed in a 95ºC water bath for 2 minutes fallowed by 95% ethanol wash for 1 minute and dried by centrifugation at 185 ×g prior its storage at room temperature (Anjum *et al*., 2003; Lucchini *et al*., 2005).

The meta-array probe sequences represented 6239 CDS comprising 4264 *E. coli* K-12 MG1655 CDS, 1101 *E. coli* EDL933 CDS, 516 *S. flexneri* 2a Sf301 CDS and a further 358 virulence-associated *E. coli* CDS derived from strains representative of different pathotypes of *E. coli*, particularly enteropathogenic *E. coli* (EPEC) and enterotoxigenic *E. coli* (ETEC). Of the 358 virulence-associated CDS, 132 were located on *E. coli* chromosomes while the remaining 226 mapped onto a range of plasmids based on published literature and GenBank submissions (Ou *et al.*, 2005).

The classified CDS (present/absent) above were then used as inputs into the on-line application ArrayOme (http://mml.sjtu.edu.cn/MobilomeFINDER/) that calculated the size of genomes based on the data from the *ShE.coli* microarray (Ou *et al.*, 2005 and Ou *et al.*, 2007). Table 5.1 represents the results of the ArrayOme outputs. The MVG sizes were compared to the sizes obtained by the PFGE (Table 4.5) and the discrepancy

between the two sizes was used as an estimation of the size of novel mobile genomes (mobilomes) specific to the test strain. The novel mobilome size was estimated to be between 36-352 kb and the non_MG1655 genome between 652-1308 kb.

Table 5.1. The ArrayOme-predicted MVG sizes of *E. coli* strains based on the data derived from the *ShE.coli* microarray.

Strain[a]	No. of CDS classified as 'Present'						Size of the *ShE.coli*-MVG [MG1655-MVG] (kb)[b]	Discrepancy between the sizes of PFGE and *ShE.coli*-MVG (kb)	Size of the non-MG1655 genome (kb)[c]
	Size of PFGE (kb)[d]	No. of MG1655 CDS (n = 4264 + 25)	No. of EDL933 specific CDS (n = 1101)	Sf301 specific CDS (n = 516)	Other chromosomal CDS (n = 132)	Plasmids borne CDS (n = 226)			
K-12 MG1655	4852	4288	58	59	17	8	4771 [4639]	81	–
E107M1	4994	3793	338	84	84	145	4958 [4162]	36	[832]
E107M2		3892	417	95	90	155	5152 [4245]	–158	[749]
E108M1	5132	3586	389	103	86	145	4801 [3943]	331	[1189]
E108M2		3742	479	113	88	153	5059 [4079]	73	[1053]
E109M1	5272	3606	436	131	94	145	4920 [3964]	352	1308
E109M2		3877	559	146	101	155	5345 [4212]	–73	1060
E110M1	5262	3583	481	128	87	138	4942 [3957]	320	1305
E110M2		3821	561	135	91	152	5272 [4174]	–10	1088
E111M1	5236	3728	540	93	75	124	5035 [4047]	201	[1189]
E111M2		4239	627	112	86	144	5722 [4584]	–486	[652]

[a] A total of 4264 *E. coli* K-12 MG1655 CDS, 1101 *E. coli* EDL933CDS, 516 *S. flexneri* 2a Sf301CDS, 132 other *E. coli* chromosomal virulence-associated CDS and 226 plasmid-borne *E.coli* virulence genes are classified as 'Present' or 'Absent' based on method 1 (twofold cutoff) or method 2 (In<2).

[b]The data corresponding to plasmid-borne genes was considered when calculating MVG sizes, that is because, the PFGE genome size was presumed to include any episomal entities and subtracting the PFGE size from the MVG size would exclude these CDS related to plasmid-borne genes from the size of the novel genome size. The ArrayOme-predicted sizes of *ShE.coli*-MVGs (left) and MG1655-MVGs (right, square brackets) are shown to the nearest kilobase (kb). The MVG sizes represent the mean of duplicates with average SD of 115 between duplicates.

[c]For strain E107 the non-MG1655 genome size between square brackets was obtained by subtracting the PFGE size after excluding the two fragments that refer to the plasmid DNA. For strains E108 and E111 the non-MG1655 genome size between square brackets was obtained by subtracting the PFGE size after adding the duplicate band - MG1655 MVG size. It is important to note that subtracting the PFGE genomes sizes for the BSI-associated *E. coli* strains from the PFGE size of MG1655 would not result in obtaining the correct size of the non-MG1655 genome. This is because direct subtraction of the PFGE sizes obtained for the BSI-associated *E. coli* strains from the PFGE genome size of MG1655 will account for only some of the CDS present in the tested strains and absent in MG1655. For example, subtracting the PFGE genome size of E107 (4994 kb) from the PFGE genome size of MG1655 (4852 kb) would result in non-MG1655 genome size of 142 kb. On the other hand, subtracting the PFGE genome size obtained for E107 (4994 kb) from the [MG1655-MVG] (4162 kb), which represents the total size of the MG1655 CDS spotted on the microarray and present in E107, the non-MG1655 genome size would be (832 kb).

[d]The lengths of the complete genomes, shown to the nearest Kilobase (kb), are based on PFGE results. For strain E107 the genome size refers to the PFGE size after excluding the two fragments that refer to the plasmid DNA. For strains E108 and E111 the genome size refers to the PFGE size after adding the duplicate band 782 and 786 kb to the genome size respectively.

5.3. Discussion

5.3.1. Estimating the MVG and the novel, non-microarray mobilome sizes using the *She.coli* meta-array data

The high reproducibility of Microarray and its accurate prediction of gene differences between strains have made it more feasible to investigate the horizontal acquisition or loss of mobile genetic elements MGEs. The *ShE.coli* meta-array was designed to produce a nonredundant microarray as it represents each gene with one copy only. This would minimize the associated problems with other microarray methods which use ratios of chromosomal DNA to generate a control signal for every spot which could lead to increased signal intensity of core genes in the reference channel and therefore, complicating subsequent data analysis. As with the majority of CGH studies the *ShE.coli* microarray was an amplicon-based microarray and not oligonucleotide-based array, which is associated with more erroneous gene absence calls. This is because the probe length is shorter in oligonucleotide-based arrays than in PCR-amplicon based arrays, and the oligonucleotide-based arrays are generally designed to represent a whole gene family which makes it perfect in detecting genes resembling the known database for a particular gene but rather unable to detect new polymorphisms that exist in nature (Denef *et al*., 2003; Daran-Lapujade *et al*., 2003; Taboada *et al.*, 2005; Ou *et al.*, 2005; Witney *et al.*, 2005).

The pON > 0.5 score had been selected in order to identify the hybridization of probes to divergent alleles and orthologous genes of different functions and at the same time maintain a reliable threshold to decrease the detection of false positives. However, the pON score had no account of any relative intensity levels, so it could call a gene present in two strains if the pON threshold was exceeded in both, even if the relative intensity level was significantly higher in one strain. To counteract these effects and decrease possible false positives produced from low signal intensity above background, the data were post-normalized and the genes were reassessed as present/absent using more conservative methods. These conservative methods would reduce false positive results introduced by highly divergent genes. This processing also counteracted many aspects related to using pON metric threshold such as the impact of the photomultiplier tube

(PMT) gain settings during scanning on the number of genes passing the pON threshold and thus being assigned as present (Stabler and Hinds., 2006). The two methods selected to assist the presence/absence of genes (natural log (In) <2 and the twofold cutoff method) were considered to be more conservative than other methods such as the GACK or 3SD, which will give greater sensitivity but at the risk of a higher number of false positives (Anjum *et al.*, 2003; Witney *et al.*, 2005). However, from the results of MVG genome sizing (Table 5.1) a mean difference of 379 kb is observed between the two methods. This difference could be attributed to the more relaxed threshold applied by the In <2 method calling for more genes as present in the test strain. This is confirmed when the reverse of the natural log of 2 is calculated. The inverse (antilogarithm) of In (2) is ~7.4 and comparing the two applied thresholds (Figure 5.4) shows that the twofold cutoff method was more conservative than the In<2 method in calling for the presence of genes. Possible reasons for the increased calling of present genes in the In<2 method could be referred to the way this threshold was selected (Anjum *et al.*, 2003). The In<2 threshold was selected to detect deletions in mutant derivatives of MG1655 in a microarray that uses MG1655 genome only as the baseline for the microarray construction while the two-fold cutoff method had been used and tested to verify the presence/absence of genes in multistrains based microarrays and so overcome limitations associated with thresholds based on one strain microarray (Witney *et al.*, 2005).

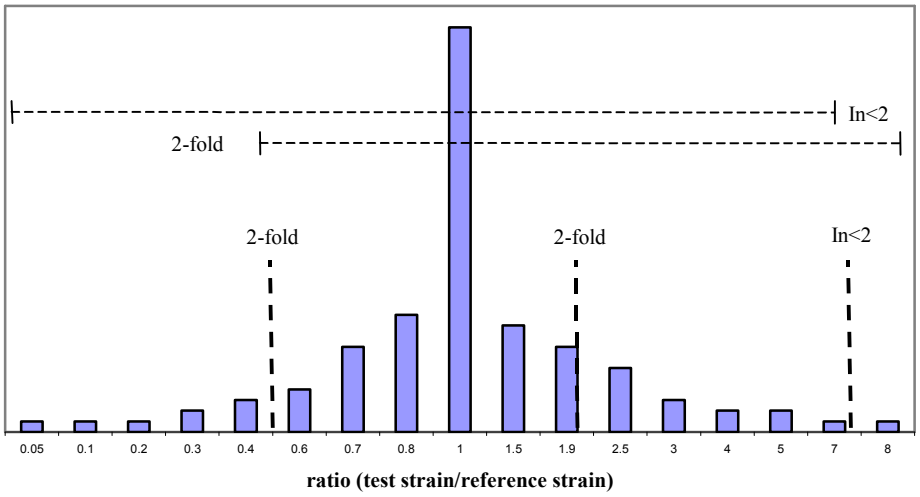


Figure 5.4. Frequency distribution of a representative data. The cutoffs used for the two methods of analysis are represented with vertical dotted lines. The horizontal dotted lines above the graph represent the distribution of genes identified as present by each method of analysis.

The normalized and classified data (present/absent) were then introduced into ArrayOme to estimate the size of MVG. Size estimations obtained by ArrayOme were previously evaluated and found to be within ~2% of accuracy (Ou *et al.*, 2005). ArrayOme had used the annotated chromosomal sequences of MG1655, *E. coli* EDL933 and *S. flexneri* Sf301 as reference templates for creating ICs. ArrayOme generated these linear ICs, using as the reference template a single genome or GenBank sequence for each IC. Each probe was mapped onto a single CDS, which in turn was mapped onto a single template only. The *ShE.coli* microarray results obtained after analysing genome content of the *E. coli* BSI isolates by ArrayOme showed that an average of 4618 CDS were deduced to be 'Present', which comprises an average of 3787 MG1655 CDS, 483 EDL933 CDS, 114 Sf301 CDS, 88 other *E. coli* chromosomal virulence-associated CDS, 146 plasmid-borne virulence genes and an average novel, non-microarray-represented mobilome of 219 kb. This last number corresponded with observations made by other groups: Ochman *et al* (2000) estimated the amount of unique DNA present in four *E. coli* strains to be between 65 to 1183 kb based on differences in chromosome length and gene content relative to *E. coli* K-12 MG1655, and Bergthorsson *et al* (1998) had found that the chromosome sizes of natural *E. coli* strains

had varied by as much as 1 Mb and were ranged from 4.5 to 5.5 Mb. Based on these limits, Ou *et al* (2005) had speculated that as much as 1.8 Mb (5.5-3.7 = 1.8 Mb) of the genome of one of these *E. coli* isolates (strain ECOR 37) may differ from that of *E. coli* K-12 and that this mobilome alone could harbour up to 1800 non-MG1655 CDS.

5.3.2. Genes Identified by the *ShE.coli* microarray

Analysing the distribution of the CDS identified by the ArrayOme in Table 5.1 and comparing it with previously identified CDS (using the same software) in other *E. coli* and *Shigella* strains (Ou *et al*., 2005) have revealed some important information regarding the distribution of these genes in the tested strains. It could be concluded that more of the virulence and plasmid borne CDS are present in these BSI-associated strains than in other *E. coli* and *Shigella* strains (Ou *et al*., 2005). Such CDS included sequences from the *E. coli* K-12 seven cryptic prophages, *CP4-6*, *DLP12*, *e14*, *Rac*, *Qin*, *CP4-44*, and *CP4-57*. Genes in the cryptic prophage regions included DNA integrases, invertases, and recombinases (e.g., *ybcK*, *b1345*, *b1374*, and *b1545*). Other laterally acquired elements such as the insertion sequence IS*1* genes, *insA* and *insB*, and genes of the Rhs elements, *rhsD* and *rhsE*, were also identified. Moreover, the plasmid-borne ipa-mxi-spa locus which is known to be essential for *Shigella* entry into both epithelial cells and macrophages was or part of it was identified in the following *E. coli* strains: E107, E108, E109, E110 and E111. The regulatory mechanism of the ipa-mxi-spa locus and the expression of the virulence phenotype of *S. flexneri* involve a complex process, in which two plasmid-borne genes, *virF* and *virB*, encode essential regulatory proteins. The VirF protein, an AraC-like transcription factor, activates *virB* and *icsA/virG*. The VirB protein would bind and activate the promoters of the entry genes (Lucchini *et al*., 2005).

A previous analysis of the core genes represented by the *ShE.coli* microarray had classified the core genes into the following functional groups (Anjum *et al*., 2003): cell division and chromosome partitioning; coenzyme metabolism; energy production and conversion; nucleotide transport and metabolism; posttranslational modification; protein turnover and chaperones; and translation, ribosome structure, and biogenesis. Screening of the (*ShE.coli* microarray) core genes present in the *E. coli* strains of our

study had showed that genes associated with metabolism, various cellular processes, and information storage and processing have been conserved and maintained in all tested strains.

Chapter 6 Final conclusions

In this study we have tested and proved that the tRIP-PCR and the SGSP-PCR methods are simple and efficient ways to identify tRNA sites occupied with GEIs. The identities of these GEIs were verified from their extremities compared with previously identified GEIs. However, the internal spans of these GEIs may differ considerably from previously identified GEIs. This is because each genomic island has its unique mosaic structure that expresses its repository of integration events and differentiates it from other similar GEIs. The shared ends identified in this study may define a biological and/or evolutionary relationship between GEI families and could facilitate a better understanding of the association between some GEIs and phages e.g., phages could be used by GEIs as a vector for transmission or as a mechanism for efficient gene/DNA sequence shuffling and subsequently a way for the generation of genome diversity. Two more strategies have been used to verify the mobile genome contents present in the studied 10 BSI-associated *E. coli* isolates, these were the physical genome sizing by the PFGE and the CGH approach.

Based on the study results, it is suggested that physical genome sizing using PFGE combined with the use of rare restriction enzymes such as I-*Ceu*I and the use of different PFGE run conditions to measure DNA fragments from different size ranges had reduced the percentage error (<5%) of the measured DNA fragments. Therefore, the use of such approaches may increase the accuracy of genome sizing. We think that more accurate genome sizing could be achievable using the same approaches mentioned above but more stringent conditions are to be imposed e.g., the use of Tn*5*Map I-*Sce*I site that linearized the bacterial genome and allow for sizing one genomic band that represents the whole genome. Although, such approaches would need a lot of optimization steps to achieve desired targets. We infer that accurate genome sizing alone could be used to demonstrate true and significant genome diversity within species and to define mobilome rich areas by comparing the sizes of genomes within the same species. Identification of such rich areas would incur for further investigations and future works.

In regards to the obtained CGH data in this study, we concluded that applying such approach could be used to indicate evidence of absolute gene differences as well as an indication of marked polymorphisms in the DNA sequences between the tested strains and reference strain. This of course could be determined by the applied thresholds used to indicate the presence/absence of genes. Therefore, the application of more complex algorithms and thresholds that utilize statistical methods will consequently impact on the quality and utility of comparative genomic hybridization CGH and array-transcriptomic data. Moreover, the use of larger numbers of sequenced genomes in such studies e.g., *E. coli* and *Salmonella* with 17 available sequenced genomes in the database up to date, could lead to more robust interpretive criteria. Such whole-genome based methods would determine the repertoire of virulence genes found in bacterial pathogens and therefore, are considered more realistic than other approaches like the whole–genome sequencing of multiple strains representing the same species (Lucchini *et al*., 2001). However, CDS microarray is unable to detect single nucleotide changes responsible for protein polymorphism and allelic variation, therefore, high-density oligonucleotide microarrays could be used to detect mutant alleles or single nucleotide polymorphisms (SNPs), as they carry shorter targets (25 nt on Affymetrix gene chips). In such microarrays, a single nucleotide difference is considered sufficient to prevent hybridization between target and probe (Lucchini *et al*., 2001).

An interesting finding of this study is that 46 GEIs identified by the sequential PCR strategy tRIP and SGSP-PCR were found to resemble CFT073-like entities. Moreover, the physical genome size obtained by the PFGE for these strains was close to the physical genome size for CFT073. Further investigations of these CFT073-like entities might reveal an important evolutionary role played by both ExPEC groups and/or that both pathogroups have descent from a common ExPEC ancestor. The fact that strain CFT073 was isolated from the blood and urine of a woman with acute pyelonephritis might explain the high number of the CFT073-like entities identified in our study (Kao *et al*., 1997).

Similar associations between different ExPEC have been reported previously. For example, the recently sequenced ExPEC APEC O1 has been found to share nucleotide sequence with UTI89, 536 and CFT073 (Johnson *et al*., 2007). APEC O1 is an avian pathogenic *E. coli* responsible for colibacillosis in poultry (Johnson *et al*., 2007). Johnson *et al* (2007) found that 87.1% of the APEC O1 ORFs were part of the common ExPEC backbone (K-12 like sequences plus common ExPEC sequences), and that 9% of these ORFs, were specific to all sequenced ExPEC strains only. In this context, an interesting finding of our tRIP and SGSP-PCR strategy is that the boundaries of some of the GEIs discovered in this study resemble APEC O1 GEI sequences e.g., the islet of E106_*selC*, E106_*serW*, E108_*serW* and E111_*serW*. Further investigation for the sequence of these GEIs might reveal more information regarding the virulence of ExPEC strains. Moreover, interrogation of these sites might provide evidence that human and avian ExPEC strains are highly similar to each other, and that a possible food-borne link exists between some APEC and UPEC strains. The origin of most UPEC strains is widely accepted to be from the colonic flora of affected individuals; however, there is no consensus as to the source of these UTI strains colonizing the gut. One possible route is the oral-faecal. Poultry might play an important role as vehicle for the transmission of avian *E. coli* from poultry to human (Rodriguez-Siek *et al*., 2005). Such view is supported by previous studies, for example, Moulin-Schouleur *et al* (2007) had found that no host specificity is observed between avian and human strains belonging to the same ECOR group and that both pathogenic strains are highly and similarly virulent for chicken. This is probably explained by the ability of both UPEC and EPEC to adapt for an extraintestinal lifestyle. Therefore, APEC strains are considered one of the most important candidates that might cause extraintestinal disease in human beings (Rodriguez-Siek *et al*., 2005). One possible way to further investigate such association between ExPEC strains is by including more genes representing the genome of UPEC and avian pathogenic *E. coli* in the *ShE.coli* microarray to make it more comprehensive. Such microarray would represent the genome of ExPEC strains and would cover more of the virulence genes associated with extraintestinal diseases.

As mentioned before, many of the covered aspects and findings in this study are subject for further investigations. For instance, the full-length sequence of 3 GEIs with novel

DNA sequences are currently being investigated by using a capture vector containing two conserved targeting sequences flanking each end of the region of interest (GEI) within the genome. On the other hand, the DNA degradation obtained with strains E102 and E103 are being further investigated to verify if a similar Dnd system exists to that discovered in *S. lividans* and *S. avermitilis*. These investigations could lead to a better understanding of the *E. coli* pan-genome. The current status of the *E. coli* pan-genome indicates that the number of the new strain specific genes identified in the *E. coli* pan-genome with every newly sequenced strain is in decline. However, 441 new genes are being added to the core genome (2865 genes) with each newly sequenced strain and therefore, the *E. coli* genome was considered to be an open pan-genome (Muzzi *et al*., 2007). For that reason, it is tempting to obtain more data concerning the *E. coli* genome structure of the tested strains in this study and to expand the number of tested strains to give more representative data concerning the *E. coli* genome.

The findings of this study could contribute to new approaches for microbial treatments such as the reverse vaccinology (Muzzi *et al*., 2007). In reverse vaccinology a cocktail of selected antigens was used to induce immunity against a specific microbial species or specific pathogenic strains by comparing the genome sequences of these strains and deducing immunological targets from the shared core or the dispensable genome, respectively (Muzzi *et al*., 2007).

In conclusion, this work describes the use of large-scale study of the sequence context of mobilome in BSI-associated *E. coli* isolates. I propose that the strategies used in this study to address and identify mobilome-rich strains could in combination with shotgun sample pyrosequencing prioritize the strains and the genomic regions that need to be sequenced. Such prioritization would avoid sequencing of hundreds of isolates to identify their novel gene pool and would reduce the cost of genomic sequencing. In this study, remarkably, despite only studying 10 *E. coli* strains, associated with a single disease type, we were able to identify at least 3 GEIs that contain novel sequences and one particular strain E105 had 13 tRNA sites occupied with GEIs. On the other hand, using our strategy MAmP we estimated the average novel, non-microarray-represented

mobilome to be 219 kb in the tested strains. This last number corresponded with observations made by other groups (Ochman *et al*., 2000).

Chapter 7 References

Abe A, de Grado M, Pfuetzner R A, Sanchez-SanMartin C, DeVinney R, Puente J L, Strynadka N C J and Finlay B B (1999). "Enteropathogenic *Escherichia coli* translocated intimin receptor, Tir, requires a specific chaperone for stable secretion." Molecular Microbiology **33**(6): 1162-1175.

Acheson D W K, Reidl J, Zhang X P, Keusch G T, Mekalanos J J and Waldor M K (1998). "*In vivo* transduction with Shiga toxin 1-encoding phage." Infection and Immunity **66**(9): 4496-4498.

Allen N L, Hilton A C, Betts R and Penn C W (2001). "Use of representational difference analysis to identify *Escherichia coli* O157-specific DNA sequences." Fems Microbiology Letters **197**(2): 195-201.

Andremont A, R. Lancar, N. A. Lê, J. M. Hattchouel, S. Baron, T. Tavakoli, M. F. Daniel, C. Tancrède, and M. G. Lê (1996). "Secular trends in mortality associated with bloodstream infections in 4268 patients hospitalized in a cancer referral center between 1975 and 1989." Clin. Microbiol. Infect **1**: 160-167.

Anjum M F, Lucchini S, Thompson A, Hinton J C D and Woodward M J (2003). "Comparative genomic indexing reveals the phylogenomics of *Escherichia coli* pathogens." Infection and Immunity **71**(8): 4674-4683.

Ardell D H and Kirsebom L A (2005). "The genomic pattern of tDNA operon expression in *E. coli*." Plos Computational Biology **1**(1): 86-99.

Aumont P, Enard C, Expert D, Pieddeloup C, Tancrede C and Andremont A (1989). "Production of hemolysin, aerobactin and enterobactin by strains of *Escherichia coli* causing bacteremia in cancer-patients, and their resistance to human-serum." Research in Microbiology **140**(1): 21-26.

Ausubel F, Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., et al (1997). Current Protocols in Molecular Biology. New York, New York: John Wiley & Sons.

Barnett M J, Fisher R F, Jones T, Komp C, Abola A P, Barloy-Hubler F, Bowser L, Capela D, Galibert F, Gouzy J, Gurjal M, Hong A, Huizar L, Hyman R W, Kahn D, Kahn M L, Kalman S, Keating D H, Palm C, Peck M C, Surzycki R, Wells D H, Yeh K C, Davis R W, Federspiel N A and Long S R (2001). "Nucleotide sequence and predicted functions of the entire *Sinorhizobium meliloti* pSymA megaplasmid." Proceedings of the National Academy of Sciences of the United States of America **98**(17): 9883-9888.

Bart A, Dankert J and van der Ende A (2000). "Representational difference analysis of *Neisseria meningitidis* identifies sequences that are specific for the hyper-virulent lineage III clone." Fems Microbiology Letters **188**(2): 111-114.

Berg D (1977). Insertion and excision of the transposable kanamycin resistance determinant Tn*5*. New York., Cold Spring Harbor Press.

Berg R (1995). "Bacterial translocation from the gastrointestinal tract." Trends Microbiol **3**: 149-154.

Bergthorsson U and Ochman H (1998). "Distribution of chromosome length variation in natural isolates of *Escherichia coli*." Molecular Biology and Evolution **15**(1): 6-16.

Bettelheim K A (1994). Biochemical characteristics of *Escherichia coli*. Wallingford, United Kingdom, CAB International.

Bettelheim K A (1994). *Escherichia coli* in domestic animals and humans. Wallingford, United Kingdom., CAB International.

Bingen E, Denamur E, Brahimi N and Elion J (1996). "Genotyping may provide rapid identification of *Escherichia coli* K1 organisms that cause neonatal meningitis." Clinical Infectious Diseases **22**(1): 152-156.

Bingen E, Picard B, Brahimi N, Mathy S, Desjardins P, Elion J and Denamur E (1998). "Phylogenetic analysis of *Escherichia coli* strains causing neonatal meningitis suggests horizontal gene transfer from a predominant pool of highly virulent B2 group strains." Journal of Infectious Diseases **177**(3): 642-650.

Bingen E H, Denamur E and Elion J (1994). "Use of ribotyping in epidemiologic surveillance of nosocomial outbreaks." Clinical Microbiology Reviews **7**(3): 311-327.

Blackwood R A, Rode C K, Pierson C L and Bloch C A (1997). "Pulsed-field gel electrophoresis genomic fingerprinting of hospital *Escherichia coli* bacteraemia isolates." Journal of Medical Microbiology **46**(6): 506-510.

Blanc-Potard A B, Tinsley C, Scaletsky I, Le Bouguenec C, Guignot J, Servin A L, Nassif X and Bernet-Camard M F (2002). "Representational difference analysis between Afa/Dr diffusely adhering *Escherichia coli* and nonpathogenic *E. coli* K-12." Infection and Immunity **70**(10): 5503-5511.

Blanco J, Blanco M, Alonso M P, Blanco J E, Gonzalez E A and Garabal J I (1992). "Characteristics of hemolytic *Escherichia coli* with particular reference to production of cytotoxic necrotizing factor type-1 (CNF1)." Research in Microbiology **143**(9): 869-878.

Blattner F R, Plunkett G, Bloch C A, Perna N T, Burland V, Riley M, ColladoVides 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**(5331): 1453-1462.

Bloch C A, Huang S H, Rode C K and Kim K S (1996). "Mapping of noninvasion Tn*phoA* mutations on the *Escherichia coli* O18:K1:H7 chromosome." Fems Microbiology Letters **144**(2-3): 171-176.

Bloch C A and Rode C K (1996). "Pathogenicity island evaluation in *Escherichia coli* K1 by crossing with laboratory strain K-12." Infection and Immunity **64**(8): 3218-3223.

Bloch C A, Rode C K, Obreque V and Russell K Y (1994). "Comparative genome mapping with mobile physical map landmarks." Journal of Bacteriology **176**(22): 7121-7125.

Bloch C A, Rode C K, Obreque V H and Mahillon J (1996). "Purification of *Escherichia coli* chromosomal segments without cloning." Biochemical and Biophysical Research Communications **223**(1): 104-111.

Blum G, Ott M, Lischewski A, Ritter A, Imrich H, Tschape H and Hacker J (1994). "Excision of large DNA regions termed pathogenicity islands from transfer-RNA-specific loci in the chromosome of an *Escherichia coli* wild-type pathogen." Infection and Immunity **62**(2): 606-614.

Bonacorsi S P P, Clermont O, Tinsley C, Le Gall I, Beaudoin J C, Elion J, Nassif X and Bingen E (2000). "Identification of regions of the *Escherichia coli* chromosome specific for neonatal meningitis-associated strains." Infection and Immunity **68**(4): 2096-2101.

Bork P (2000). "Powers and pitfalls in sequence analysis: The 70% hurdle." Genome Research **10**(4): 398-400.

Bosl M and Kersten H (1991). "A novel RNA product of the *tyrT* operon of *Escherichia coli*." Nucleic Acids Research **19**(21): 5863-5870.

Boybek A, Ray T D, Evans M C and Dyson P J (1998). "Novel site-specific DNA modification in *Streptomyces*: analysis of preferred intragenic modification sites present in a 5.7 kb amplified DNA sequence." Nucleic Acids Research **26**(14): 3364-3371.

Boyd E F and Hartl D L (1998). "Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a phylogenetically clustered distribution." Journal of Bacteriology **180**(5): 1159-1165.

Brzuszkiewicz E, Bruggemann H, Liesegang H, Emmerth M, Oschlager T, Nagy G, Albermann K, Wagner C, Buchrieser C, Emody L, Gottschalk G,

Hackert J and Dobrindt U (2006). "How to become a uropathogen: Comparative genomic analysis of extraintestinal pathogenic *Escherichia coli* strains." Proceedings of the National Academy of Sciences of the United States of America **103**(34): 12879-12884.

Buchrieser C, Prentice M and Carniel E (1998). "The 102-kilobase unstable region of *Yersinia pestis* comprises a high-pathogenicity island linked to a pigmentation segment which undergoes internal rearrangement." Journal of Bacteriology **180**(9): 2321-2329.

Buchrieser C, Weagant S D and Kaspar C W (1994). "Molecular characterization of *Yersinia enterocolitica* by pulsed-field gel-electrophoresis and hybridization of DNA fragments to *ail* and pYV probes." Applied and Environmental Microbiology **60**(12): 4371-4379.

Burland V, Shao Y, Perna N T, Plunkett G, Sofia H J and Blattner F R (1998). "The complete DNA sequence and analysis of the large virulence plasmid of *Escherichia coli* O157: H7." Nucleic Acids Research **26**(18): 4196-4204.

Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann M L and Brussow H (2003). "Phage as agents of lateral gene transfer." Current Opinion in Microbiology **6**(4): 417-424.

Canil C, Rosenshine I, Ruschkowski S, Donnenberg M S, Kaper J B and Finlay B B (1993). "Enteropathogenic *Escherichia coli* decreases the transepithelial electrical-resistance of polarized epithelial monolayers." Infection and Immunity **61**(7): 2755-2762.

Capela D, Barloy-Hubler F, Gouzy J, Bothe G, Ampe F, Batut J, Boistard P, Becker A, Boutry M, Cadieu E, Dreano S, Gloux S, Godrie T, Goffeau A, Kahn D, Kiss E, Lelaure V, Masuy D, Pohl T, Portetelle D, Puhler A, Purnelle B, Ramsperger U, Renard C, Thebault P, Vandenbol M, Weidner S and Galibert F (2001). "Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021." Proceedings of the National Academy of Sciences of the United States of America **98**(17): 9877-9882.

Carniel E, Guilvout I and Prentice M (1996). "Characterization of a large chromosomal "high-pathogenicity island" in biotype 1B *Yersinia enterocolitica*." Journal of Bacteriology **178**(23): 6743-6751.

Chan P T, Ohmori H, Tomizawa J and Lebowitz J (1985). "Nucleotide-sequence and gene organization of ColE1 DNA." Journal of Biological Chemistry **260**(15): 8925-8935.

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 Research **32**: D296-D299.

Cheetham B F and Katz M E (1995). "A role for bacteriophages in the evolution and transfer of bacterial virulence determinants." Molecular Microbiology **18**(2): 201-208.

Chen S L, Hung C S, Xu J A, Reigstad C S, Magrini V, Sabo A, Blasiar D, Bieri T, Meyer R R, Ozersky P, Armstrong J R, Fulton R S, Latreille J P, Spieth J, Hooton T M, Mardis E R, Hultgren S J and Gordon J I (2006). "Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: A comparative genomics approach." Proceedings of the National Academy of Sciences of the United States of America **103**(15): 5977-5982.

Cherifi A, Contrepois M, Picard B, Goullet P, Derycke J, Fairbrother J and Barnouin J (1990). "Factors and markers of virulence in *Escherichia coli* from human septicemia." Fems Microbiology Letters **70**(3): 279-283.

Clermont O, Bonacorsi S and Bingen E (2000). "Rapid and simple determination of the *Escherichia coli* phylogenetic group." Applied and Environmental Microbiology **66**(10): 4555-4558.

Cole S T and Saintgirons I (1994). "Bacterial genomics." Fems Microbiology Reviews **14**(2): 139-160.

Colleaux L, Dauriol L, Betermier M, Cottarel G, Jacquier A, Galibert F and Dujon B (1986). "Universal code equivalent of a yeast mitochondrial intron reading frame is expressed into *Escherichia coli* as a specific double strand endonuclease." Cell **44**(4): 521-533.

Colleaux L, Dauriol L, Galibert F and Dujon B (1988). "Recognition and cleavage site of the intron-encoded *omega* transposase." Proceedings of the National Academy of Sciences of the United States of America **85**(16): 6022-6026.

Corkill J E, Graham R, Hart C A and Stubbs S (2000). "Pulsed-field gel electrophoresis of degradation-sensitive DNAs from *Clostridium difficile* PCR ribotype 1 strains." Journal of Clinical Microbiology **38**(7): 2791-2792.

Correa L and Pittet D (2000). "Problems and solutions in hospital-acquired bacteraemia." Journal of Hospital Infection **46**(2): 89-95.

Cravioto A, Gross R J, Scotland S M and Rowe B (1979). "Adhesive factor found in strains of *Escherichia coli* belonging to the traditional infantile enteropathogenic serotypes." Current Microbiology **3**(2): 95-99.

Daran-Lapujade P, Daran J M, Kotter P, Petit T, Piper M D W and Pronk J T (2003). "Comparative genotyping of the *Saccharomyces cerevisiae* laboratory strains S288C and CEN.PK113-7D using oligonucleotide microarrays." Fems Yeast Research **4**(3): 259-269.

David Greenwood R C B S a J F P (2002). Medical Microbiology a guide to microbial infections: pathogenesis, immunity, laboratory diagnosis and control. UK, Churchill Livingstone.

Denef V J, Park J, Rodrigues J L M, Tsoi T V, Hashsham S A and Tiedje J M (2003). "Validation of a more sensitive method for using spotted oligonucleotide DNA microarrays for functional genomics studies on bacterial communities." Environmental Microbiology **5**(10): 933-943.

Dobrindt U (2005). "(Patho-)genomics of *Escherichia coli*." International Journal of Medical Microbiology **295**(6-7): 357-371.

Dobrindt U, Agerer F, Michaelis K, Janka A, Buchrieser C, Samuelson M, Svanborg C, Gottschalk G, Karch H and Hacker J (2003). "Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays." Journal of Bacteriology **185**(6): 1831-1840.

Dobrindt U, Blum-Oehler G, Nagy G, Schneider G, Johann A, Gottschalk G and Hacker J (2002). "Genetic structure and distribution of four pathogenicity islands (PAI I_{-536} to PAI IV_{536}) of uropathogenic *Escherichia coli* strain 536." Infection and Immunity **70**(11): 6365-6372.

Dobrindt U and Hacker J (2001). "Whole genome plasticity in pathogenic bacteria." Current Opinion in Microbiology **4**(5): 550-557.

Dobrindt U, Hochhut B, Hentschel U and Hacker J (2004). "Genomic islands in pathogenic and environmental microorganisms." Nature Reviews Microbiology **2**(5): 414-424.

Don R H, Cox P T, Wainwright B J, Baker K and Mattick J S (1991). "Touchdown PCR to circumvent spurious priming during gene amplification." Nucleic Acids Research **19**(14): 4008-4008.

Donnenberg M S and Nataro J P (1995). Methods for studying adhesion of diarrheagenic *Escherichia coli*. Adhesion of Microbial Pathogens. **253:** 324-336.

Dougan G, Haque A, Pickard D, Frankel G, O'Goara P and Wain J (2001). "The *Escherichia coli* gene pool." Current Opinion in Microbiology **4**(1): 90-94.

Douglas Marchuk M D, Ann Saulino and Francis S. Collins (1990). "Construction of T-vectors, a rapid and general system for direct cloning of unmodified PCR products." Nucleic Acids Research **19**(5): 1154.

Dozois C M, Daigle F and Curtiss R (2003). "Identification of pathogen-specific and conserved genes expressed *in vivo* by an avian pathogenic *Escherichia coli* strain." Proceedings of the National Academy of Sciences of the United States of America **100**(1): 247-252.

Dujon B, Belfort M, Butow R A, Jacq C, Lemieux C, Perlman P S and Vogt V M (1989). "Mobile introns - definition of terms and recommended nomenclature." Gene **82**(1): 115-118.

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

Edwards P R, and W. H. Ewing (1972). Identification of Enterobacteriaceae. Minneapolis, Minn, Burgess Publishing Co.

Eisen J A, J. F. Heidelberg, O. White, and S. L. Salszberg (2000). "Evidence for symmetric chromosomal inversions around the replication origin in bacteria." Genome Biol **1**: 0011.1–0011.9.

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*." Molecular Biology and Evolution **21**(6): 1085-1094.

Evans M and Dyson P (1993). "Pulsed-field gel-electrophoresis of *streptomyces lividans* DNA." Trends in Genetics **9**(3): 72-72.

Fonstein M and Haselkorn R (1995). "Physical mapping of bacterial genomes." Journal of Bacteriology **177**(12): 3361-3369.

Fonstein M a H (1998). Encyclopedias of bacterial genomes. New York, CHAPMAN and HALL.

Fournier M J and Ozeki H (1985). "Structure and organization of the transfer ribonucleic-acid genes of *Escherichia coli* k-12." Microbiological Reviews **49**(4): 379-397.

Francetic O and Pugsley A P (1996). "The cryptic general secretory pathway (*gsp*) operon of *Escherichia coli* K-12 encodes functional proteins." Journal of Bacteriology **178**(12): 3544-3549.

Fukiya S, Mizoguchi H, Tobe T and Mori H (2004). "Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* strains revealed by comparative genomic hybridization microarray." Journal of Bacteriology **186**(12): 3911-3921.

Fux C A, Shirtliff M, Stoodley P and Costerton J W (2005). "Can laboratory reference strains mirror 'real-world' pathogenesis?" Trends in Microbiology **13**(2): 58-63.

Gadberry M D, Malcomber S T, Doust A N and Kellogg E A (2005). "Primaclade - a flexible tool to find conserved PCR primers across multiple species." Bioinformatics **21**(7): 1263-1264.

Gal-Mor O and Finlay B B (2006). "Pathogenicity islands: a molecular toolbox for bacterial virulence." Cellular Microbiology **8**(11): 1707-1719.

Galibert F, Finan T M, Long S R, Puhler A, Abola P, Ampe F, Barloy-Hubler F, Barnett M J, Becker A, Boistard P, Bothe G, Boutry M, Bowser L, Buhrmester J, Cadieu E, Capela D, Chain P, Cowie A, Davis R W, Dreano S, Federspiel N A, Fisher R F, Gloux S, Godrie T, Goffeau A, Golding B, Gouzy J, Gurjal M, Hernandez-Lucas I, Hong A, Huizar L, Hyman R W, Jones T, Kahn D, Kahn M L, Kalman S, Keating D H, Kiss E, Komp C, Lalaure V, Masuy D, Palm C, Peck M C, Pohl T M, Portetelle D, Purnelle B, Ramsperger U, Surzycki R, Thebault P, Vandenbol M, Vorholter F J, Weidner S, Wells D H, Wong K, Yeh K C and Batut J (2001). "The composite genome of the legume symbiont *Sinorhizobium meliloti*." Science **293**(5530): 668-672.

Gauthier A, Turmel M and Lemieux C (1991). "A group-I intron in the chloroplast large subunit ribosomal-RNA gene of *Chlamydomonas eugametos* encodes a double-strand endonuclease that cleaves the homing site of this intron." Current Genetics **19**(1): 43-47.

Germon P, Roche D, Melo S, Mignon-Grasteau S, Dobrindt U, Hacker J, Schouler C and Moulin-Schouleur M (2007). "tDNA locus polymorphism and ecto-chromosomal DNA insertion hot-spots are related to the phylogenetic group of *Escherichia coli* strains." Microbiology-Sgm **153**: 826-837.

Goryachev A B, MacGregor P F and Edwards A M (2001). "Unfolding of microarray data." Journal of Computational Biology **8**(4): 443-461.

Grant S G N, Jessee J, Bloom F R and Hanahan D (1990). "Differential plasmid rescue from transgenic mouse DNAs into *Escherichia coli* methylation-restriction mutants." Proceedings of the National Academy of Sciences of the United States of America **87**(12): 4645-4649.

Guha M and Mackman N (2001). "LPS induction of gene expression in human monocytes." Cellular Signalling **13**(2): 85-94.

Guyer D M, Kao J S and Mobley H L T (1998). "Genomic analysis of a pathogenicity island in uropathogenic *Escherichia coli* CFT073: Distribution of homologous sequences among isolates from patients with

pyelonephritis, cystitis, and catheter-associated bacteriuria and from fecal samples." Infection and Immunity **66**(9): 4411-4417.

Habib A and Tabata M (2004). "Oxidative DNA damage induced by HEPES (2-[4-(2-hydroxyethyl)-1-piperazinyl]ethanesulfonic acid) buffer in the presence of Au(III)." Journal of Inorganic Biochemistry **98**(11): 1696-1702.

Hacker J, Bender L, Ott M, Wingender J, Lund B, Marre R and Goebel W (1990). "Deletions of chromosomal regions coding for fimbriae and hemolysins occur *in vitro* and *in vivo* in various extraintestinal *Escherichia coli* isolates." Microbial Pathogenesis **8**(3): 213-225.

Hacker J and Carniel E (2001). "Ecological fitness, genomic islands and bacterial pathogenicity - A Darwinian view of the evolution of microbes." Embo Reports **2**(5): 376-381.

Hacker J and Kaper J B (2000). "Pathogenicity islands and the evolution of microbes." Annual Review of Microbiology **54**: 641-679.

Hacker J, Kaper, J.B. (2002). "Pathogenicity islands and the evolution of pathogenic microbes." Curr. Top. Microbiol. Immunol **264**: 1,2.

Hanish J and McClelland M (1991). "Enzymatic cleavage of a bacterial chromosome at a transposon-inserted rare site." Nucleic Acids Research **19**(4): 829-832.

Hansen-Wester I and Hensel M (2002). "Genome-based identification of chromosomal regions specific for *Salmonella* spp." Infection and Immunity **70**(5): 2351-2360.

Hautala T, Syrjala H, Lehtinen V, Kauma H, Kauppila J, Kujala P, Pietarinen I, Ylipalosaari P and Koskela M (2005). "Blood culture Gram stain and clinical categorization based empirical antimicrobial therapy of bloodstream infection." International Journal of Antimicrobial Agents **25**(4): 329-333.

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 Research **8**(1): 11-22.

He X, Ou H Y, Yu Q, Zhou X, Wu J, Liang J, Zhang W, Rajakumar K and Deng Z (2007). "Analysis of a genomic island housing genes for DNA S-modification system in *Streptomyces lividans* 66 and its counterparts in

other distantly related bacteria." Molecular Microbiology **65**(4): 1034-1048.

Heath J D, Perkins J D, Sharma B and Weinstock G M (1992). "*Not*I genomic cleavage map of *Escherichia coli* k-12 strain MG1655." Journal of Bacteriology **174**(2): 558-567.

Henderson I R, Navarro-Garcia F and Nataro J P (1998). "The great escape: structure and function of the autotransporter proteins." Trends in Microbiology **6**(9): 370-378.

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*." Journal of Bacteriology **172**(11): 6175-6181.

Hilali F, Ruimy R, Saulnier P, Barnabe C, Lebouguenec C, Tibayrenc M and Andremont A (2000). "Prevalence of virulence genes and clonality in *Escherichia coli* strains that cause bacteremia in cancer patients." Infection and Immunity **68**(7): 3983-3989.

Hill C W and Gray J A (1988). "Effects of chromosomal inversion on cell fitness in *Escherichia coli* k-12." Genetics **119**(4): 771-778.

Hiller B, Frey B and Schumann W (1994). "Tn*5*Map, a transposon for the rapid mapping of restriction sites in plasmids." Fems Microbiology Letters **115**(2-3): 151-155.

Hiramatsu K, Katayama Y, Yuzawa H and Ito T (2002). "Molecular genetics of methicillin-resistant *Staphylococcus aureus*." International Journal of Medical Microbiology **292**(2): 67-74.

Hobman J, Penn C, and Pallen M (2007). "Laboratory strains of *Escherichia coli*: model citizens or deceitful delinquents growing old disgracefully?" Molecular Microbiology **64**(4): 881-885.

Honeycutt R J, McClelland M and Sobral B W S (1993). "physical map of the genome of *Rhizobium meliloti* 1021." Journal of Bacteriology **175**(21): 6945-6952.

Hou Y M (1999). "Transfer RNAs and pathogenicity islands." Trends in Biochemical Sciences **24**(8): 295-298.

Hsiao W, Wan I, Jones S J and Brinkman F S L (2003). "IslandPath: aiding detection of genomic islands in prokaryotes." Bioinformatics **19**(3): 418-420.

Hueck C J (1998). "Type III protein secretion systems in bacterial pathogens of animals and plants." Microbiology and Molecular Biology Reviews **62**(2): 379-433.

Ideker T, Thorsson V, Siegel A F and Hood L E (2000). "Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data." Journal of Computational Biology **7**(6): 805-817.

Inuzuka M and Helinski D R (1978). "Requirement of a plasmid-encoded protein for replication *in vitro* of plasmid R6K." Proceedings of the National Academy of Sciences of the United States of America **75**(11): 5381-5385.

Ito T, Katayama Y, Asada K, Mori N, Tsutsumimoto K, Tiensasitorn C and Hiramatsu K (2001). "Structural comparison of three types of *staphylococcal cassette* chromosome *mec* integrated in the chromosome in methicillin-resistant *Staphylococcus aureus*." Antimicrobial Agents and Chemotherapy **45**(5): 1323-1336.

Ito T, Katayama Y and Hiramatsu K (1999). "Cloning and nucleotide sequence determination of the entire *mec* DNA of pre-methicillin-resistant *Staphylococcus aureus* N315." Antimicrobial Agents and Chemotherapy **43**(6): 1449-1458.

Jain R, Rivera M C, Moore J E and Lake J A (2002). "Horizontal gene transfer in microbial genome evolution." Theoretical Population Biology **61**(4): 489-495.

Janka A, Bielaszewska M, Dobrindt U and Karch H (2002). "Identification and distribution of the enterohemorrhagic *Escherichia coli* factor for adherence (*efa1*) gene in sorbitol-fermenting *Escherichia coli* O157: H." International Journal of Medical Microbiology **292**(3-4): 207-214.

Janke B, Dobrindt U, Hacker J and Blum-Oehler G (2001). "A subtractive hybridisation analysis of genomic differences between the uropathogenic *E. coli* strain 536 and the *E. coli* K-12 strain MG1655." Fems Microbiology Letters **199**(1): 61-66.

Johnson J R and Stell A L (2000). "Extended virulence genotypes of *Escherichia coli* strains from patients with urosepsis in relation to phylogeny and host compromise." Journal of Infectious Diseases **181**(1): 261-272.

Johnson T J, Kariyawasam S, Wannemuehler Y, Mangiamele P, Johnson S J, Doetkott C, Skyberg J A, Lynne A M, Johnson J R and Nolan L K (2007). "The genome sequence of avian pathogenic *Escherichia coli* strain O1: K1: H7 shares strong similarities with human extraintestinal pathogenic *E. coli* genomes." Journal of Bacteriology **189**(8): 3228-3236.

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*." International Journal of Medical Microbiology **294**(2-3): 103-113.

Jumasbilak E, Maugard C, Michauxcharachon S, Allardetservent A, Perrin A, Ocallaghan D and Ramuz M (1995). "Study of the organization of the genomes of *Escherichia coli*, *Brucella melitensis* and *Agrobacterium tumefaciens* by insertion of a unique restriction site." Microbiology **141**: 2425-2432.

Kado C I and Liu S T (1981). "Rapid procedure for detection and isolation of large and small plasmids." Journal of Bacteriology **145**(3): 1365-1373.

Kao J S, Stucker D M, Warren J W and Mobley H L T (1997). "Pathogenicity island sequences of pyelonephritogenic *Escherichia coli* CFT073 are associated with virulent uropathogenic strains." Infection and Immunity **65**(7): 2812-2820.

Kaper JB M I, Nataro JP (1999). Pathogenicity islands and other mobile genetic elements of diarrheagenic *Escherichia coli*. Washington, DC, Am. Soc. Microbiol.

Karch H, Schubert S, Zhang D, Zhang W, Schmidt H, Olschlager T and Hacker J (1999). "A genomic island, termed high-pathogenicity island, is present in certain non-O157 shiga toxin-producing *Escherichia coli* clonal lineages." Infection and Immunity **67**(11): 5994-6001.

Kenny B D R, Stein M, Reinscheid DJ, Frey EA, Finlay BB (1997). "Enteropathogenic *E. coli* (EPEC) transfers its receptor for intimate adherence into mammalian cells." Cell **91**: 511-20.

Kinashi H, Shimaji M and Sakai A (1987). "Giant linear plasmids in *Streptomyces* which code for antibiotic biosynthesis genes." Nature **328**(6129): 454-456.

Kolter R, Inuzuka M and Helinski D R (1978). "Trans-complementation-dependent replication of a low-molecular weight origin fragment from plasmid R6K." Cell **15**(4): 1199-1208.

Koonin E V (2000). "How many genes can make a cell: The minimal-gene-set concept." Annual Review of Genomics and Human Genetics **1**: 99-116.

Koort J M K, Lukinmaa S, Rantala M, Unkila E and Siitonen A (2002). "Technical improvement to prevent DNA degradation of enteric pathogens in pulsed-field gel electrophoresis." Journal of Clinical Microbiology **40**(9): 3497-3498.

Kothapalli S, Nair S, Alokam S, Pang T, Khakhria R, Woodward D, Johnson W, Stocker B A D, Sanderson K E and Liu S L (2005). "Diversity of genome structure in *Salmonella enterica* serovar typhi populations." Journal of Bacteriology **187**(8): 2638-2650.

Krawiec S and Riley M (1990). "Organization of the bacterial chromosome." Microbiological Reviews **54**(4): 502-539.

Lalioui L and Le Bouguenec C (2001). "*afa-8* Gene cluster is carried by a pathogenicity island inserted into the $tRNA^{Phe}$ of human and bovine pathogenic *Escherichia coli* isolates." Infection and Immunity **69**(2): 937-948.

Lanata C F, Kaper J B, Baldini M M, Black R E and Levine M M (1985). "Sensitivity and specificity of DNA probes with the stool blot technique for detection of *Escherichia coli* enterotoxins." Journal of Infectious Diseases **152**(5): 1087-1090.

Lawrence J G (2005). "Horizontal and vertical gene transfer: the life history of pathogens." Contrib Microbiol **12**: 255–271.

Lawrence J G and Ochman H (1997). "Amelioration of bacterial genomes: Rates of change and exchange." Journal of Molecular Evolution **44**(4): 383-397.

Lebourgeois P, Lautier M, Mata M and Ritzenthaler P (1992). "New tools for the physical and genetic-mapping of *Lactococcus* strains." Gene **111**(1): 109-114.

Lee J S, An G, Friesen J D and Fiil N P (1981). "Location of the *tufB* promoter of *Escherichia coli* - cotranscription of *tufB* with 4 transfer-RNA genes." Cell **25**(1): 251-258.

Lennon E and Decicco B T (1991). "Plasmids of *Pseudomonas cepacia* strains of diverse origins." Applied and Environmental Microbiology **57**(8): 2345-2350.

Li Z W and Deutscher M P (2002). "RNase E plays an essential role in the maturation of *Escherichia coli* tRNA precursors." RNA-A Publication of the RNA Society **8**(1): 97-109.

Lior H (1996). Classification of *Escherichia coli*. Wallingford, United Kingdom, CAB International.

Lisitsyn N, Lisitsyn N and Wigler M (1993). "Cloning the differences between 2 complex genomes." Science **259**(5097): 946-951.

Liu S L, Hessel A and Sanderson K E (1993). "Genomic mapping with I-*Ceu*I, an intron-encoded endonuclease specific for genes for ribosomal-RNA,

in *Salmonella* spp, *Escherichia coli*, and other bacteria." Proceedings of the National Academy of Sciences of the United States of America **90**(14): 6874-6878.

Liu S L and Sanderson K E (1995). "Genomic cleavage map of *Salmonella typhi* ty2." Journal of Bacteriology **177**(17): 5099-5107.

Lloyd A L, Rasko D A and Mobley H L T (2007). "Defining genomic islands and uropathogen-specific genes in uropathogenic *Escherichia coli*." Journal of Bacteriology **189**(9): 3532-3546.

Lou Q Y, Chong S K F, Fitzgerald J F, Siders J A, Allen S D and Lee C H (1997). "Rapid and effective method for preparation of fecal specimens for PCR assays." Journal of Clinical Microbiology **35**(1): 281-283.

Lucchini S, Liu H, Jin O, Hinton J, and Yu J (2005). "Transcriptional adaptation of *Shigella flexneri* during infection of macrophages and epithelial cells: Insights into the strategies of a cytosolic bacterial pathogen." Infection and Immunity **73**(1): 88-102.

Lucchini S, Thompson A, and Hinton J (2001). "Microarrays for microbiologists." Microbiology **147**: 1403-1414.

Malloff C A, Fernandez R C and Lam W L (2001). "Bacterial comparative genomic hybridization: A method for directly identifying lateral gene transfer." Journal of Molecular Biology **312**(1): 1-5.

Mantri Y and Williams K P (2004). "Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities." Nucleic Acids Research **32**: D55-D58.

Marshall P, Davis T B and Lemieux C (1994). "The I-*Ceu*I endonuclease - purification and potential role in the evolution of *Chlamydomonas* group-I introns." European Journal of Biochemistry **220**(3): 855-859.

Marshall P and Lemieux C (1992). "The I-*Ceu*I endonuclease recognizes a sequence of 19 base-pairs and preferentially cleaves the coding strand of the *Chlamydomonas moewusii* chloroplast large subunit ribosomal-RNA gene." Nucleic Acids Research **20**(23): 6401-6407.

Maslow J N, Whittam T S, Gilks C F, Wilson R A, Mulligan M E, Adams K S and Arbeit R D (1995). "Clonal relationships among bloodstream isolates of *Escherichia coli*." Infection and Immunity **63**(7): 2409-2417.

Maurelli A T, Fernandez R E, Bloch C A, Rode C K and Fasano A (1998). ""Black holes" and bacterial pathogenicity: A large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*." Proceedings of the National Academy of Sciences of the United States of America **95**(7): 3943-3948.

McClelland M, Jones R, Patel Y and Nelson M (1987). "Restriction endonucleases for pulsed field-mapping of bacterial genomes." Nucleic Acids Research **15**(15): 5985-6005.

McClelland M, Kessler L G and Bittner M (1984). "Site-specific cleavage of DNA at 8-base-pair and 10-base-pair sequences." Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences **81**(4): 983-987.

Medini D, Donati C, Tettelin H, Masignani V and Rappuoli R (2005). "The microbial pan-genome." Current Opinion in Genetics & Development **15**(6): 589-594.

Mekalanos V L M a J J (1988). "A Novel suicide vector and its use in construction of insertion mutations: Osmoregulation of outer membrane proteins and virulence determinants in *Vibrio cholerae* requires toxR." Journal of Bacteriology **170**(6): 2575-2583.

Melkerson-Watson L J, Rode C K, Zhang L X, Foxman B and Bloch C A (2000). "Integrated genomic map from uropathogenic *Escherichia coli* J96." Infection and Immunity **68**(10): 5933-5942.

Middendorf B, Hochhut B, Leipold K, Dobrindt U, Blum-Oehler G and Hacker J (2004). "Instability of pathogenicity islands in uropathogenic *Escherichia coli* 536." Journal of Bacteriology **186**(10): 3086-3096.

Miki. T, Okada. N, and Danbara. H, (2004), “Two Periplasmic disulfide oxidoreductase, DsbA and SrgA, target outer membrane protein SpiA, a component of the *Salmonella* pathogenicity island 2 type III secretion system” The Journal of Biological Chemistry **279**(3): 34631-34642

Miller V L and Mekalanos J J (1988). "A novel suicide vector and its use in construction of insertion mutations - osmoregulation of outer membrane proteins and virulence determinants in *Vibrio cholerae* requires *toxR*." Journal of Bacteriology **170**(6): 2575-2583.

Mills M and Payne S M (1995). "Genetics and regulation of heme iron transport in *Shigella dysenteriae* and detection of an analogous system in *Escherichia coli* O157-H7." Journal of Bacteriology **177**(11): 3004-3009.

Miyazaki J, Ba-Thein W, Kumao T, Akaza H and Hayashi H (2002). "Identification of a type III secretion system in uropathogenic *Escherichia coli*." Fems Microbiology Letters **212**(2): 221-228.

Mobley H L T, Green D M, Trifillis A L, Johnson D E, Chippendale G R, Lockatell C V, Jones B D and Warren J W (1990). "Pyelonephritogenic *Escherichia coli* and killing of cultured human renal proximal tubular

epithelial cells: Role of hemolysin in some strains." Infection and Immunity **58**(5): 1281-1289.

Mokady D, Gophna U and Ron E Z (2005). "Extensive gene diversity in septicemic *Escherichia coli* strains." Journal of Clinical Microbiology **43**(1): 66-73.

Monteilhet C, Perrin A, Thierry A, Colleaux L and Dujon B (1990). "Purification and characterization of the *in vitro* activity of I-*Sce*-I, a novel and highly specific endonuclease encoded by a group-I intron." Nucleic Acids Research **18**(6): 1407-1413.

Moran N A (2002). "Microbial minimalism: Genome reduction in bacterial pathogens." Cell **108**(5): 583-586.

Morelle G (1989). "A plasmid extraction procedure on a miniprep scale." FOCUS **11**: 7-8.

Moulin-Schouleur M, Reperant M, Laurent S, Bree A, Mignon-Grasteau S, Germon P, Rasschaert D and Schouler C (2007). "Extraintestinal pathogenic *Escherichia coli* strains of avian and human origin: Link between phylogenetic relationships and common virulence patterns." Journal of Clinical Microbiology **45**(10): 3366-3376.

Morschhauser J, Vetter V, Emody L and Hacker J (1994). "Adhesin regulatory genes within large, unstable DNA regions of pathogenic *Escherichia coli* - cross-talk between different adhesin gene clusters." Molecular Microbiology **11**(3): 555-566.

Mushegian A R and Koonin E V (1996). "A minimal gene set for cellular life derived by comparison of complete bacterial genomes." Proceedings of the National Academy of Sciences of the United States of America **93**(19): 10268-10273.

Muzzi. A, Masignani. V, and Rappuoli. R, (2007),"The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials" Drug Discovery Today **12**(11/12):429-439.

Narimatsu. M, Noiri. Y, Itoh. S, Noguchi. N, Kawahara. T, and Ebisu. S, (2004), "Essential Role for the *gtfA* gene encoding a putative glycosyltransferase in the adherence of *Porphyromonas gingivalis*" Infection and Immunity **72**(5): 2698-2702

Nataro J P and Kaper J B (1998). "Diarrheagenic *Escherichia coli*." Clinical Microbiology Reviews **11**(1): 142-201.

Nelson M, Christ C and Schildkraut I (1984). "Alteration of apparent restriction endonuclease recognition specificities by DNA methylases." Nucleic Acids Research **12**(13): 5165-5173.

Nelson M, Raschke E and McClelland M (1993). "Effect of site-specific methylation on restriction endonucleases and DNA modification methyltransferases." Nucleic Acids Research **21**(13): 3139-3154.

Nixon. J, Wang. A, Field. J, Morrison. H, McArthur. A, Sogin. M, Loftus. B, and Samuelson. J, (2002), "Evidence for Lateral Transfer of Genes Encoding Ferredoxins, Nitroreductase, NADH oxidase, and alcohol dehydrogenase 3 from anaerobic prokaryotes to *Giardia lamblia* and *Entamoeba histolytica*" Eukaryotic Cell **1**(2):181-190.

Ochman H and Jones I B (2000). "Evolutionary dynamics of full genome content in *Escherichia coli*." Embo Journal **19**(24): 6637-6643.

Ochman H, Lawrence J G and Groisman E A (2000). "Lateral gene transfer and the nature of bacterial innovation." Nature **405**(6784): 299-304.

Ochman H and Selander R K (1984). "Standard reference strains of *Escherichia coli* from natural populations." Journal of Bacteriology **157**(2): 690-693.

Oelschlaeger T A, Zhang D, Schubert S, Carniel E, Rabsch W, Karch H and Hacker J (2003). "The high-pathogenicity island is absent in human pathogens of *Salmonella enterica* subspecies I but present in isolates of subspecies III and VI." Journal of Bacteriology **185**(3): 1107-1111.

Opal S M, Cross A S, Gemski P and Lyhte L W (1990). "Aerobactin and alpha-hemolysin as virulence determinants in *Escherichia coli* isolated from human blood, urine, and stool." Journal of Infectious Diseases **161**(4): 794-796.

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 Research **34**(1).

Ou H Y, Smith R, Lucchini S, Hinton J, Chaudhuri R R, Pallen M, Barer M R and Rajakumar K (2005). "ArrayOme: a program for estimating the sizes of microarray-visualized bacterial genomes." Nucleic Acids Research **33**(1).

Ou H Y H, X. Harrison, E. M. Kulasekara, B. R. Thani, A. Bin. Kadioglu, A. Lory, S. Hinton, J. C. D. Barer, M. R. Deng, Z. Rajakumar, K. (2007). "MobilomeFINDER: web-based tools for *in silico* and experimental discovery of bacterial genomic islands." Nucleic Acids Research **35**: W97-W104.

Overton T W, Whitehead R, Li Y, Snyder L A S, Saunders N J, Smith H and Cole J A (2006). "Coordinated regulation of the *Neisseria gonorrhoeae*-truncated denitrification pathway by the nitric oxide-sensitive repressor, NsrR, and nitrite-insensitive NarQ-NarP." Journal of Biological Chemistry **281**(44): 33115-33126.

Page R D M (1996). "TreeView: An application to display phylogenetic trees on personal computers." Computer Applications in the Biosciences **12**(4): 357-358.

Parham N J, Pollard S J, Chaudhuri R R, Beatson S A, Desvaux M, Russell M A, Ruiz J, Fivian A, Vila J and Henderson I R (2005). "Prevalence of pathogenicity island II_{CFT073} genes among extraintestinal clinical isolates of Escherichia coli." Journal of Clinical Microbiology **43**(5): 2425-2434.

Patel Y, Vancott E, Wilson G G and McClelland M (1990). "Cleavage at the 12-base-pair sequence 5'-TCTAGATCTAGA-3' using M.*Xba*I (TCTAG^{m6}A) methylation and *Dpn*I (G^{m6}A/TC) cleavage." Nucleic Acids Research **18**(6): 1603-1607.

Perna N T, Plunkett G, 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, Limk A, Dimalanta E T, Potamousis K D, Apodaca J, Anantharaman T S, Lin J Y, Yen G, Schwartz D C, Welch R A and Blattner F R (2001). "Genome sequence of enterohaemorrhagic *Escherichia coli* O157: H7." Nature **409**(6819): 529-533.

Perrin A, Buckle M and Dujon B (1993). "Asymmetrical recognition and activity of the I-*Sce*I endonuclease on its site and on intron-exon junctions." Embo Journal **12**(7): 2939-2947.

Picard B, Garcia J S, Gouriou S, Duriez P, Brahimi N, Bingen E, Elion J and Denamur E (1999). "The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection." Infection and Immunity **67**(2): 546-553.

Picard B and Goullet P (1988). "Correlation between electrophoretic type-b1 and type-b2 of carboxylesterase-b and host-dependent factors in *Escherichia coli* septicemia." Epidemiology and Infection **100**(1): 51-61.

Qiang B Q, McClelland M, Poddar S, Spokauskas A and Nelson M (1990). "The apparent specificity of *Not*I (5'-GCGGCCGC-3') is enhanced by M.*Fnu*DII or M.*Bep*I methyltransferases (5'-^{m}CGCG-3') - cutting bacterial chromosomes into a few large pieces." Gene **88**(1): 101-105.

Rajakumar K, Sasakawa C and Adler B (1997). "Use of a novel approach, termed island probing, identifies the *Shigella flexneri she* pathogenicity

island which encodes a homolog of the immunoglobulin A protease-like family of proteins." Infection and Immunity **65**(11): 4606-4614.

Rajashekara G, Glasner J D, Glover D A and Splitter G A (2004). "Comparative whole-genome hybridization reveals genomic islands in *Brucella* species." Journal of Bacteriology **186**(15): 5040-5051.

Rasko D A, Phillips J A, Li X and Mobley H L T (2001). "Identification of DNA sequences from a second pathogenicity island of uropathogenic *Escherichia coli* CFT073: Probes specific for uropathogenic populations." Journal of Infectious Diseases **184**(8): 1041-1049.

Ray T, Weaden J and Dyson P (1992). "Tris-dependent site-specific cleavage of *Streptomyces lividans* DNA." Fems Microbiology Letters **96**(2-3): 247-252.

Reacher M H, Shah A, Livermore D M, Wale M C J, Graham C, Johnson A P, Heine H, Monnickendam M A, Barker K F, James D and George R C (2000). "Bacteraemia and antibiotic resistance of its pathogens reported in England and Wales between 1990 and 1998: trend analysis." British Medical Journal **320**(7229): 213-216.

Reiter W D, Palm P and Yeats S (1989). "Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements." Nucleic Acids Research **17**(5): 1907-1914.

Ritter A, Gally D L, Olsen P B, Dobrindt U, Friedrich A, Klemm P and Hacker J (1997). "The Pai-associated *leuX* specific $tRNA^{Leu}_{5}$ affects type 1 fimbriation in pathogenic *Escherichia coli* by control of FimB recombinase expression." Molecular Microbiology **25**(5): 871-882.

Rocke D M and Durbin B (2001). "A model for measurement error for gene expression arrays." Journal of Computational Biology **8**(6): 557-569.

Rode C K, Melkerson-Watson L J, Johnson A T and Bloch C A (1999). "Type-specific contributions to chromosome size differences in *Escherichia coli*." Infection and Immunity **67**(1): 230-236.

Rode C K, Obreque V H and Bloch C A (1995). "New tools for integrated genetic and physical analyses of the *Escherichia coli* chromosome." Gene **166**(1): 1-9.

Rodriguez-Siek K, Giddings C, Doetkott C, Johnson T, Fakhr M and Nolan L (2005). "Comparison of *Escherichia coli* isolates implicated in human urinary tract infection and avian colibacillosis." Microbiology **151**: 2097-2110.

Romling U and Tummler B (2000). "Achieving 100% typeability of *Pseudomonas aeruginosa* by pulsed-field gel electrophoresis." Journal of Clinical Microbiology **38**(1): 464-465.

Rowley K B, Elford R M, Roberts I and Holmes W M (1993). "*In vivo* regulatory responses of 4 *Escherichia coli* operons which encode leucyl-transfer RNAs." Journal of Bacteriology **175**(5): 1309-1315.

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." Proceedings of the National Academy of Sciences of the United States of America **97**(26): 14668-14673.

Salyers A W, D. (2002). Bacterial pathogenesis, a molecular approach. Washington, D.C, ASM PRESS.

Schmidt H and Hensel M (2004). "Pathogenicity islands in bacterial pathogenesis." Clinical Microbiology Reviews **17**(1): 14-56.

Schnell R, Abdulkarim F, Kalman M and Isaksson L A (2003). "Functional EF-Tu with large C-terminal extensions in an *E. coli* strain with a precise deletion of both chromosomal *tuf* genes." Febs Letters **538**(1-3): 139-144.

Schouler C, Koffmann F, Amory C, Leroy-Setrin S and Moulin-Schouleur M (2004). "Genomic subtraction for the identification of putative new virulence factors of an avian pathogenic *Escherichia coli* strain of O2 serogroup." Microbiology-Sgm **150**: 2973-2984.

Schubert S, Picard B, Gouriou S, Heesemann J and Denamur E (2002). "*Yersinia* high-pathogenicity island contributes to virulence in *Escherichia coli* causing extraintestinal infections." Infection and Immunity **70**(9): 5335-5337.

Schubert S, Rakin A, Fischer D, Sorsa J and Heesemann J (1999). "Characterization of the integration site of *Yersinia* high-pathogenicity island in *Escherichia coli*." Fems Microbiology Letters **179**(2): 409-414.

Schubert S, Rakin A, Karch H, Carniel E and Heesemann J (1998). "Prevalence of the "high-pathogenicity island" of *Yersinia* species among *Escherichia coli* strains that are pathogenic to humans." Infection and Immunity **66**(2): 480-485.

Scott. D, Grossmann. G, Tame. J, Byron. O, Wilson. K, and Otto. B, (2002). “Low resolution solution structure of apo form of *Escherichia coli* haemoglobin protease Hbp.” Journal of Molecular Biology **315**: 1179-1187.

Scott F W and Pitt T L (2004). "Identification and characterization of transmissible *Pseudomonas aeruginosa* strains in cystic fibrosis patients in England and Wales." Journal of Medical Microbiology **53**(7): 609-615.

Selander R K, Caugant D A, Ochman H, Musser J M, Gilmour M N and Whittam T S (1986). "Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics." Applied and Environmental Microbiology **51**(5): 873-884.

Selander R K, D. A. Caugant, and T. S. Whittam (1987). Genetic structure and variation in natural populations of *Escherichia coli*. Washington, D.C., American Society for Microbiology.

Shinder G and Gold M (1988). "The Nu1 subunit of bacteriophage-lambda terminase binds to specific sites in *cos* DNA." Journal of Virology **62**(2): 387-392.

Shu S E, Setianingrum E, Zhao L C, Li Z Y, Xu H X, Kawamura Y and Ezaki T (2000). "I-*Ceu*I fragment analysis of the *Shigella* species: evidence for large-scale chromosome rearrangement in *S. dysenteriae* and *S. flexneri*." Fems Microbiology Letters **182**(1): 93-98.

Simon R, Priefer U and Puhler A (1983). "A broad host range mobilization system for *in vivo* genetic-engineering - transposon mutagenesis in gram-negative bacteria." Bio-Technology **1**(9): 784-791.

Smith M J, J; Landers, T; Jordan, J (1990). "High efficiency bacterial electroporation: 1 x 10^{10} *E. coli* transformants/µg." FOCUS **12**(2): 38-40.

Snyder L A S and Saunders N J (2006). "The majority of genes in the pathogenic *Neisseria* species are present in non-pathogenic *Neisseria lactamica*, including those designated as 'virulence genes'." Bmc Genomics **7**.

Snyder L C, W. (2003). Molecular Genetics of Bacteria. USA, ASM PRESS.

Sorsa L J, Dufke S and Schubert S (2004). "Identification of novel virulence-associated loci in uropathogenic *Escherichia coli* by suppression subtractive hybridization." Fems Microbiology Letters **230**(2): 203-208.

Spitz J, Yuhan R, Koutsouris A, Blatt C, Alverdy J and Hecht G (1995). "Enteropathogenic *Escherichia coli* adherence to intestinal epithelial monolayers diminishes barrier function." American Journal of Physiology-Gastrointestinal and Liver Physiology **31**(2): G374-G379.

Stabler R and Hinds J (2006). "The majority of genes in the pathogenic *Neisseria* species are present in non-pathogenic *Neisseria lactamica*, including those designated as virulence genes: response." Bmc Genomics **7**.

Stacyphipps S, Mecca J J and Weiss J B (1995). "Multiplex PCR assay and simple preparation method for stool specimens detect enterotoxigenic *Escherichia coli* DNA during course of infection." Journal of Clinical Microbiology **33**(5): 1054-1059.

Stephenson F H (2003). Calculation for molecular biology and biotechnology, a guide to mathematics in the laboratory. USA, ACADEMIC PRESS.

Stocki S L, Babiuk L A, Rawlyk N A, Potter A A and Allan B J (2002). "Identification of genomic differences between *Escherichia coli* strains pathogenic for poultry and *E. coli* K-12 MG1655 using suppression subtractive hybridization analysis." Microbial Pathogenesis **33**(6): 289-298.

Strassheim D, Park J S and Abraham E (2002). "Sepsis: current concepts in intracellular signaling." International Journal of Biochemistry & Cell Biology **34**(12): 1527-1533.

Suwanto A and Kaplan S (1992). "Chromosome transfer in *Rhodobacter sphaeroides*: Hfr formation and genetic-evidence for two unique circular chromosomes." Journal of Bacteriology **174**(4): 1135-1145.

Tancrede C H and Andremont A O (1985). "Bacterial translocation and gram-negative bacteremia in patients with hematological malignancies." Journal of Infectious Diseases **152**(1): 99-103.

Thierry A, Perrin A, Boyer J, Fairhead C, Dujon B, Frey B and Schmitz G (1991). "Cleavage of yeast and bacteriophage-T7 genomes at a single site using the rare cutter endonuclease I-*Sce*-I." Nucleic Acids Research **19**(1): 189-190.

Thompson J D, Gibson T J, Plewniak F, Jeanmougin F and Higgins D G (1997). "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools." Nucleic Acids Research **25**(24): 4876-4882.

Thong K L, Puthucheary S D and Pang T (1997). "Genome size variation among recent human isolates of *Salmonella typhi*." Research in Microbiology **148**(3): 229-235.

Torres A G and Payne S M (1997). "Haem iron-transport system in enterohaemorrhagic *Escherichia coli* O157:H7." Molecular Microbiology **23**(4): 825-833.

Turmel M, Otis C, Cote V and Lemieux C (1997). "Evolutionarily conserved and functionally important residues in the I-*Ceu*I homing endonuclease." Nucleic Acids Research **25**(13): 2610-2619.

van Passel M W J, Bart A, Waaijer R J A, Luyf A C M, van Kampen A H C and van der Ende A (2004). "An *in vitro* strategy for the selective isolation of anomalous DNA from prokaryotic genomes." Nucleic Acids Research **32**(14).

Wang L R, Chen S, Xu T G, Taghizadeh K, Wishnok J S, Zhou X F, You D L, Deng Z X and Dedon P C (2007). "Phosphorothioation of DNA in bacteria by *dnd* genes." Nature Chemical Biology **3**(11): 709-710.

Wattiau P, Woestyn S and Cornelis G R (1996). "Customized secretion chaperones in pathogenic bacteria." Molecular Microbiology **20**(2): 255-262.

Weinstein M P (1997). "The clinical significance of positive blood cultures in the 1990s: A prospective comprehensive evaluation of the microbiology, epidemiology, and outcome of bacteremia and fungemia in adults." Clinical Infectious Diseases **24**(4): 584-602.

Welch R A, Burland V, Plunkett G, 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 T, Donnenberg M S and Blattner F R (2002). "Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*." Proceedings of the National Academy of Sciences of the United States of America **99**(26): 17020-17024.

Welsh J and McClelland M (1990). "Fingerprinting genomes using PCR with arbitrary primers." Nucleic Acids Research **18**(24): 7213-7218.

Whitehead R N, Overton T W, Snyder L A S, McGowan S J, Smith H, Cole J A and Saunders N J (2007). "The small FNR regulon of *Neisseria gonorrhoeae*: comparison with the larger *Escherichia coli* FNR regulon and interaction with the NarQ-NarP regulon." Bmc Genomics **8**.

Whittam T S (1996). Genetic variation and evolutionary processes in natural populations of *Escherichia coli*. Washington, D.C, American Society for Microbiology.

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." Infection and Immunity **61**(5): 1619-1629.

Wick L M, Qi W H, Lacher D W and Whittam T S (2005). "Evolution of genomic content in the stepwise emergence of *Escherichia coli* O157: H7." Journal of Bacteriology **187**(5): 1783-1791.

Willenbrock H, Petersen A, Sekse C, Kiil K, Wasteson Y and Ussery D W (2006). "Design of a seven-genome *Escherichia coli* microarray for comparative genomic profiling." Journal of Bacteriology **188**(22): 7713-7721.

Williams K P (2002). "Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies." Nucleic Acids Research **30**(4): 866-875.

Witney A A, Marsden M, Holden T G, 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." Applied and Environmental Microbiology **71**(11): 7504-7514.

Wolfgang M C, Kulasekara B R, Liang X Y, Boyd D, Wu K, Yang Q, Miyada C G and Lory S (2003). "Conservation of genome content and virulence determinants among clinical and environmental isolates of *Pseudomonas aeruginosa*." Proceedings of the National Academy of Sciences of the United States of America **100**(14): 8484-8489.

Wong K K and McClelland M (1992). "Dissection of the *Salmonella typhimurium* genome by use of a Tn*5* derivative carrying rare restriction sites." Journal of Bacteriology **174**(11): 3807-3811.

Yamamoto T, Kaneko M, Changchawalit S, Serichantalergs O, Ijuin S and Echeverria P (1994). "Actin accumulation associated with clustered and localized adherence in *Escherichia coli* isolated from patients with diarrhea." Infection and Immunity **62**(7): 2917-2929.

Zhang Y S, Yakrus M A, Graviss E A, Williams-Bouyer N, Turenne C, Kabani A and Wallace R J (2004). "Pulsed-field gel electrophoresis study of *Mycobacterium abscessus* isolates previously affected by DNA degradation." Journal of Clinical Microbiology **42**(12): 5582-5587.

Zhou X F, He X Y, Li A Y, Lei F, Kieser T and Deng Z X (2004). "*Streptomyces coelicolor* A3(2) lacks a genomic island present in the chromosome of *Streptomyces lividans* 66." Applied and Environmental Microbiology **70**(12): 7110-7118.

Zhou X F, He X Y, Liang J D, Li A Y, Xu T G, Kieser T, Helmann J D and Deng Z X (2005). "A novel DNA modification by sulphur." Molecular Microbiology **57**(5): 1428-1438.

Zuber U and Schumann W (1991). "Tn*5cos*: a transposon for restriction mapping of large plasmids using phage lambda terminase." Gene **103**(1): 69-72.

Chapter 8 Appendix

Preparation of SOC medium (100 ml):

SOB medium (autoclaved)	98 ml
2M glucose (autoclaved)	1 ml
2M Mg^{2+} stock (autoclaved)	1 ml

The Mg^{2+} stock was made as a 100 ml solution containing 20.33g $MgCl_2.6H_2O$ and 24.65g $MgSO_4.7H_2O$.

Preparation of SOB medium (400 ml):

Tryptone (OXOID)	8g
Yeast extract Bacto	2g
NaCl	0.234g
KCl	0.074g

Preparation of 2×YT medium pH 7.0 (1000 ml):

Bacto-tryptone	16g
Yeast extract Bacto	10g
NaCl	5g

Preparation of 5×M9 medium (500 ml):

Na_2HPO_4	15g
KH_2PO_4	7.5g
NH_4Cl	2.5g
NaCl	1.25g
$CaCl_2$	7.5mg

Table 8.1. The SGSP-PCR amplicons produced by U and T3 (vector primer) in MG1655

tRNA site	*Bam*H I	*Eco*R I	*Eco*R V	*Hin*c II	*Hin*d III	*Pst* I	*Sal* I
*aspV*1	6.3	0.7	0.8	0.5	1.0	22.9	12.6
thrW	5.3	6.5	1.4	3.7	4.2	0.4	22.1
serW	4.4	1.5	2.6	1.4	1.9	0.9	16.6
serT	5.1	1.4	3.2	4.9	27.5	0.9	19.0
serX	14.3	8.0	2.4	0.2	2.6	3.3	0.8
serU	14.8	2.0	2.1	1.6	0.4	7.7	16.0
*asnT*1	5.2	8.2	3.7	0.5	5.1	3.4	9.1
*asnV*1	3.0	6.4	1.8	2.2	7.2	4.2	15.2
argW	6.7	1.6	2.3	0.5	1.8	0.4	15.5
ssrA	10.8	1.4	0.3	1.1	1.2	1.9	5.8
metV	1.9	31.1	1.7	1.0	7.5	2.5	1.0
glyU	6.7	0.9	7.8	0.7	2.0	9.8	3.7
pheV	15.8	19.4	2.0	0.2	10.5	0.2	8.1
selC	18.9	1.3	0.2	2.2	15.7	4.4	9.1
*pheU*1	13.6	5.4	0.4	0.3	14.6	4.1	0.4
leuX	25.3	10.5	0.3	1.1	3.2	2.6	18.1
*aspV*2	6.4	0.8	0.9	0.6	1.0	22.9	12.6
*pheU*2	13.5	5.3	0.3	0.2	14.5	4.0	0.3
*asnV*2	2.0	5.4	0.8	1.2	6.2	3.2	14.2
*aspV*3	6.2	0.5	0.6	0.4	0.8	22.7	12.4
*asnT*3	5.1	8.1	3.6	0.3	5.0	3.2	8.9

[a]Numbers after some tRNA sites refer to different U primers used in the SGSP-PCR.

Table 8.2. The SGSP-PCR amplicons produced by D and T3 (vector primer) in MG1655

tRNA site[a]	*Bam*H I	*Eco*R I	*Eco*R V	*Hin*c II	*Hin*d III	*Pst* I	*Sal* I
*aspV*1	3.4	2.3	2.5	1.6	2.3	5.8	6.7
thrW	14.2	3.9	4.2	3.4	2.8	4.0	4.0
serW	12.0	0.1	8.6	0.2	1.8	0.6	21.6
serT	0.9	0.4	2.9	1.2	18.3	13.1	1.2
serX	2.2	5.3	1.1	0.3	1.6	2.3	0.3
serU	7.5	10.5	2.2	0.4	1.5	5.7	11.4
asnT	5.5	2.6	5.1	1.2	1.6	5.2	1.2
*asnV*1	4.5	7.3	0.3	3.8	2.5	7.5	9.3
argW	6.4	3.4	1.5	1.4	1.6	3.3	17.4
ssrA	9.8	4.8	3.1	1.7	1.6	8.7	3.7
metV	6.9	10.7	2.5	0.6	14.5	7.6	0.6
glyU	5.9	1.4	0.7	0.4	2.0	2.8	9.0
pheV	13.8	17.4	5.4	2.1	11.9	0.7	2.1
selC	16.6	1.7	2.9	0.9	11.1	13.7	4.0
*pheU*1	10.2	5.5	0.5	0.4	13.3	4.1	0.6
leuX	8.9	6.0	2.9	0.7	15.6	0.8	0.7
*aspV*2	3.6	2.5	2.7	1.8	2.5	6.0	6.9
*pheU*2	10.2	5.5	0.5	0.4	13.3	4.1	0.6
*asnV*2	4.0	6.9	2.6	3.4	2.0	7.1	8.9
*aspV*3	3.6	2.5	2.7	1.8	2.5	6.0	6.9
*asnT*3	5.5	2.6	5.1	1.2	1.6	5.2	1.2

[a]numbers after some tRNA sites refer to different D primers used in the SGSP-PCR.

Table 8.3. Summary of available sequence data for GEIs identified by the tRIP-PCR and SGSP-PCR strategies

Strain-tRNA gene	SGSP-PCR amplicon	Total sequence length (bp)	Length of GEI-specific sequence (bp)	Blastn hit(s) of GEI-specific sequence (length of match [bp])	Blastn results	NCBI GeneBank accession number
E105-*selC*	#U-#T3 (1.5kb) *Sal*I (#T3)	442	297	EAEC042 (161)	e-value=4e-19 bit score =97.6 ID= 82%	In-completed genome sequence
				EAEC042 (136)	e-value=2e-18 bit score =95.6 ID= 83%	
E109-*selC*	#U-#T3 (1.5 kb) *Sal*I (#U)	654	494	EAEC042 (312)	e-value=e-161 bit score =571 ID= 98%	In-completed genome sequence
				EAEC042 phage integrase (182)	e-value=2e-93 bit score =345 ID= 98%	
E110-*selC*	#D-#T3 (2.0kb) *Hin*cII (#D)	734	235	CFT073 hypo.protein (734)	e-value=0.0 bit score =1345 ID= 99%	AE014075.1
E102- *glyU*	#D-#T3 (0.6kb) *Eco*RV (#D)	574	482	EDL933 (574)	e-value=0.0 bit score =976 ID= 97%	AE005174.2
E104- *glyU*	#U-#T3 (2.5kb) *Hin*dIII (#KS)	558	558	EDL933 (558)	e-value=0.0 bit score =1026 ID= 99%	AE005174.2
E107- *glyU*	#U-#T3 (2.7kb) *Hin*dIII (#KS)	314	314	EDL933 (314)	e-value=5e-150 bit score =538 ID= 97%	AE005174.2
E108-*serU*	#U-#T3 (1.0kb) *Sal*I (#KS)	247	247	CFT073 (247)	e-value=2e-123 bit score =449 ID= 99%	AE014075.1
E109- *serU*	#U-#T3 (1.0kb) *Sal*I (#T3)	334	334	CFT073 (334)	e-value=3e-157 bit score =562 ID= 96%	AE014075.1
E111- *serU*	#U-#T3 (2.0kb) *Bam*HI (#KS)	750	473	EAEC042 integrase prophage CP-9330(172)	e-value=2e-123 bit score =449 ID= 99%	AE014075.1
				EAEC042 (301)	e-value=2e-123 bit score =449 ID= 99%	

Strain-tRNA gene	SGSP-PCR amplicon	Total sequence length (bp)	Length of GEI-specific sequence (bp)	Blastn hit(s) of GEI-specific sequence (length of match [bp])	Blastn results	NCBI GeneBank accession number
E106-*asnV*	#D-#T3 (1.5kb) *Hin*dIII (#KS)	738	738	CFT073 transposase and hypo. protein (738)	e-value=0.0 bit score =1358 ID= 99%	AE014075.1
E108-*asnV*	#D-#T3 (1.5kb) *Hin*dIII (#KS)	750	750	CFT073 transposase and hypo. protein (750)	e-value=0.0 bit score =1369 ID= 99%	AE014075.1
E108-*asnV*	#D-#T3 (1.5kb) *Hin*dIII (#D)	291	0	CFT073 protein erfK/srfK precursor (291)	e-value=2e-133 bit score =483 ID= 96%	AE014075.1
E109-*asnV*	#U-#T3 (3.0kb) *Hin*dIII (#U)	448	0	CFT073 notrogen assimilation protein (448)	e-value=0.0 bit score =800 ID= 98%	AE014075.1
E109-*asnV*	#D-#T3 (1.5kb) *Hin*dIII (#KS)	747	747	CFT073 transposase and hypo. protein (747)	e-value=0.0 bit score =1360 ID= 99%	AE014075.1
E103-*thrW*	#U-#T3 (2.3kb) *Sal*I (#T3)	72	72	CFT073 (72)	e-value=2e-118 bit score =430 ID= 99%	AE014075.1
E105-*thrW*	#U-#T3 (0.9kb) *Eco*RV (#U)	750	665	CFT073 CP4-like integrase (750)	e-value=0.0 bit score =1375 ID= 99%	AE014075.1
E105-*thrW*	#U-#T3 (2.5kb) *Hin*cII (#U)	750	662	CFT073 CP4-like integrase (750)	e-value=0.0 bit score =1369 ID= 99%	AE014075.1
E107-*thrW*	#U-#T3 (2.4kb) *Sal*I (#T3)	222	222	MG1655 oxidoreductase (222)	e-value=2e-102 bit score =379 ID= 97%	U00096.2
E107-*thrW*	#U-#T3 (2.4kb) *Sal*I (#U)	865	865	MG1655 transcriptional regulator and conserved protein (865)	e-value=0.0 bit score =1493 ID= 98%	U00096.2
E110-*thrW*	#U-#T3 (2.4kb) *Pst*I (#KS)	600	600	CFT073 hypo. Protein and haemoglobin protease (600)	e-value=0.0 bit score =1103 ID= 99%	AE014075.1
E111-*thrW*	#U-#T3 (1.2kb) *Eco*RV (#U)	750	665	MG1655 CP4-6 prophage (750)	e-value=0.0 bit score =1380 ID= 99%	U00096.2
E104-*metV*	#D-#T3 (2.5kb) *Bam*HI (#KS)	242	242	CFT073 PTS Maltose and glucose-specific IIAC component antiterminator (242)	e-value=1e-119 bit score =436 ID= 99%	AE014075.1

Strain-tRNA gene	SGSP-PCR amplicon	Total sequence length (bp)	Length of GEI-specific sequence (bp)	Blastn hit(s) of GEI-specific sequence (length of match [bp])	Blastn results	NCBI GeneBank accession number
E105-*metV*	#D-#T3 (3.0kb) *Bam*HI (#KS)	613	613	CFT073 PTS Maltose and glucose-specific IIAC component antiterminator (613)	e-value=0.0 bit score =1112 ID= 99%	AE014075.1
E106-*metV*	#D-#T3 (3.0kb) *Bam*HI (#KS)	793	793	CFT073 PTS Maltose and glucose-specific IIAC component antiterminator (793)	e-value=0.0 bit score =1452 ID= 99%	AE014075.1
E106-*metV*	#U-#T3 (1.0kb) *Hin*dIII (#KS)	279	279	CFT073 (279)	e-value=2e-133 bit score =483 ID= 97%	AE014075.1
E108-*metV*	#U-#T3 (1.0kb) *Hin*dIII (#KS)	517	364	CFT073 (517)	e-value=0.0 bit score =915 ID= 98%	AE014075.1
E109-*metV*	#U-#T3 (2.0kb) *Eco*RI (#U)	729	729	CFT073 hypo. Protein and conserved hypo. protein (729)	e-value=0.0 bit score =1319 ID= 99%	AE014075.1
E110-*metV*	#D-#T3 (3.0kb) *Bam*HI (#KS)	684	684	CFT073 PTS Maltose and glucose-specific IIAC component antiterminator (684)	e-value=0.0 bit score =1247 ID= 99%	AE014075.1
E110-*metV*	#U-#T3 (1.0kb) *Hin*dIII (#KS)	709	354	CFT073 membrane-bound lytic murein transglycosylase A precursor (709)	e-value=0.0 bit score =1303 ID= 99%	AE014075.1
E111-*metV*	#D-#T3 (0.6kb) *Hin*cII (#D)	463	220	CFT073 N-acetylmuramoyl-L-alanine amiC precursor (463)	e-value=0.0 bit score =850 ID= 99%	AE014075.1
E102-*serX*	#U-#T3 (1.6kb) *Eco*RI (#KS)	575	575	*S. boydii* Sb227 conserved hypo. Protein (575)	e-value=0.0 bit score =821 ID= 92%	CP000036.1
E103-*serX*	#U-#T3 (1.6kb) *Eco*RI (#KS)	295	295	*S. boydii* Sb227 (290)	e-value=6e-109 bit score =401 ID= 91%	CP000036.1
E103-*serX*	#U-#T3 (0.9kb) *sal*I (#T3)	222	222	*S. boydii* Sb227 conserved hypo. Protein (222)	e-value=8e-97 bit score =361 ID= 95%	CP000036.1
E104-*serX*	#U-#T3 (0.7kb) *Hin*dIII (#KS)	280	194	CFT073 hypo.protein (280)	e-value=1e-125 bit score =457 ID= 96%	AE014075.1

Strain-tRNA gene	SGSP-PCR amplicon	Total sequence length (bp)	Length of GEI-specific sequence (bp)	Blastn hit(s) of GEI-specific sequence (length of match [bp])	Blastn results	NCBI GeneBank accession number
E104-*serX*	#U-#T3 (0.9kb) *Eco*RV (#U)	750	221	CFT073 hypo.protein YcdT (534)	e-value=0.0 bit score =944 ID= 98%	AE014075.1
				CFT073 (221)	e-value=6e-108 bit score =399 ID= 99%	
E105-*serX*	#U-#T3 (1.5kb) *Pst*I (#KS)	685	618	CFT073 conserved hypo.protein and hypo. Protein (618)	e-value=0.0 bit score =1109 ID= 99%	AE014075.1
E110-*serX*	#U-#T3 (1.5kb) *Eco*RV (#U)	750	219	CFT073 conserved hypo.protein and hypo. Protein (750)	e-value=0.0 bit score =1375 ID= 99%	AE014075.1
E111-*serX*	#U-#T3 (0.5kb) *Sal*I (#T3)	316	51	MG1655 (313)	e-value=8e-148 bit score =531 ID= 97%	U00096.2
E102-*argW*	#U-#T3 (1.8kb) *Bam*HI (#KS)	421	420	CFT073 hypo.protein and hypo. Protein ydeU (420)	e-value=0.0 bit score =739 ID= 98%	AE014075.1
E103-*argW*	#U-#T3 (1.9kb) *Bam*HI (#KS)	750	749	CFT073 hypo.protein and hypo. Protein ydeU (749)	e-value=0.0 bit score =1314 ID= 98%	AE014075.1
E104-*argW*	#U-#T3 (0.5kb) *Pst*I (#KS)	256	0	EDL933 putative transport (256)	e-value=5e-129 bit score =468 ID= 99%	AE005174.2
E104-*argW*	#D-#T3 (1.5kb) *Sal*I (#T3)	719	418	EDL933 sucrose specific transcriptional regulator D-serine permease (718)	e-value=0.0 bit score =1275 ID= 98%	AE005174.2
E105-*argW*	#U-#T3 (1.6kb) *Bam*HI (#KS)	224	220	CFT073 hypo. Protein ydeU (220)	e-value=4e-105 bit score =388 ID= 98%	AE014075.1
E106-*argW*	#U-#T3 (3.0kb) *Eco*RV (#U)	694	425	EDL933 putative transport and putative prophage integrase (690)	e-value=0.0 bit score =1173 ID= 97%	AE005174.2
E107-*argW*	#U-#T3 (1.3kb) *Sal*I (#T3)	750	674	EDL933 sucrose permease (674)	e-value=0.0 bit score =1129 ID= 96%	AE005174.2

Strain-tRNA gene	SGSP-PCR amplicon	Total sequence length (bp)	Length of GEI-specific sequence (bp)	Blastn hit(s) of GEI-specific sequence (length of match [bp])	Blastn results	NCBI GeneBank accession number
E110-*argW*	#U-#T3 (1.5kb) *Eco*RV (#U)	722	451	CFT073 hypo.protein yfdC and hypo. Protein (722)	e-value=0.0 bit score =1323 ID= 99%	AE014075.1
E107-*leuX*	#U-#T3 (1.5kb) *Eco*RI (#KS)	856	426	*S. flexneri* 2a Sf301 putative P4-type integrase and putative oxidoreductase (855)	e-value=0.0 bit score =1238 ID= 92%	AE005674.1
E110-*leuX*	#D-#T3 (3.0kb) *Eco*Ri (#KS)	576	576	CFT073 hypo.protein and hypo. Protein (576)	e-value=0.0 bit score =1064 ID= 100%	AE014075.1
E110-*leuX*	#D-#T3 (2.5kb) *Hin*dIII (#KS)	890	890	CFT073 hypo.protein and hypo. Protein (890)	e-value=0.0 bit score =1637 ID= 99%	AE014075.1
E111-*leuX*	#D-#T3 (1.0kb) *Eco*RV (#KS)	270	182	CFT073 hypo.protein and hypo. Protein yjhS precursor (266)	e-value=9e-127 bit score =460 ID= 97%	AE014075.1
E102-*asnT*	#D-#T3 (1.5kb) *Hin*dIII (#KS)	369	369	CFT073 hypo.protein and shikimate transporter (369)	e-value=6e-180 bit score =638 ID= 97%	AE014075.1
E103-*asnT*	#D-#T3 (1.5kb) *Hin*dIII (#KS)	524	524	CFT073 hypo.protein and shikimate transporter (524)	e-value=0.0 bit score =946 ID= 99%	AE014075.1
E104-*asnT*	#D-#T3 (1.5kb) *Hin*dIII (#KS)	750	750	MG1655 putative glycosyltransferase and shikimate transporter (750)	e-value=0.0 bit score =1336 ID= 98%	U00096.2
E105-*asnT*	#D-#T3 (2.5kb) *Eco*RI (#KS)	615	615	CFT073 hypo.protein and hypo. Protein (615)	e-value=0.0 bit score =1062 ID= 97%	AE014075.1
E106-*asnT*	#D-#T3 (1.5kb) *Hin*cII (#D)	750	765	CFT073 shikimate transporter (750)	e-value=0.0 bit score =1375 ID= 99%	AE014075.1
E107-*asnT*	#D-#T3 (1.6kb) *Hin*dIII (#KS)	614	614	CFT073 hypo.protein and shikimate transporter (614)	e-value=0.0 bit score =1134 ID= 100%	AE014075.1
E110-*asnT*	#D-#T3 (1.6kb) *Hin*dIII (#KS)	945	945	CFT073 hypo.protein and shikimate transporter (945)	e-value=0.0 bit score =1746 ID= 100%	AE014075.1

Strain-tRNA gene	SGSP-PCR amplicon	Total sequence length (bp)	Length of GEI-specific sequence (bp)	Blastn hit(s) of GEI-specific sequence (length of match [bp])	Blastn results	NCBI GeneBank accession number
E111-*asnT*	#D-#T3 (1.2kb) *Sal*I (#D)	570	567	EDL933 putative transport protein, shikimate (570)	e-value=0.0 bit score =990 ID= 97%	AE005174.2
E105-*aspV*	#D-#T3 (2.0kb) *Pst*I (#KS)	750	440	EAECO42 (440)	e-value=0.0 bit score =866 ID= 99%	In-completed genome sequence
E107-*aspV*	#D-#T3 (1.9kb) *Hin*dIII (#D)	750	750	E24377A hyrolase, carbon-nitrogen family (750)	e-value=0.0 bit score =1380 ID= 99%	CP000800.1
E111-*aspV*	#D-#T3 (2.2kb) *Eco*RV (#D)	748	371	EAECO42 (371)	e-value=0.0 bit score =706 ID= 98%	In-completed genome sequence
E102-*pheV*	#U-#T3 (2.0kb) *Eco*RI (#KS)	697	693	CFT073 prophage P4 integrase (693)	e-value=0.0 bit score =1253 ID= 99%	AE014075.1
E103-*pheV*	#U-#T3 (2.0kb) *Eco*RI (#KS)	215	215	CFT073 prophage P4 integrase (215)	e-value=4e-99 bit score =368 ID= 97%	AE014075.1
E104-*pheV*	#U-#T3 (1.2kb) *Pst*I (#KS)	560	560	CFT073 prophage P4 integrase (560)	e-value=0.0 bit score =941 ID= 96%	AE014075.1
E105-*pheV*	#D-#T3 (1.5kb) *Eco*RI (#KS)	443	443	MG1655 predicted inner membrane lipoprotein (443)	e-value=1e-167 bit score =597 ID= 91%	U00096.2
E105-*pheV*	#D-#T3 (1.3kb) *Bam*HI (#KS)	496	300	MG1655 predicted inner membrane lipoprotein (300)	e-value=5e-122 bit score =446 ID= 93%	U00096.2
E106-*pheV*	#D-#T3 (1.7kb) *Bam*HI (#KS)	707	669	MG1655 predicted inner membrane lipoprotein and glycolate transporter (705)	e-value=0.0 bit score =1083 ID= 94%	U00096.2
E107-*pheV*	#U-#T3 (1.9kb) *Sal*I (#T3)	750	750	*Shigella flexneri* 2a str.301 putative P4-type integrase (750)	e-value=0.0 bit score =1297 ID= 97%	AE005674.1
E108-*pheV*	#D-#T3 (1.5kb) *Bam*HI (#KS)	750	677	MG1655 predicted inner membrane lipoprotein and glycolate transporter (740)	e-value=0.0 bit score =1136 ID= 94%	U00096.2

Strain-tRNA gene	SGSP-PCR amplicon	Total sequence length (bp)	Length of GEI-specific sequence (bp)	Blastn hit(s) of GEI-specific sequence (length of match [bp])	Blastn results	NCBI GeneBank accession number
E109-*pheV*	#U-#T3 (1.2kb) *Eco*RV (#U)	156	156	CFT073 (156)	e-value=2e-70 bit score =272 ID= 98%	AE014075.1
E110-*pheV*	#U-#T3 (2.0kb) *Eco*RI (#KS)	373	373	CFT073 Prophage P4 integrase (373)	e-value=0.0 bit score =676 ID= 99%	AE014075.1
E111-*pheV*	#U-#T3 (2.6kb) *Eco*RV (#U)	712	493	CFT073 hypo. protein (493)	e-value=0.0 bit score =824 ID= 96%	AE014075.1
E102-*ssrA*	#D-#T3 (2.0kb) *Hin*dIII (#KS)	289	289	EAECO42 (289)	e-value=e-160 bit score =565 ID= 99%	In-completed genome sequence
E103-*ssrA*	#D-#T3 (2.0kb) *Hin*dIII (#KS)	295	287	EAECO42 (287)	e-value=e-159 bit score =563 ID= 99%	In-completed genome sequence
E104-*ssrA*	#D-#T3 (2.0kb) *Hin*dIII (#KS)	376	376	EAECO42 (376)	e-value=0.0 bit score =729 ID= 99%	In-completed genome sequence
E105-*ssrA*	#D-#T3 (0.6kb) *Eco*RI (#D)	502	383	CFT073 (442)	e-value=0.0 bit score =789 ID= 98%	AE014075.1
E105-*ssrA*	#D-#T3 (0.3kb) *Sal*I (#D)	281	169	CFT073 (229)	e-value=3e-112 bit score =412 ID= 99%	AE014075.1
E106-*ssrA*	#D-#T3 (0.5kb) *Eco*RI (#D)	516	383	CFT073 (444)	e-value=0.0 bit score =793 ID= 98%	AE014075.1
E106-*ssrA*	#D-#T3 (0.3kb) *Sal*I (#D)	280	168	CFT073 (228)	e-value=2e-109 bit score =403 ID= 98%	AE014075.1
E106-*ssrA*	#D-#T3 (1.4kb) *Pst*I (#D)	750	580	CFT073 (640)	e-value=0.0 bit score =1138 ID= 98%	AE014075.1
E106-*ssrA*	#D-#T3 (1.5kb) *Hin*dIII (#D)	723	580	CFT073 (638)	e-value=0.0 bit score =1134 ID= 98%	AE014075.1
E107-*ssrA*	#U-#T3 (1.0kb) *Hin*dIII (#KS)	834	3	CFT073 (700)	e-value=0.0 bit score =1014 ID= 92%	AE014075.1

Strain-tRNA gene	SGSP-PCR amplicon	Total sequence length (bp)	Length of GEI-specific sequence (bp)	Blastn hit(s) of GEI-specific sequence (length of match [bp])	Blastn results	NCBI GeneBank accession number
E108-*ssrA*	#D-#T3 (0.6kb) *Eco*RI (#D)	501	383	CFT073 (441)	e-value=0.0 bit score =787 ID= 98%	AE014075.1
E108-*ssrA*	#D-#T3 (0.3kb) *Sal*I (#D)	279	169	CFT073 (227)	e-value=4e-111 bit score =409 ID= 99%	AE014075.1
E108-*ssrA*	#D-#T3 (1.4kb) *Pst*I (#D)	750	580	CFT073 (638)	e-value=0.0 bit score =1134 ID= 98%	AE014075.1
E108-*ssrA*	#D-#T3 (1.5kb) *Hin*dIII (#D)	750	579	CFT073 (638)	e-value=0.0 bit score =1125 ID= 98%	AE014075.1
E109-*ssrA*	#U-#T3 (1.1kb) *Eco*RI (#KS)	693	4	MG1655 (541)	e-value=0.0 bit score =835 ID= 94%	U00096.2
				MG1655 (173)	e-value=2e-52 bit score =215 ID= 89%	
E110-*ssrA*	#U-#T3 (1.0kb) *Sal*I (#T3)	811	152	CFT073 (583)	e-value=0.0 bit score =935 ID= 95%	AE014075.1
E111-*ssrA*	#D-#T3 (0.6kb) *Eco*RI (#D)	513	383	CFT073 (442)	e-value=0.0 bit score =789 ID= 98%	AE014075.1
E111-*ssrA*	#D-#T3 (0.4kb) *Sal*I (#D)	279	169	CFT073 (227)	e-value=4e-111 bit score =409 ID= 99%	AE014075.1
E111-*ssrA*	#D-#T3 (1.4kb) *Hin*dIII (#D)	750	592	CFT073 (650)	e-value=0.0 bit score =1125 ID= 97%	AE014075.1

Table 8.4
Putative t-RNA sites for PAI by t-RIP-PCR and the results of the SGSP-PCR

#	t-RNA site	SGSP-PCR products within a putative PAI Size of SGSP-PCR product U# /T3#									seq to walk into GI
		Restriction Enzymes									
		EcoR I	Sal I	Pst I	Hind III	BamH I	Sau3 I	EcoRV	HincII	Kpn I	
	K12/metV					1.8					
1	E104/metV	N/PCR	N/PCR	N/D	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
2	E105/metV	N/PCR	0.9	N/D	1	3.2 & 2	N/D	N/D	N/D	N/D	
3	E106/metV	N/PCR	N/PCR	N/D	1	3.2	N/D	N/D	N/D	N/D	
4	E108/metV	2	N/PCR	N/D	1	3.2	N/D	N/D	N/D	N/D	481 bp
5	E109/metV	2	N/PCR	N/D	1	3.2	N/D	N/D	N/D	N/D	
6	E110/metV	2	3.6	N/D	1	3.2	N/D	N/D	N/D	N/D	
7	E111/metV	2	N/PCR	N/D	1	3.2 & 2	N/D	1.4	0.7	N/D	
	K12/asnV					3					
8	E105/asnV	0.7	N/PCR	N/D	3.5	3	N/D	N/D	N/D	N/D	
9	E106/asnV	0.7	N/PCR	N/D	3	N/PCR	N/D	N/D	N/D	N/D	
10	E108/asnV	0.7	N/PCR	N/D	3	N/PCR	N/D	N/D	N/D	N/D	1376 bp
11	E109/asnV	0.7	N/PCR	N/D	3	N/PCR	N/D	N/D	N/D	N/D	
12	E110/asnv	0.7	N/PCR	N/D	3	N/PCR	N/D	N/D	N/D	N/D	
	K12/serU				0.5						
13	E105/serU	N/PCR	1	N/D	N/PCR	0.5	N/D	N/D	N/D	N/D	
14	E108/serU	N/PCR	1	N/D	N/PCR	0.5	N/D	N/D	N/D	N/D	200bp
15	E109/serU	2 faint	1	N/D	N/PCR	0.5	N/D	N/D	N/D	N/D	
16	E111/serU	N/PCR	6 faint	N/D	N/PCR	2	N/D	N/D	N/D	N/D	
	K12/aspV				1						
17	E102/aspV	N/PCR	1	N/PCR	1	1	N/D	N/D	N/D	N/D	
18	E103/aspV	N/PCR	1	N/PCR	1	1	N/D	N/D	N/D	N/D	
19	E104/aspV	N/PCR	1	N/PCR	1	1	N/D	N/D	N/D	N/D	
20	E105/aspV	1	1	N/PCR	1	1	0.9	N/PCR	N/PCR	N/D	
21	E106/aspV	1	1	N/PCR	1	1	N/D	N/D	N/D	N/D	550 bp
22	E107/aspV	N/PCR	1	N/PCR	1	1	N/D	N/D	N/D	N/D	
23	E108/aspV	1	1	N/PCR	1	1	N/D	N/D	N/D	N/D	
24	E109/aspV	T/R	T/R	N/PCR	T/R	T/R	N/D	N/D	N/D	N/D	
25	E110/aspV	1.5	1	N/PCR	2	1	N/D	N/D	N/D	N/D	
26	E111/aspV	T/R	T/R	N/PCR	T/R	T/R	0.9	N/PCR	N/PCR	N/D	
	K12/serW	1.4			N/PCR						
27	E104/serW	N/PCR	1.2 &1.4	1.9	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
28	E105/serW	N/PCR	1.2	N/PCR	N/PCR	N/PCR	N/PCR	1.9	2.1	N/D	
29	E106/serW	N/PCR	1.2	N/PCR	N/PCR	N/PCR	N/D	1.9	2.2	N/PCR	676 bp
30	E108/serW	N/PCR	1.2	N/PCR	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
31	E111/serW	N/PCR	1.2	N/PCR	N/PCR	N/PCR	N/PCR	2	2.2	N/D	
32	E105/serT	0.6	0.7	N/PCR	1.2	2.3	N/D	N/D	N/D	N/D	170 bp
	K12/argW	1.5									
33	E102/argW	3.7	N/PCR	N/PCR	N/PCR	1.8	N/D	N/D	N/D	N/D	
34	E103/argW	3.7	N/PCR	0.5	N/PCR	1.9	N/D	N/D	N/D	N/D	
35	E104/argW	N/PCR	N/PCR	0.5	4	N/PCR	N/D	N/D	N/D	N/D	328 bp
36	E105/argW	3.8	N/PCR	0.4	N/PCR	1.6	N/D	N/D	N/D	N/D	
37	E106/argW	N/PCR	2	N/PCR	N/PCR	N/PCR	N/D	3	0.7	N/PCR	
38	E107/argW	0.88	1.3	0.4	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
39	E110/argW	N/D	N/D	N/D	N/D	N/D	N/D	1.5	0.5	N/PCR	
	K12/pheV										
40	E102/pheV	2	N/PCR	N/PCR	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
41	E103/pheV	2	N/PCR	0.1 & 1.2	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
42	E104/pheV	2	N/PCR	1.2	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
43	E105/pheV	0.5	2.5	1.9	1.9	N/PCR	N/D	N/D	N/D	N/D	293 bp
44	E106/pheV	N/PCR	N/PCR	N/PCR	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
45	E107/pheV	N/PCR	1.9	N/PCR	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
46	E108/pheV	N/PCR	2.8	2	2	N/PCR	N/D	N/D	N/D	N/D	
47	E109/pheV	0.5	N/PCR	N/PCR	N/PCR	N/PCR	N/D	1.2	N/PCR	N/PCR	
48	E110/pheV	2	2.5	1.1	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
49	E111/pheV	N/PCR	N/PCR	2.1	2.1	N/PCR	N/D	2.6	N/PCR	N/D	
	K12/serX					N/PCR					

#	t-RNA site	SGSP-PCR products within a putative PAI Size of SGSP-PCR product D# /T3#									seq to walk into GI
		Restriction Enzymes									
		EcoR I	Sal I	Pst I	Hind III	BamH I	Sau3 I	EcoRV	HincII	Kpn I	
	K12/metV		0.6								
1	E104/metV	N/PCR	0.6		N/PCR	2.5	N/D	N/D	N/D	N/D	
2	E105/metV	N/PCR	0.6		N/PCR	3	N/D	N/D	N/D	N/D	
3	E106/metV	N/PCR	0.6		N/PCR	3	N/D	N/D	N/D	N/D	
4	E108/metV	N/PCR	0.6		N/PCR	2.5	N/D	N/D	N/D	N/D	898 bp
5	E109/metV	N/PCR	0.6		N/PCR	2.5	N/D	N/D	N/D	N/D	
6	E110/metV	N/PCR	0.6		N/PCR	3	N/D	N/D	N/D	N/D	
7	E111/metV	N/PCR	0.6		N/PCR	2.5	N/D	1.4	0.6	N/D	
	K12/asnV				2.5						
8	E105/asnV	N/PCR	N/PCR		N/PCR	N/PCR	N/D	N/D	N/D	N/D	
9	E106/asnV	N/PCR	N/PCR		1.5	N/PCR	N/D	N/D	N/D	N/D	
10	E108/asnV	N/PCR	N/PCR		1.5	N/PCR	N/D	N/D	N/D	N/D	602 bp
11	E109/asnV	N/PCR	N/PCR		1.5	N/PCR	N/D	N/D	N/D	N/D	
12	E110/asnV	N/PCR	N/PCR		1.5	N/PCR	N/D	N/D	N/D	N/D	
	K12/serU				1.5						
13	E105/serU	N/PCR	N/PCR		N/PCR	N/PCR	N/D	N/D	N/D	N/D	
14	E108/serU	N/PCR	N/PCR		N/PCR	N/PCR	N/D	N/D	N/D	N/D	1590 bp
15	E109/serU	N/PCR	N/PCR		N/PCR	N/PCR	N/D	N/D	N/D	N/D	
16	E111/serU	N/PCR	N/PCR		N/PCR	N/PCR	N/D	N/D	N/D	N/D	
	K12/aspV	2.5 D#2			2.2						
17	E102/aspV	N/PCR	N/PCR	N/PCR	1.5 D#2	N/PCR	N/D	N/PCR	3	N/PCR	
18	E103/aspV	N/PCR	N/PCR	N/PCR	N/PCR	N/PCR	N/D	N/PCR	N/PCR	N/PCR	
19	E104/aspV	3	N/PCR	N/PCR	0.6	3.5	N/D	N/PCR	1.6	0.6	
20	E105/aspV	N/PCR	0.5	2	N/PCR	N/PCR	N/D	N/PCR	N/PCR		
21	E106/aspV	N/PCR	N/PCR	0.4 D#2	1.5 D#2	N/PCR	N/D	2.5	2.6	N/PCR	215 bp
22	E107/aspV	1.9	N/PCR	N/PCR	1.9	N/PCR	N/D	N/D	N/D	N/D	
23	E108/aspV	N/PCR	0.5 & 0.8	0.4 D#2	3	1.9	N/D	2.5	2.6	N/PCR	
24	E109/aspV	N/PCR	N/PCR	N/PCR	3	1.9	N/D	2.5	2.6	N/PCR	
25	E110/aspV	N/PCR	N/PCR	N/PCR	2.2	N/PCR	N/D	N/D	2	N/D	
26	E111/aspV	N/PCR	N/PCR	N/PCR	3	N/PCR	N/D	2.2	2.3	N/D	
	K12/serW				1.7						
27	E104/serW	N/PCR	0.6	0.6	1.7	N/PCR	N/D	N/D	N/D	N/D	
28	E105/serW	0.8 &2	3	2.3	1.7	N/PCR	N/PCR	multi	multi	N/D	
29	E106/serW	N/PCR	0.5 &3	2.4	N/PCR	N/PCR	N/D	multi	multi	N/PCR	746 bp
30	E108/serW	0.8	0.7	2.3	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
31	E111/serW	N/PCR	N/PCR	2.3	T/R	T/R	N/D	N/D	N/D	N/D	
32	E105/serT	N/PCR	1.3	1.2	N/PCR	2	N/D	N/D	N/D	N/D	136 bp
	K12/argW				1.5						
33	E102/argW	N/PCR	N/PCR	N/PCR	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
34	E103/argW	N/PCR	N/PCR	1	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
35	E104/argW	N/PCR	1.5	N/PCR	3.4	2.4	N/D	N/D	N/D	N/D	822 bp
36	E105/argW	N/PCR	N/PCR	1	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
37	E106/argW	N/PCR	N/PCR	N/PCR	N/PCR	N/PCR	N/D	1.6	3.7	N/PCR	
38	E107/argW	N/PCR	1.4	N/PCR	N/PCR	2.3	N/D	N/D	N/D	N/D	
39	E110/argW	N/D	N/D	N/D	N/D	N/D	N/D	1.6	1.5	N/PCR	
	K12/pheV		2								
40	E102/pheV	N/PCR	2.4	N/PCR	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
41	E103/pheV	N/PCR	2.5	1.7 &4.4	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
42	E104/pheV	N/PCR	1.8	1.7	2.7	N/PCR	N/D	N/D	N/D	N/D	
43	E105/pheV	1.5	1.7 & 2	N/PCR	N/PCR	1.3	N/D	N/D	N/D	N/D	715 bp
44	E106/pheV	N/PCR	&1.5&2	N/PCR	N/PCR	1.7	N/D	N/D	N/D	N/D	
45	E107/pheV	N/PCR	1.8&2.5	N/PCR	N/PCR	N/PCR	N/D	N/D	N/D	N/D	
46	E108/pheV	1.9	1.4&2.4	N/PCR	N/PCR	1.5	N/D	N/D	N/D	N/D	
47	E109/pheV	1.6	multi	multi	multi	1.3	N/D	N/PCR	1.9	3	
48	E110/pheV	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	
49	E111/pheV	N/D	N/D	N/D	N/D	N/D	N/D	N/PCR	1.4 & 1.6	N/D	
	K12/serX		0.3								

50	E102/serX	1.6	0.9	N/PCR	3	N/PCR	N/D	N/D	N/D	N/D		50	E102/serX	T/R	T/R	T/R	T/R	T/R	N/D	N/D	N/D	N/D	
51	E103/serX	1.6	0.9	N/PCR	N/PCR	N/PCR	N/D	N/D	N/D	N/D		51	E103/serX	T/R	T/R	T/R	T/R	T/R	N/D	N/D	N/D	N/D	
52	E104/serX	N/PCR	0.4	N/PCR	0.7	N/PCR	N/D	0.9	0.2	N/PCR	266 bp	52	E104/serX	T/R	T/R	T/R	T/R	T/R	N/D	multi	multi	1.5	103 bp
53	E105/serX	N/PCR	N/PCR	1.5	N/PCR	N/PCR	N/D	N/D	N/D	N/D		53	E105/serX	T/R	T/R	T/R	N/D	N/D	N/D	N/D	N/D	N/D	
54	E106/serX	N/PCR	N/PCR	1.5	N/PCR	N/PCR	N/D	1.3	0.2	1.2		54	E106/serX	N/D	N/D	N/D	N/D	N/D	N/D	multi	multi	0.5 & 1.2	
55	E110/serX	N/PCR	N/PCR	1.6	N/PCR	N/PCR	N/D	1.5	0.2	1.2		55	E110/serX	N/D	N/D	N/D	N/D	N/D	N/D	multi	multi	0.5 & 1.2	
56	E111/serX	N/PCR	0.5	N/PCR	0.7	N/PCR	N/D	N/D	N/D	N/D		56	E111/serX	N/D	N/D	N/D	N/D	N/D	N/D				
	k12/asnT			T/R									k12/asnT				1.5						
57	E102/asnt	T/R	T/R	T/R	T/R	T/R	N/D	N/D	N/D	N/D		57	E102/asnt	1.5 & 2.5	1	1.5	1.5	1.5	N/D	N/D	N/D	N/D	
58	E103/asnT	T/R	T/R	T/R	T/R	T/R	N/D	N/D	N/D	N/D		58	E103/asnT	2.5	1 & 1.5	N/PCR	1.5	1.5	N/D	N/D	N/D	N/D	
59	E104/asnT	T/R	T/R	T/R	T/R	T/R	N/D	N/D	N/D	N/D	462 bp	59	E104/asnT	N/PCR	0.5 & 1.5	1.5	1.5	1.5	N/D	N/D	N/D	N/D	58 bp
60	E105/asnT	T/R	T/R	T/R	N/D	N/D	N/D	N/D	N/D	N/D		60	E105/asnT	1.5 & 2.5	1 & 1.5	1 & 1.5	1.6	1.6	N/D	N/D	N/D	N/D	
61	E106/asnT	N/D	N/D	N/D	N/D	N/D	N/D	multi	multi	multi		61	E106/asnT	1.6	1.2&1.6	1.6	1.6	1.6	N/D	1.6	1.5	1.6	
62	E107/asnT	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D		62	E107/asnT	1.6&2.7	1.2&1.6	0.7&1.6	1.6	1.1&1.6	N/D	N/D	N/D	N/D	
63	E110/asnT	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D		63	E110/asnT	1.6&2.6	1.1&1.6	1.1&1.6	1.6	1.1&1.6	N/D	N/D	N/D	N/D	
64	E111/asnT	N/D	N/D	N/D	N/D	N/D	N/D	multi	0.4	N/D		64	E111/asnT	1.6&2.6	1.2	1.6	1.6	1.3 & 1.6	N/D	1.6	1.2	N/D	
	K12/thrW	N/PCR											K12/thrW						N/D	N/D	N/D	N/D	
65	E103/thrW	N/PCR	2.3	N/PCR	2.3	N/PCR	N/D	N/D	N/D	N/D		65	E103/thrW	multi	multi	multi	multi	multi	N/D	N/D	N/D	N/D	
62	E105/thrW	N/PCR	0.7	0.7 & 2.4	0.7	N/PCR	N/D	0.9	2.5	N/D	144 bp	62	E105/thrW	multi	multi	multi	multi	multi	N/D	N/PCR	N/PCR	N/D	200 bp
63	E107/thrW	N/PCR	2.4	2.4	0.5	N/PCR	N/D	N/D	N/D	N/D		63	E107/thrW	multi	multi	multi	multi	multi	N/D	N/D	N/D	N/D	
66	E110/thrW	N/PCR	5	2.4	0.8	N/PCR	N/D	N/D	N/D	N/D		66	E110/thrW	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	
67	E111/thrW	N/PCR	2.4	N/PCR	N/PCR	4.7	N/D	1.2	2.2	N/D		67	E111/thrW	N/D	N/D	N/D	N/D	N/D	N/D	1.7	N/PCR	N/D	
	K12/selC	1.2											K12/selC	1.7									
68	E105/selC	N/PCR	1.5	N/PCR	2.5	N/PCR	N/D	N/D	N/D	N/D	775 bp	68	E105/selC	multi	multi	multi	multi	multi	N/D	N/D	N/D	N/D	566 bp
69	E109/selC	N/PCR	1.5	N/PCR	N/D	N/D	N/D	N/D	N/D	N/D		69	E109/selC	multi	multi	multi	N/D	N/D	N/D	N/D	N/D	N/D	
70	E110/selC	N/D	N/D	N/D	N/D	N/D	N/D	0.2	0.4	2.3		70	E110/selC	N/D	N/D	N/D	N/D	N/D	N/D	2.2	2	multi	
	K12/ssrA	1.3			1.2								K12/ssrA	D#3 0.7			1.6						
71	E102/ssrA	N/PCR	N/PCR	N/PCR	1.7	2.3	N/D	N/D	N/D	N/D		71	E102/ssrA	N/PCR	N/PCR	N/PCR	2	N/PCR	N/D	N/D	N/D	N/D	
72	E103/ssrA	N/PCR	N/PCR	N/PCR	1.7	2.3	N/D	N/D	N/D	N/D		72	E103/ssrA	N/PCR	N/PCR	N/PCR	2	N/PCR	N/D	N/D	N/D	N/D	
73	E104/ssrA	N/PCR	N/PCR	N/PCR	N/PCR	2.3	N/D	N/D	N/D	N/D		73	E104/ssrA	N/PCR	N/PCR	N/PCR	2	N/PCR	N/D	N/D	N/D	N/D	
74	E105/ssrA	N/PCR	N/PCR	1.7	N/PCR	N/PCR	N/PCR	0.25	1.6	N/D	807 bp	74	E105/ssrA	0.6	0.3	1.3	N/PCR	N/PCR	N/PCR	1.3 D# 0	0.6 D#3	N/D	182 bp
75	E106/ssrA	N/PCR	N/PCR	N/PCR	N/PCR	N/PCR	N/D	0.25	N/PCR	N/PCR		75	E106/ssrA	0.5	0.3	1.4	1.5	N/PCR	N/D	D#3 1.6	D#3 0.3	D#3 0.9	
76	E107/ssrA	N/PCR	N/PCR	2.3	1	N/PCR	N/D	N/D	N/D	N/D		76	E107/ssrA	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	
77	E108/ssrA	N/PCR	N/PCR	N/PCR	N/PCR	N/PCR	N/D	0.25	N/PCR	N/PCR		77	E108/ssrA	0.6	0.3	1.4	1.5	N/PCR	N/D	D#3 1.6	D#3 0.3	D#3 0.9	
78	E109/ssrA	1.1	1.2	N/PCR	N/PCR	N/PCR	N/D	N/D	N/D	N/D		78	E109/ssrA	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	
79	E110/ssrA	N/PCR	1	N/PCR	N/PCR	N/PCR	N/D	N/D	N/D	N/D		79	E110/ssrA	N/PCR	N/PCR	N/D	N/D	N/D	N/D	N/D	N/D	N/D	
80	E111/ssrA	N/D	N/D	N/D	N/D	N/D	N/D	N/PCR	N/PCR	N/D		80	E111/ssrA	0.6	0.4	N/PCR	1.4	PCR D#3 1.2		N/PCR D3#	PCR D#3	N/D	
	K12/glyU	0.9			2								K12/glyU			2.8							
81	E102/glyU	N/PCR	N/PCR	N/PCR	N/PCR	N/PCR	N/D	multi	0.4 & 0.6	N/PCR		81	E102/glyU	N/D	N/D	N/D	N/D	N/D	N/D	0.6	0.2	N/PCR	
82	E104/glyU	N/PCR	N/PCR	2.3	2.5	N/PCR	N/D	N/D	N/D	N/D		82	E104/glyU	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	
83	E107/glyU	N/PCR	2	N/PCR	2.7	N/PCR	N/D	N/D	N/D	N/D		83	E107/glyU	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	
	K12/leuX			2.6									K12/leuX			0.8							
84	E102/leuX	N/PCR	1.2	N/PCR	N/PCR	N/PCR	N/D	N/D	N/D	N/D		84	E102/leuX	3	0.7	0.7	1.5	N/PCR	N/D	N/D	N/D	N/D	
85	E103/leuX	N/PCR	0.8	2.9	N/PCR	N/PCR	N/D	N/D	N/D	N/D		85	E103/leuX	3	0.7	0.7	1.5	N/PCR	N/D	N/D	N/D	N/D	
86	E104/leuX	N/PCR	1.4	1.9&3	N/PCR	N/PCR	N/D	N/D	N/D	N/D	1386 bp	86	E104/leuX	N/PCR	0.7	0.7	1.5	N/PCR	N/D	N/D	N/D	N/D	828 bp
87	E105/leuX	N/PCR	1.8&3	1.4			N/D	0.3	2.3	N/D		87	E105/leuX	3	0.7	0.7	N/PCR	N/PCR	N/D	1	0.4	N/D	
88	E106/leuX	N/PCR	2.6	1.2	3	N/PCR	N/D	N/D	N/D	N/D		88	E106/leuX	N/PCR	0.7	N/PCR	3	N/PCR	N/D	N/D	N/D	N/D	
89	E107/leuX	1.5	1	N/PCR	N/PCR	N/PCR	N/D	N/D	N/D	N/D		89	E107/leuX	N/PCR	0.7	N/PCR	N/PCR	0.8	N/D	N/D	N/D	N/D	
90	E110/leuX	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D		90	E110/leuX	3	0.7	1.5	2.5	0.7	N/D	N/D	N/D	N/D	
91	E111/leuX	N/PCR	N/PCR	1.2	N/PCR	N/PCR	N/D	0.3	1.5	N/D		91	E111/leuX	3	0.7	1.2	N/D	N/D	N/D	1	0.4	N/D	

[a] N/PCR: No PCR product; N/D: Not done; T/R: The PCR result is to be repeated