

Characterising the mobile genome of *Shigella*

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

by

James David Lonnen BSc (Hons)
Department of Infection, Immunity and Inflammation
University of Leicester

December 2007

Statement of Originality

This accompanying thesis submitted for the degree of PhD entitled “Characterising the mobile genome of *Shigella*” 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 October 2003 and October 2006.

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

My sincere gratitude goes to Dr. Kumar Rajakumar for his unwavering enthusiasm, support and for being a true inspiration throughout my project and to Prof. Mike Barer for his continual support, advice and words of wisdom. Also, thanks to both Kumar's and Mike's families for some great days out.

A special thank you to Dr. Primrose Freestone for her compassion, advice and amazing attitude to science.

A huge thanks to Dr. Rebecca Smith, Dr. Natalie Garton, Dr. Hong-Yu Ou, Dr. Richard Haigh, Dr. Simon Kilvington, Dr. Martha Clokie, Dr. Jonathan Hales, Sheila, Pam, Mo and Elizabeth for all of their help over the last four years.

Thanks to all of my Lab 136/212 fellows past and present and to all of my other friends/colleagues at the University of Leicester, especially Niran, Wayne, Rob Free, Paddy, Ewan, Barbara, Anna, Helen, Adam, Jeni, Cordula, Claire, Sarah, Sonia and Sergio, Aline, Kim, Mafalda, Vitor and Rob Hardwick.

To Katerina and Tomis, thanks for all of your love, time and pastitsio.

To Jon and Kate, thank you so much for your friendship and all of the wonderful times.

Thanks to Ed and Dave, for being fantastic mates.

To the Tfenkgi family, thank you all so much for your love and support.

Cheers to all my friends from back home, especially to Tony and Rita, Mike and Rachael and Addam.

Merci ktir to Amo George, Tante Laudy, Amo Joseph, Tante Elise and the entire El-Rachkidy family, love you all, can't wait to be with you again.

Thanks to my sister Kate, her husband Dave and my beautiful niece Jasmine for the lovely times back home.

To my parents, I cannot thank you enough, for all the love, advice, for always being there and for supporting my every decision, you are the best.

Most importantly, to Rana, my beautiful wife, thank you so much for all of your advice and support throughout my project, thank you albi for being so patient over the last few months, inti hyete and I love you more than words can say.

Abstract

Shigella spp. are pathogenic variants of *Escherichia coli* that cause bacillary dysentery, resulting in over 1 million deaths per year. Across *E. coli* bacteria, the dynamic process of acquisition and loss of GIs, especially those associated with virulence (pathogenicity islands [PAIs]) is a driving force behind the emergence of new pathogenic strains. In *Shigella* only 5 GIs have been well characterised and there is currently no effective vaccine. Therefore the development of an efficient screen to detect GIs in unsequenced *Shigella* strains could be highly informative.

Nineteen *Shigella* strains were probed for the presence of GIs using a high throughput PCR screen (tRIP), across 16 tRNA gene integration hotspots. Putative GIs were then investigated using a chromosome walking technique (SGSP-PCR). Representative PCR amplicons were sequenced to get a snapshot of the islands contents. Islands of particular interest were characterised further using allelic exchange and marker rescue to capture clones that harbour larger portions of the GI and subsequent sequencing and analysis.

Using SGSP-PCR, 81% of the putative GI occupied tRNA loci were characterised, and sequencing analysis found they all contain island DNA, indicating that tRIP followed by SGSP-PCR is a robust strategy for GI discovery in unsequenced strains; also this method should be applicable to a broad range of microorganisms.

At least 54% of the islands identified harbour phage-like integrase genes, strongly supporting the notion that many of these elements arose following acquisition of horizontally acquired integrative GIs. The frequent presence of integrase genes also highlights the potential role of bacteriophage in the original and/or ongoing dissemination of island DNA in *Shigella*.

Only one novel GI was discovered; it has classic prophage-like features and contains completely novel DNA, indicating that while *Shigella* has a plastic genome, it is a highly specialised human pathogen that has undergone considerable pathoadaptive genome reduction.

The major development from this study is evidence that a number of key *Shigella* virulence determinants are independently mobile and not only localised to a single family of islands; this significantly increases their potential to spread by HGT across *Shigella* and could contribute to the rapid emergence of new endemic strains.

Abbreviations

%	Percent
°C	Degrees Centigrade
μg	Micrograms
μl	Microlitres
μM	Micromolar
μF	Microfaradays
Ap	Ampicillin
Ap ^r	Ampicillin resistant
ATP	Adenosine 5'-triphosphate
bp	Base pairs
Cm	Chloramphenicol
Cm ^r	Chloramphenicol resistant
CTAB	Cetrimonium bromide
DNA	Deoxyribonucleic acid
ddH ₂ O	Double-distilled water
dNTPs	Deoxynucleosides
DR	Direct repeat
EDTA	Ethylenediamine tetraacetic acid
ExPEC	Extraintestinal Pathogenic <i>Escherichia coli</i>
fg	Femtograms
g	Grams
<i>g</i>	Relative centrifugal force
HGT	Horizontal gene transfer
hr	Hours
ID	Identity
IPTG	Isopropyl-β-D-thiogalactosylpyranoside
kb	Kilobases
Km	Kanamycin
Km ^r	Kanamycin resistant
kV	Kilovolts
l	Litres
LA	Luria agar
LB	Luria broth
LPS	Lipopolysaccharide

M	Molar
MCS	Multiple cloning site
mg	Milligrams
min	Minutes
ml	Millilitres
mm	Millimetres
mM	Millimolar
ng	Nanograms
NaCl	Sodium chloride
nH ₂ O	Nanopure water
OD	Optical density
PCR	Polymerase Chain Reaction
Pg	Picograms
psi	Pounds per square inch
RNA	Ribonucleic acid
rpm	Revolutions per minute
sec	Seconds
Sm	Streptomycin
Sm ^r	Streptomycin resistant
Tc	Tetracycline
Tc ^r	Tetracycline resistant
TE	Tris-EDTA
tRIP	tRNA site interrogation for PAIs, prophages and other GIs
SDS	Sodium Dodecyl Sulfate
SGSP-PCR	Single genome-specific primer-PCR
SH	Subtractive hybridisation
v/v	By volume
w/v	Weight by volume
X-gal	5-bromo-4-chloro-3-indolyl-b-D-galactopyranoside
Ω	Ohms

Contents

Title Page	i
Statement of Originality	ii
Acknowledgement	iii
Abstract	iv
Abbreviations	v

Table of contents

1.0	Introduction.....	1
1.1	The core and mobile genomes	1
1.2	Genomic islands.....	1
1.3	Pathogenicity islands	2
1.4	Structure of genomic islands.....	3
1.5	Horizontal transfer of genomic islands.....	5
1.5.1	Natural transformation.....	5
1.5.2	Conjugation.....	5
1.5.3	Transduction	6
1.6	tRNA loci as integration sites for genomic islands.....	8
1.7	Signatures used to discover genomic islands <i>in silico</i>	10
1.8	Genomic island discovery in unsequenced strains	12
1.9	<i>Shigella</i>	14
1.10	Epidemiology.....	15
1.11	<i>Shigella</i> Pathogenesis	16
1.12	Genomic islands in <i>Shigella</i>	18
1.13	Treatment of <i>Shigella</i> infections.....	20
1.14	The aims of this study	21
2.0	Materials and methods	22
2.1	Bioinformatics	22

2.1.1	<i>E. coli</i> genomes analysed.....	22
2.1.2	tRNA loci chosen to study	22
2.1.3	tRNA core flanking regions.....	23
2.2	tRNA site Interrogation for PAIs, prophages and other GIs (tRIP).....	24
2.2.1	The principle of tRIP	24
2.2.2	Multiple sequence alignments and primer design.....	25
2.2.3	tRIP Primer locations.....	26
2.3	<i>Shigella</i> , <i>E. coli</i> strains and plasmids used in this study.....	28
2.3.1	<i>Shigella</i> strains chosen for the tRIP screen.....	28
2.4	Genomic DNA extraction	32
2.5	Plasmid extraction.....	33
2.5.1	Modified alkaline lysis method.....	33
2.5.2	GenElute™ miniprep kit (Sigma).....	33
2.6	Restriction endonuclease digestion.....	34
2.7	DNA ligation.....	34
2.8	Polymerase Chain reaction (PCR)	35
2.8.1	Standard PCR.....	36
2.8.2	Hot-start, ‘touchdown’ PCR	36
2.8.3	Band-stab PCR.....	37
2.8.4	Colony PCR	37
2.8.5	Splicing by overlap extension (SOE) PCR.....	37
2.9	Agarose gel electrophoresis	38
2.10	Gel extraction.....	38
2.11	DNA cleanup from agarose gel	38
2.12	DNA sequencing.....	39
2.13	Preparation of eletrocompetent cells.....	39
2.13.1	Large-scale preparation of <i>E. coli</i> cells	39

2.13.2	Small-scale preparation of <i>E. coli</i> and <i>Shigella</i> cells.....	39
2.14	Electroporation.....	40
2.15	Southern hybridization.....	40
2.15.1	Restriction digestion and gel electrophoresis	40
2.15.2	Southern transfer.....	40
2.15.3	Labelling of the probe.....	41
2.15.4	Prehybridisation and hybridisation	42
2.15.5	Detection.....	42
2.16	Allelic exchange	43
2.16.1	Conjugation by filter mating.....	44
2.16.2	Sucrose selection.....	44
2.17	Bacterial identification using the API [®] 20 E test system (bioMeriueX.Inc)	45
3.0	Development and optimisation of the tRIP screen	46
3.1	tRIP	46
3.1.1	tRIP sensitivity and optimisation.....	46
3.2	tRIP results.....	51
3.2.1	<i>In Silico</i> tRIP results	51
3.2.2	Orientation of flanking regions.....	53
3.2.3	tRIP screen results across the <i>Shigella</i> strains	54
4.0	Characterisation of the extremities of tRNA associated islands identified by tRIP.	57
4.1	Interrogation of tRNA borne GIs using SGSP-PCR.....	57
4.1.1	Enzyme choice	58
4.2	<i>In silico</i> SGSP-PCR	60
4.3	SGSP-PCR regime	60
4.3.1	Touchdown PCR.....	62
4.3.2	Optimal vector primer choice	62

4.3.3	Use of hot-start <i>Taq</i> polymerase	63
4.4	SGSP-PCR after optimisation.....	65
4.5	Band-stab PCR.....	65
4.6	Negative SGSP-PCRs	66
4.6.1	Additional libraries	66
4.6.2	<i>Int</i> -PCR	66
4.6.3	Southern hybridisation	66
4.7	Sequencing and analysis	66
4.7.1	tRIP PCR amplicon sequencing.....	67
4.7.2	SGSP-PCR amplicon sequencing	67
4.7.3	Sequence runs	68
4.8	Blast searches using the NCBI database	69
5.0	Overall results of the characterisation of island DNA across <i>Shigella</i>	71
5.1	Matrix of results across the <i>Shigella</i> strains	71
5.2	Occupancy of tRNA loci across <i>Shigella</i>	73
5.3	Breakdown of putatively GI occupied tRNA loci	74
5.4	Island families observed across <i>Shigella</i>	76
5.5	Summary of island characterisation across <i>Shigella</i>	76
5.6	S109 (<i>S. flexneri</i> 4a) re-classification to an <i>E. coli</i>	78
5.7	S108 (<i>S. flexneri</i> 3a) re-classification to a <i>S. sonnei</i>	79
6.0	Island characterisation at the <i>argW</i> locus.....	80
6.1	<i>argW</i> overall results	81
6.1.1	<i>S. flexneri</i>	82
6.1.2	<i>S. sonnei</i>	84
6.1.3	The <i>argW</i> associated sucrose metabolism and D-serine metabolism genes	87
6.1.4	The S120 <i>argW</i> D-arm is CFT073-like	88
6.1.5	<i>S. dysenteriae</i>	91

6.2	The S116 (<i>S. boydii</i> 1 strain) <i>argW</i> novel prophage-like GI	93
6.2.1	SGSP-PCR Results	93
6.3	S116 <i>argW</i> island probing	96
6.3.1	Tagging of the <i>argW</i> UF	96
6.3.2	Choice of marker	96
6.3.3	<i>argW</i> UF region chosen for allelic exchange.....	97
6.3.4	Construction of the <i>argW</i> Km ^r tagged UF	99
6.3.5	Insertion and orientation of the Km ^r cassette in the <i>argW</i> UF region.....	100
6.3.6	Insertion and orientation of the mutant <i>argW</i> UF region in pDS132	105
6.3.7	Generation of X102 and X103: ampicillin resistant derivatives of S116	108
6.3.8	X102 allelic exchange.....	109
6.3.9	X102 sucrose selection	110
6.3.10	Screening of X102 potential transconjugants.	110
6.3.11	Refinement of the sucrose selection method	116
6.4	Marker rescue	118
6.4.1	Rescue of clones harbouring the Km ^r cassette.....	119
6.4.2	Restriction map of the X106 <i>argW</i> GI U-arm	122
6.4.3	X106 (S116 Km ^r strain) chromosome walking.....	123
6.4.4	X106 <i>argW</i> GI integrase walking	126
6.4.5	Schematic of the S116 <i>argW</i> associated novel prophage-like GI.....	127
7.0	tRNA loci harbouring GIs that contribute to the virulence of <i>Shigella</i>.....	129
7.1	<i>leuX</i>	130
7.1.1	<i>S. flexneri</i>	131
7.1.2	<i>S. sonnei</i>	133
7.1.3	The <i>fec</i> locus is independently mobile.....	133
7.1.4	<i>leuX</i> island family 4 is a prophage restricted to some <i>S. boydii</i> strains.....	135
7.1.5	<i>S. boydii</i> and <i>S. dysenteriae</i> strains that harbour <i>leuX</i> island family 1	137

7.1.6	Diversity of other <i>leuX</i> associated GIs	142
7.2	A <i>sigA</i> gene is harboured on the Sb227 distal <i>leuX</i> GI.....	142
7.2.1	The <i>sigA</i> gene may be independently mobile	143
7.3	S101 <i>leuX</i> island probing.....	145
7.3.1	<i>leuX</i> UF splicing overlap extension (SOE) PCR	146
7.3.2	S101 allelic exchange	152
7.4	X101 marker rescue	154
7.5	<i>serU</i>	162
7.5.1	The <i>S. flexneri serU</i> associated prophage-like GI	161
7.5.2	The <i>S. boydii</i> and <i>S. dysenteriae</i> strains harbour the same prophage-like GI at the <i>serU</i> locus	163
7.5.3	The <i>S. boydii</i> and <i>S. dysenteriae serU</i> GI D-arms	166
7.6	The S102 (<i>S. dysenteriae</i> 9 strain) novel sequence	170
7.7	Linkage of the <i>ipaH</i> genes with the <i>serU</i> associated prophages in <i>Shigella</i>	172
7.8	The <i>Shigella serU</i> prophage is a selfish element.....	173
7.9	<i>aspV</i>	175
7.9.1	<i>aspV</i> U-arm Southern hybridisation	176
7.9.2	Southern hybridisation results	179
7.9.3	The Sf301 <i>aspV</i> associated GI.....	181
7.9.4	Putative functions of the <i>Shigella aspV</i> associated <i>sciABCDE</i> gene cluster ..	182
7.9.5	The <i>Shigella tauABCD</i> gene cluster	183
7.9.6	The <i>S. flexneri</i> novel <i>Rhs</i> elements	183
7.9.7	<i>S. sonnei</i>	186
7.9.8	The <i>aspV</i> associated GIs are mosaic elements.....	188
7.10	<i>thrW</i>	189
7.10.1	Diverse O-serotype converting bacteriophages are associated with <i>thrW</i> in <i>Shigella</i>	190

7.10.2	The S118 (<i>S. boydii</i> 3) <i>thrW</i> prophage	195
7.10.3	The S107 (<i>S. flexneri</i> 2b strain) <i>thrW</i> associated mosaic prophage.....	197
7.10.4	The <i>Shigella thrW</i> associated integrase genes are found in the opposite orientation	199
8.0	Discussion	200
8.1	tRIP and SGSP-PCR as strategies for GI detection and characterisation.....	200
8.2	GI diversity across <i>Shigella</i>	200
8.3	Key genomic islands identified across <i>Shigella</i> in this study	202
8.4	Independently mobile virulence determinants in <i>Shigella</i>	204
8.5	The <i>aspV</i> associated ‘ <i>sci</i> ’ island could encode novel virulence factors	205
8.6	Improvement of the tRIP screen	205
8.7	The power of tRIP when used with complementary <i>in silico</i> GI discovery methods ..	206
8.8	Future work.....	207
8.9	Conclusion	208
A1.0	Appendix 1	A1-1
A2.0	Appendix 2	A2-1
A2.1	Primers used in this study	A2-1
A2.2	<i>In silico</i> SGSP-PCR results.....	A2-7
A2.3	Details of the different size tRIP amplicons generated.....	A2-9
A2.4	Details SGSP-PCR derived island sequences that are defined by tRIP only.....	A2-12
A2.5	Details of the generation of the <i>leuX</i> UF suicide constructs	A2-16
A2.5.1	Insertion and orientation of the Km ^r cassette in the <i>leuX</i> UF region	A2-16
A2.5.2	Insertion and orientation of the mutant <i>leuX</i> UF region in pDS132.....	A2-18
A2.6	Details of the island characterisation across all other tRNA loci known to be hotspots for GI insertion across <i>E. coli</i>	A2-21
A2.6.1	<i>serW</i>	A2-21

A2.6.2 <i>glyU</i>	A2-29
A2.6.3 S104 (<i>S. flexneri</i> 1a strain) <i>glyU</i> U# Results	A2-39
A2.6.4 <i>pheU</i> and <i>selC</i>	A2-44
A2.6.5 <i>pheV</i>	A2-54
A2.6.6 <i>metV</i>	A2-62
A2.6.7 <i>ssrA</i>	A2-69
A2.6.8 <i>serX</i>	A2-75
A2.6.9 <i>asnT</i>	A2-80
A2.7 <i>asnV</i> Results	A2-86
A2.7.1 Inversion of the <i>asnV</i> UF region in <i>E. coli</i> CFT073	A2-86
A2.7.2 Inversion of the <i>asnV</i> UF region in <i>S. sonnei</i> 046	A2-86
References	I

List of tables

Table 1.1. Examples of genomic islands in bacteria.....	4
Table 1.2. The median of <i>Shigella</i> isolates in developed and developing countries ^a	16
Table 1.3. tRNA associated genomic islands characterised in <i>Shigella</i>	19
Table 2.1 The sixteen tRNA loci designated as hotspots for GI insertion after analysis of four of the completely sequenced <i>E. coli</i> genomes.	23
Table 2.2. Locations of the conserved U and D tRIP primers relative to the putative GI boundary.	28
Table 2.3. <i>Shigella</i> , <i>E. coli</i> strains and plasmids used in this study	29
Table 3.1. <i>In silico</i> tRNA site interrogation for PAIs (tRIP) of the published complete <i>E. coli</i> and <i>Shigella</i> genomes, across the sixteen tRNA loci designated as hotspots for GI insertion.	52
Table 3.2. tRNA site interrogation for PAIs (tRIP) screen across the 16 tRNA loci designated as GI insertion hotspots in twenty unsequenced <i>Shigella</i> strains representative of the four ‘species’ and two sequenced <i>E. coli</i> control strains.....	55

Table 3.3. tRIP screen across the 16 tRNA loci designated as GI insertion hotspots in twenty <i>Shigella</i> strains representative of the four ‘species’, with the tRIP amplicon sizes indicated..	56
Table 4.1. The five enzymes used to generate genomic libraries for SGSP-PCR.....	60
Table 4.2. pBluescript KS II (+) primer matches across the K12 MG1655 chromosome.....	63
Table 4.3. <i>Phred</i> scores and the corresponding base call error probabilities.	68
Table 5.1. Overall results of the characterisation of island DNA in nineteen <i>Shigella</i> strains across 16 tRNA loci that are known hotspots for GI insertion.	72
Table 5.2. Island families found across nineteen <i>Shigella</i> strains at sixteen tRNA loci that are known to be hotspots for GI insertion in <i>E. coli</i>	76
Table 5.3. Site by site breakdown of the tRIP screen results and island characterisation across nineteen <i>Shigella</i> strains at sixteen tRNA loci that are known to be hotspots for GI insertion in <i>E. coli</i>	77
Table 6.1. SGSP-PCR results of the <i>argW</i> tRIP negative strain-tRNA loci.....	81
Table 6.2. Restriction fragments produced by the four possible pJL5/ <i>Nsi</i> I::Km ^r / <i>Nsi</i> I constructs.	103
Table 6.3. <i>Sph</i> I restriction fragments produced by the <i>argW</i> UF region suicide plasmid constructs.	107
Table 7.1. SGSP-PCR results of <i>leuX</i> tRIP negative strain-tRNA loci	130
Table 7.2. Blastn comparison of the Sb227 P4-like integrase gene found 12.1 kb into the <i>leuX</i> associated GI, against the other complete <i>E. coli</i> and <i>Shigella</i> genomes.	140
Table 7.3. SGSP-PCR results of <i>serU</i> tRIP negative strain-tRNA loci.....	162
Table 7.4. SGSP-PCR results of the <i>aspV</i> tRIP negative strain-tRNA loci.....	175
Table 7.5. SGSP-PCR results of the <i>thrW</i> tRIP negative strain-tRNA loci.....	189
Table 8.1. Key <i>Shigella</i> islands identified in this study.....	203
Table A2. 1. tRIP and SGSP-PCR Primers	A2-1
Table A2. 2. Other primers used in this study	A2-4

Table A2. 3. Length of the <i>in silico</i> SGSP-PCR products with the U-T7 primers across the 16 tRNA loci in the <i>E. coli</i> K12 MG1655 chromosome	A2-7
Table A2. 4. Length of the <i>in silico</i> SGSP-PCR products with the D-T7 primers at the 16 tRNA loci in the <i>E. coli</i> K12 MG1655 chromosome	A2-8
Table A2. 5. tRIP-positive amplicons of different length to the original six control <i>E. coli</i> and <i>Shigella</i> strains.....	A2-9
Table A2. 6. Island sequences obtained by sequencing of SGSP-PCR amplicons that are defined as island DNA by the tRIP method only and their corresponding GC contents and nucleotide matches to the <i>E. coli</i> and <i>Shigella</i> genomes available in the NCBI database.	A2-12
Table A2. 7. Occupancy of <i>serW</i> and <i>serX</i> in the original four complete <i>E. coli</i> genomes and the respective integrase locations of each island.	A2-22
Table A2. 8. SGSP-PCR results of the <i>glyU</i> tRIP-negative strain-tRNA loci.....	A2-29
Table A2. 9. SGSP-PCR results of the <i>pheU</i> tRIP-negative strain-tRNA loci.....	A2-44
Table A2. 10. SGSP-PCR results of the <i>selC</i> tRIP-negative strain-tRNA loci	A2-45
Table A2. 11. SGSP-PCR results of the <i>pheV</i> tRIP negative strain-tRNA loci	A2-54
Table A2. 12. SGSP-PCR results of the <i>metV</i> tRIP negative strain-tRNA loci	A2-62
Table A2. 13. Contents of the tRIP defined 6347 bp Sb227 <i>metV</i> islet.....	A2-68
Table A2. 14. SGSP-PCR results of the <i>ssrA</i> tRIP negative strain-tRNA loci.....	A2-69
Table A2. 15. SGSP-PCR results of the <i>serX</i> tRIP negative strain-tRNA-loci.....	A2-75
Table A2. 16. SGSP-PCR results of the <i>asnT</i> tRIP negative strain-tRNA loci.....	A2-80
Table A2. 17 The initial <i>asnV</i> tRIP results showing the fourteen <i>Shigella</i> strains screened with the original U and D primers.	A2-90
Table A2. 18 The final <i>asnV</i> tRIP screen results after screening with the U2 and D2 primers.	A2-92

Table of figures

Figure 1.1. Schematic of a typical genomic island.	3
Figure 2.1. The cognate tRNA loci in a strain that harbours no GI and a strain harbouring a GI.	24
Figure 2.2 tRNA site Interrogation for PAIs and other GIs (tRIP).....	25
Figure 2.3 Multiple sequence alignment (MSA) of the conserved <i>serU</i> upstream flanking regions (UFs) in 6 sequenced genomes.	26
Figure 2.4. Schematic showing the minimum distance from the GI end of the conserved flanking regions that the U and D primers were located at each tRNA locus.	27
Figure 3.1. Agarose gel of an initial <i>serU</i> tRIP screen using over 200 ng of genomic DNA as template.....	47
Figure 3.2. Agarose gel showing the <i>serU</i> tRIP sensitivity assay using different concentrations of <i>E. coli</i> K12 MG1655 genomic DNA as template.	49
Figure 3.3. Agarose gel of the <i>serU</i> tRIP PCR screen after dilution of the template DNA to 10-15 ng/ μ l.....	50
Figure 3.4. Alternative arrangements of the conserved upstream flanking region (UF) and conserved downstream flanking region (DF) relative to each other, found by the <i>in silico</i> tRIP screen of the six <i>E. coli</i> and <i>Shigella</i> genomes.	53
Figure 4.1. Single Genome-Specific Primer-PCR (SGSP-PCR).....	58
Figure 4.2. Distribution of <i>Hind</i> III fragments across the <i>E. coli</i> K12 MG1655 chromosome.	59
Figure 4.3. <i>serU</i> U# SGSP-PCR, showing multiple amplicons.....	61
Figure 4.4. An optimised <i>leuX</i> U#-T7# SGSP-PCR indicating the generation of specific amplicons.	65
Figure 4.5. The system for sequencing of SGSP-PCR amplicons.....	67
Figure 5.1. The proportion of empty and putatively GI occupied tRNA loci as indicated by the tRIP screen at 16 tRNA loci across nineteen <i>Shigella</i> strains.	73

Figure 5.2. The proportions and final classification of each tRIP-negative strain-tRNA locus.	74
Figure 6.1. <i>argW</i> U-arm SGSP-PCR results for the <i>S. flexneri</i> strains.	83
Figure 6.2. <i>argW</i> SGSP-PCR results for the <i>S. sonnei</i> strains.	85
Figure 6.3. <i>argW</i> D-arm SGSP-PCR results for S120 (<i>S. boydii</i> 7 strain)	90
Figure 6.4. <i>argW</i> SGSP-PCR results for S101 (<i>S. dysenteriae</i> 3 strain)	92
Figure 6.5. Schematic of the S116 (<i>S. boydii</i> 1) derived sequences from the <i>argW</i> U and D primer SGSP-PCRs compared with the Sb227 chromosome.	94
Figure 6.6. Schematic showing the start of the S116 novel GI.	95
Figure 6.7. The Sf301 <i>argW</i> UF showing the positions of the primers used to amplify the region used for homologous recombination with the S116 <i>argW</i> UF.	98
Figure 6.8. pJL5	100
Figure 6.9. View of the Sf301 <i>argW</i> UF region taken from Artemis showing the optimal orientation for the Km ^r cassette	101
Figure 6.10. The four possible conformations of the insert DNA in the pJL5/ <i>Nsi</i> I::Km ^r cassette/ <i>Nsi</i> I constructs.	102
Figure 6.11. Agarose gel showing the <i>Sal</i> I and <i>Pst</i> I digests of four potential pJL5/ <i>Nsi</i> I::Km ^r cassette/ <i>Nsi</i> I constructs.	104
Figure 6.12. pJL6 and pJL7.	105
Figure 6.13. pJL8 and pJL9: pDS132 derived suicide constructs used to deliver the mutant <i>argW</i> UF region to S116.	106
Figure 6.14. Agarose gel showing the <i>Sph</i> I digests of four potential pJL8/pJL9 candidates.	107
Figure 6.15. Schematics showing the <i>argW</i> UF region with the presence (a) or absence (b) of the Km ^r cassette, and the respective <i>in silico</i> PCR amplicons using Sf301 as the template genome.	111

Figure 6.16. Agarose gel showing the results of the colony PCRs on both pJL8 and pJL9 derived potential transconjugant X102 Shigellas.	112
Figure 6.17. The single and double crossover events that could occur at the <i>argW</i> UF region in the chromosome of the pJL9 derived X102 transconjugant Shigellas.	114
Figure 6.18. Potential merodiploids derived from X102 conjugations with pJL9.	116
Figure 6.19. Agarose gel showing the results of the colony PCR on X104 derived potential double crossover transconjugant Shigellae.	117
Figure 6.20. The Principle of marker rescue.	119
Figure 6.21. Agarose gel showing the sizes of the insert fragments harboured by the X106 <i>EcoRI</i> and <i>BamHI</i> marker rescue clones.	120
Figure 6.22. Agarose gel showing the restriction patterns of the X106 <i>EcoRI</i> and <i>HindIII</i> marker rescue clones.	121
Figure 6.23. Relative positions of select restriction sites and fragment sizes around the X106 <i>argW</i> tRNA locus.	122
Figure 6.24. Agarose gel showing the sizes of amplicons generated by primer walking across the X106 <i>argW</i> associated GI.	123
Figure 6.25. Agarose gel showing the sizes of the restriction fragments harboured by the X106 <i>SacI</i> marker rescue clone.	124
Figure 6.26. Schematic showing an <i>in silico</i> PCR using the X106 <i>int</i> F and T7 primers and pJL18 as the template to walk further into the novel X106 <i>argW</i> island.	126
Figure 6.27. Agarose gel showing the pJ118 marker rescue clone walking PCR using the X106 <i>int</i> F and T7 primers to walk further into the X106 <i>argW</i> associated novel integrase-like element.	127
Figure 6.28. The S116 (<i>S. boydii</i> 1) <i>argW</i> associated novel GI.	128
Figure 7.1. <i>leuX</i> U-arm SGSP-PCR results for the <i>S. flexneri</i> strains	132
Figure 7.2. <i>leuX</i> U-arm SGSP-PCR results for all of the <i>S. sonnei</i> strains.	134

Figure 7.3. <i>leuX</i> U-arm SGSP-PCR results for <i>S. boydii</i> strains belonging to <i>leuX</i> island family 4.....	136
Figure 7.4. <i>leuX</i> U-arm SGSP-PCR results for the <i>S. boydii</i> and <i>S. dysenteriae</i> strains belonging to <i>leuX</i> island family 1	139
Figure 7.5. Sequence results of the <i>leuX</i> island family 1-like strains compared with the CFT073 <i>leuX</i> associated GI.	141
Figure 7.6. Schematic showing the location and context of the <i>sigA</i> region in the completed <i>Shigella</i> genomes available in the NCBI database.	144
Figure 7.7. The Sf301 <i>leuX</i> UF, showing the positions of the primers used to amplify the region used for homologous recombination with the S101 <i>leuX</i> UF.....	147
Figure 7.8. Agarose gel showing the two first round PCR amplicons used to generate the <i>leuX</i> UF SOE PCR product.	147
Figure 7.9. Schematic showing the principle of SOE PCR and the generation of the modified <i>leuX</i> UF region containing a central <i>NsiI</i> site and flanked by <i>XbaI</i> sites.	149
Figure 7.10. Agarose gel showing the 1054 bp SOE PCR product.	150
Figure 7.11. pJL1	151
Figure 7.12. Agarose gel showing the results of the colony PCRs on the S101 potential transconjugants.	153
Figure 7.13. Agarose gel showing the sizes of the insert fragments harboured by each of the X101 <i>leuX</i> marker rescue clones pJL10 and pJL11.....	154
Figure 7.14. Agarose gel showing the restriction patterns of the X101 <i>BamHI</i> and <i>HindIII</i> marker rescue clones pJL10 and pJL11.	155
Figure 7.15. The S101 (<i>S. dysenteriae</i> 3 strain) Km ^r derivative X101, marker rescue clone end sequence analysis.	157
Figure 7.16. <i>serU</i> D-arm SGSP-PCR results for the <i>S. flexneri</i> strains belonging to <i>serU</i> island family 1	162

Figure 7.17. <i>serU</i> U-arm SGSP-PCR results for the <i>S. boydii</i> and <i>S. dysenteriae</i> strains belonging to <i>serU</i> island family 2.....	164
Figure 7.18. <i>serU</i> SGSP-PCR results showing strains that harbour the <i>yodB</i> region.....	167
Figure 7.19. View of the <i>E. coli</i> K12 MG1655 chromosome, showing the 1013 bp region in the <i>serU</i> DF that is deleted in Sf301, Sf2457T and CFT073.....	168
Figure 7.20. Agarose gel of a <i>serU</i> U# SGSP-PCR showing the 2.8 kb and 0.6 kb amplicons produced by the S102 (<i>S. dysenteriae</i> 9 strain)/ <i>Pst</i> I library as indicated by the white arrow.	170
Figure 7.21. Blastn analysis of the S101 (<i>S. dysenteriae</i> 3 strain) novel sequence discovered by SGSP-PCR with the <i>serU</i> U#.	171
Figure 7.22. <i>Hind</i> III restriction map of the region surrounding the Sf301 <i>aspV</i> tRNA locus.	177
Figure 7.23. Relative positions of the <i>aspV</i> UF probe primers on the K12 MG1655 chromosome.....	178
Figure 7.24. Photograph of the <i>aspV</i> Southern hybridisation membrane using the DIG system	179
Figure 7.25. The Sf301 <i>sci</i> island as defined by Jin <i>et al.</i> , 2002, and the corresponding tRIP defined GI found associated with the <i>aspV</i> locus in Sf301.	182
Figure 7.26. <i>aspV</i> D-arm SGSP-PCR results for the <i>S. flexneri</i> strains belonging to <i>aspV</i> island family 1	185
Figure 7.27. <i>aspV</i> D-arm SGSP-PCR results for the <i>S. sonnei</i> strains belonging to <i>aspV</i> island family 2.....	187
Figure 7.28. <i>thrW</i> U-arm SGSP-PCR results for the <i>S. flexneri</i> strains belonging to <i>thrW</i> island family 1	191
Figure 7.29. <i>thrW</i> U-arm SGSP-PCR results for the <i>Shigella</i> strains belonging to <i>thrW</i> island family 2.....	192

Figure 7.30. <i>thrW</i> U-arm SGSP-PCR results for the <i>S. boydii</i> strains belonging to <i>thrW</i> island family 2	193
Figure 7.31. <i>thrW</i> U-arm SGSP-PCR results for the <i>Shigella</i> strains designated as <i>thrW</i> island family 3	194
Figure 7.32. Detailed schematic of the S118 (<i>S. boydii</i> 3 strain) <i>thrW</i> U# SGSP-PCR amplicon derived sequence.	196
Figure 7.33. The S107 (<i>S. flexneri</i> 2b strain) <i>thrW</i> U# SGSP-PCR amplicon derived mosaic phage-like sequences	198
Figure A2. 1. View of Sf301 <i>leuX</i> UF region taken from Artemis.....	A2-17
Figure A2. 2. pJL2.	A2-18
Figure A2. 3. Agarose gel showing pJL2 digested with <i>Xba</i> I.	A2-19
Figure A2. 4. pJL3 and pJL4: pDS132 derived suicide constructs used to deliver the mutant <i>leuX</i> UF region to S101.	A2-20
Figure A2. 5. Schematic showing the possible conformations that a GI harboured integrase gene could be in.	A2-23
Figure A2. 6. Positions of the <i>int</i> -PCR primers on the <i>serX/serW</i> associated integrase gene in the sequenced genomes and SRL PAI.	A2-24
Figure A2. 7. Agarose gel of the <i>serW int</i> -PCR.	A2-25
Figure A2. 8. <i>serW</i> D# - P4I_R <i>int</i> -PCR sequencing results for S101 (<i>S. dysenteriae</i> 3 strain).	A2-27
Figure A2. 9. <i>glyU</i> D-arm SGSP-PCR results for the <i>S. sonnei</i> strains.....	A2-32
Figure A2. 10. <i>glyU</i> U-arm SGSP-PCR results for the <i>S. sonnei</i> strain S113.....	A2-33
Figure A2. 11. <i>glyU</i> D-arm SGSP-PCR results for the <i>S. flexneri</i> strains S104, S105, S106 and S107 and the three <i>S. dysenteriae</i> strains.	A2-36
Figure A2. 12. Schematic showing a comparison between the island DNA at <i>glyU</i> in EDL933, Ss046 and Sf301.....	A2-37

Figure A2. 13. S104 (<i>S. flexneri</i> 1a strain) <i>glyU</i> U-arm SGSP-PCR results.	A2-40
Figure A2. 14. <i>pheU</i> U-arm SGSP-PCR results for the <i>S. boydii</i> and <i>S. dysenteriae</i> strains.	A2-48
Figure A2. 15. <i>selC</i> U-arm SGSP-PCR results for the <i>S. flexneri</i> and <i>S. sonnei</i> strains ..	A2-49
Figure A2. 16. <i>selC</i> sequence results of larger than expected tRIP amplicons	A2-53
Figure A2. 17. <i>S. flexneri</i> strains that yielded U-arm amplicons that walked into the <i>pheV</i> associated Sf301 8.4 kb ‘flanking GI’ (<i>pheV</i> island family 2).	A2-57
Figure A2. 18. <i>pheV</i> U-arm SGSP-PCR results for the <i>S. dysenteriae</i> , <i>S. boydii</i> strains and <i>S.</i> <i>flexneri</i> 6 strain.	A2-60
Figure A2. 19. <i>metV</i> SGSP-PCR results for the <i>S. boydii</i> , <i>S. dysenteriae</i> and <i>S. flexneri</i> 6 strains	A2-65
Figure A2. 20. Schematic showing a comparison of the island DNA at <i>metV</i> in CFT073 and Sb227. Figure is not to scale.	A2-66
Figure A2. 21. <i>ssrA</i> SGSP-PCR results for the <i>S. sonnei</i> , <i>S. boydii</i> , <i>S. dysenteriae</i> and <i>S.</i> <i>flexneri</i> 6 strains.	A2-73
Figure A2. 22. <i>serX</i> SGSP-PCR results for the <i>S. sonnei</i> , <i>S. boydii</i> , and <i>S. dysenteriae</i> strains.	A2-79
Figure A2. 23. <i>asnT</i> D# SGSP-PCR results for the <i>S. flexneri</i> strains belonging to island family 3.	A2-83
Figure A2. 24. SGSP-PCR results for the <i>S. dysenteriae</i> strains and S120 (<i>S. boydii</i> 7) .	A2-84
Figure A2. 25. Schematic showing the <i>asnV</i> - <i>asnW</i> inversion in <i>E. coli</i> CFT073 relative to <i>E.</i> <i>coli</i> K12 MG1655.	A2-88
Figure A2. 26. Schematic showing the <i>asnV</i> UF inversion in <i>S. sonnei</i> 53G (unfinished) and <i>S. sonnei</i> Ss046 relative to <i>E. coli</i> K12 MG1655.	A2-89

Addendum 1

CD enclosed on inside back cover.

1.0 Introduction

1.1 The core and mobile genomes

As the number of bacterial strains with completely sequenced genomes increases and the use of comparative genomics technologies such as microarrays becomes more prevalent, it has become evident that related bacteria such as those that are classified into the same 'species' all have the same genetic backbone, which is known as the core genome. This in general is comprised of genes that are only found on the bacterial chromosome and they encode proteins involved in standard metabolic functions. However, in addition to this core set of genes, more and more bacterial strains have been found to harbour non-essential genes which are not always present in the genomes of other representatives of the same 'species' and in some instances are completely unique to a particular strain. This 'optional' DNA is known as the mobile genome as it can be passed from strain to strain by horizontal gene transfer (HGT), the mechanisms by which this occurs are described in section 1.5. The mobile genome is therefore extremely diverse and is comprised of elements such as plasmids, transposons, insertion sequences (IS), intact and degenerate prophages and genomic islands (GIs). *Escherichia coli* are the front runners in genodiversity, with up to 35% of their DNA being variable from strain to strain (Fukuya *et al.*, 2004).

1.2 Genomic islands

Genomic islands (GIs) are defined as large (> 10 kilobase) multi-gene blocks of DNA that although non-essential, often encode specific effector molecules that in some way provide a selective advantage to the host bacterium, enabling it to survive and grow in a new ecological niche. Some GIs enable the host cell to survive and grow in more nutrient limiting or extreme conditions, these GIs are sometimes called fitness islands (Preston *et al.*, 1998). An example of a fitness island is the second largest GI characterised to date; the 500 kilobase (kb) *Mesorhizobium loti* 'symbiosis island' that along with other genes, encodes genes required for

root nodule formation, nitrogen fixation and vitamin synthesis in leguminous plants.

This GI therefore encodes functions that provides nutrients for both the plant and the host bacterium and enables the bacteria to grow in a variety of niches (Sullivan and Ronson, 1998).

For a list of some more previously characterised GIs and the functions they encode, see Table

1.1

1.3 Pathogenicity islands

Some GIs encode toxins, invasion apparatus or host defence protection systems that allow the bacterium to cause disease in a particular plant or animal host, therefore promoting the survival and dissemination of the pathogen. This subset of GIs are called pathogenicity islands (PAIs) and were first discovered by Jörg Hacker and his colleagues in uropathogenic *E. coli* (UPEC) (Knapp *et al.*, 1986). Two PAIs were found that carry genes that encode haemolysins and P-related fimbriae in the *E. coli* strain 536, these were designated PAI I and PAI II and are 70 kb and 90 kb long respectively. Deletion of either of the PAIs resulted in a reduction in the pathogenicity of the host bacterium, proving that these elements contribute to the virulence of *E. coli* strain 536. These PAIs were also found to be inserted on the chromosome adjacent to the tRNA genes *selC* and *leuX* (Blum *et al.*, 1994, Hacker *et al.*, 1990). Earlier studies had shown that tRNA loci act as insertion sites for bacteriophages and plasmids in bacterial chromosomes (Reiter *et al.*, 1989) and this work by Hacker and colleagues was the first to show that GIs also insert at tRNA loci (see section 1.6 also for more details on the association of GIs with tRNA loci).

For a GI to be designated as a PAI, it must have one distinguishing feature to all other GIs; that it is present in the genome of a virulent strain, but absent from the genomes of avirulent members of the same ‘species’ or related ‘species’ (Schmidt and Hensel, 2004).

Numerous PAIs have since been characterised across bacteria, the majority being found in Gram-negative pathogens, especially across the Enterobacteriaceae. However, PAIs have also

been discovered in Gram-positive pathogens, indicating that they are present in all groups of pathogens. Table 1.1 shows some examples of PAIs.

1.4 Structure of genomic islands

Genomic islands are usually 10 – 200 kb long (< 10 kb they are known as ‘genomic islets’); they usually harbour genes that encode specific functions which in turn confer a selective advantage to the bacterial host. GIs are often found associated with tRNA loci; they often harbour intact or truncated bacteriophage-like integrase genes and other intact or cryptic phage-like elements and mobility genes such as insertion sequences (IS). The presence of these elements shows that GIs are acquired by HGT from foreign sources. Other features that indicate that GIs are acquired from foreign sources, include signatures such as different guanine plus cytosine (GC) contents to the core chromosomal DNA, different codon usage and dinucleotide bias (see section 1.7 for more details). GIs are also often flanked by direct repeats (DRs) which are generally 7-30 bp long. Figure 1.1 shows what a ‘classic’ GI looks like.

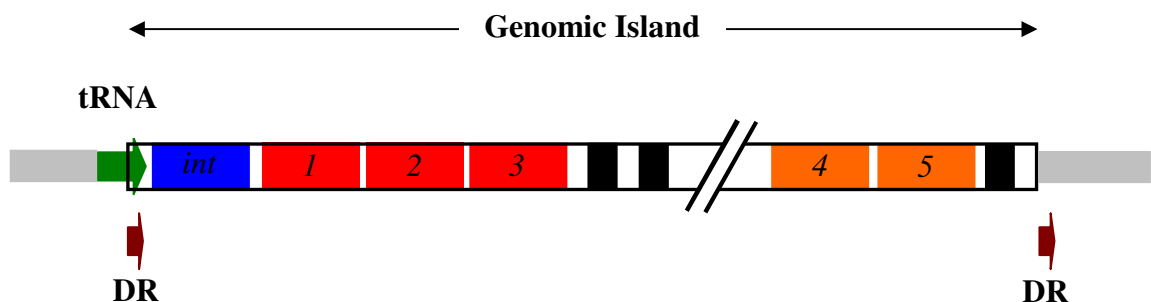


Figure 1.1. Schematic of a typical genomic island.

Core DNA is represented by the grey regions, the tRNA gene and its orientation is shown by the green arrow, and the long rectangular box represents the GI. The coloured boxes within the GI are different genes; *int* stands for integrase gene, *1*, *2* and *3* are genes that encode specific effectors, *4* and *5* encode other effector molecules. The black boxes represent mobility genes and the deep red arrows represent the flanking direct repeats (DRs).

Table 1.1. Examples of genomic islands in bacteria

Name and description	Functions encoded	Organism	Size (kb)	Ass. tRNA	References
<i>clc</i> element Fitness island	Degradation of phenolic compounds	<i>Pseudomonas putida</i>	105	<i>gly</i>	(Ravatn <i>et al.</i> , 1998)
MAI Fitness island	Magnetosome (magnetic phenotype)	<i>Magnetospirillum gryphiswaldense</i>	130	-	(Schubbe <i>et al.</i> , 2003, Ullrich <i>et al.</i> , 2005)
PAI-II Pathogenicity island	Haemolysin, P-fimbriae	<i>E. coli</i> 536 (UPEC)	190	<i>leuX</i>	(Blum <i>et al.</i> , 1994)
LEE Pathogenicity island	Type III secretion, invasion	<i>E. coli</i> O157:H7 (EHEC)	43	<i>selC</i>	(Perna <i>et al.</i> , 2001)
HPI Pathogenicity island	Yersiniabactin (iron uptake)	<i>Yersinia pestis</i> / <i>Y. pseudotuberculosis</i>	36	<i>asn</i>	(Buchrieser <i>et al.</i> , 1998)
CTnscr94 Fitness island	Sucrose uptake	<i>Salmonella senftenberg</i>	100	<i>pheV</i>	(Hochhut <i>et al.</i> , 1997)
VPI Pathogenicity island	Adhesin, receptor for CTXΦ phage	<i>Vibrio cholerae</i>	39.5	<i>ssrA</i>	(Miller, 2003)
SPI-1 Pathogenicity island	Type III secretion system	<i>Salmonella typhimurium</i>	40	-	(Galan, 1996)

Ass., associated; EHEC, enterohaemorrhagic *E. coli*; UPEC, uropathogenic *E. coli*.

1.5 Horizontal transfer of genomic islands

Point mutations and small genomic rearrangements contribute to strain variation and are therefore important in the evolution of bacteria, whereas the acquisition of a single GI in some cases can enable a bacterium to colonise a completely new niche, or even convert a previously non-pathogenic strain into a pathogen, resulting in an evolutionary ‘quantum leap’ (Groisman and Ochman, 1996). Therefore, the mechanisms by which GIs move from strain to strain and even from species to species are of great importance to our understanding of genome plasticity and bacterial evolution. There are three mechanisms that enable HGT between bacteria, these are detailed below.

1.5.1 Natural transformation

Some bacteria have been found to be able to take up exogenous DNA from the environment, which then becomes incorporated into the genome; this is known as natural transformation and was first observed in *Streptococcus pneumoniae*. It has since been observed in *Bacillus subtilis*, *Haemophilus influenzae* and *Neisseria gonorrhoeae* and studies have shown that the process is growth stage dependent and affected by environmental factors. In *Neisseria* spp., this process is responsible for high levels of HGT and transformation efficiencies as high as 1 in 100 have been reported in *N. gonorrhoeae* (Sparling, 1966). For an excellent review on natural transformation in bacteria see (Dubnau, 1999). Whether natural transformation occurs in other species is unknown as it is very difficult to observe under *in vitro* conditions.

1.5.2 Conjugation

Conjugation is the transfer of DNA from a donor cell to a recipient cell by intimate cell-cell contact via conjugation apparatus; in Gram-negative bacteria it occurs widely and the conjugation process has been well characterised. The apparatus by which conjugation is initiated are known as the pili, these are extracellular filaments that are secreted by the donor cell, they attach to the recipient and then bring the two cells into direct contact. Circular DNA

elements can then pass from donor to recipient, this includes plasmids and ‘integrative and conjugative elements’ (ICEs) (Burrus *et al.*, 2002), which in turn carry the genes that encode the conjugation apparatus and the transfer system. When DNA passes from donor to recipient, only a single strand of the conjugative element is transferred, the two single strands present in both the donor and recipient cells then replicate to form double stranded circular elements. Therefore the plasmid DNA is copied, creating two potential donors. In Gram-positive bacteria the conjugation process is less well known and a model for the way in which cell-cell contact occurs has only recently been proposed (Abajy *et al.*, 2007).

Some circular elements can also integrate into the bacterial chromosome by recombination to form episomal GIs; an example of this is the 100 kb *Vibrio cholerae* SXT ICE that carries four antibiotic resistance genes and is a mosaic that comprises plasmid and phage-like genes (Beaber *et al.*, 2002). This element is also able to excise precisely from the chromosome into the circular form and via conjugation has disseminated widely throughout *V. cholerae* in Asia. This also explains why some GIs often harbour plasmid-like sequences, as they are remnants of previous insertion events by circular elements that have since become locked into the chromosome.

1.5.3 Transduction

Transduction is the mechanism by which bacterial DNA is transferred from one cell to another by a bacteriophage. After a bacteriophage has infected a bacterial cell, two events can occur, the lytic cycle or lysogenic cycle, where each results in a different mechanism of ‘capture’ of bacterial DNA. The lytic cycle leads to generalised transduction and the lysogenic cycle can result in specialised transduction, these are detailed below.

Generalised transduction

During the lytic cycle the host bacteria’s genome is broken up, as the formation of new phage particles occurs, sometimes a fragment of the bacterial DNA is accidentally packaged into a phage head along with the phage DNA (Schmidt and Hensel, 2004). This process is

completely random, any part of the bacterial DNA can be packaged into the phage head and the amount of bacterial DNA that can be packaged depends on the total size of the phage genome. If the newly formed phage then infects another cell, the injected bacterial sequence may be incorporated into the new hosts' genome by recombination, providing there is enough homology between the transduced bacterial DNA and the recipients DNA. This is known as generalised transduction. Even though generalised transduction occurs at a low frequency (generally around 1 in 10^5 (Lederberg *et al.*, 1951)), in some genera the number of phages that can perform generalised transduction can be high; around 99% of *Salmonella* phages can perform generalised transduction (Schicklmaier *et al.*, 1998), indicating that they play a significant role in HGT.

Specialised transduction

Virulent phages always undergo the lytic cycle, whereas temperate phages such as phage λ , which infects *E. coli*, can either enter the lytic cycle or it can integrate stably into the bacterial chromosome by recombination at its *attP* (attachment) site with the specific *attB* site on the bacterial chromosome. This is known as the lysogenic cycle and the chromosomally harboured phage is then known as a prophage. Two phage encoded proteins are necessary for stable lysogeny to occur, Int and CI. Int, which is encoded by the *int* gene enables the phage to integrate into the host chromosome and CI, encoded by the *cI* gene represses the promoters for expression of the lytic cycle genes (Court *et al.*, 2007).

Under certain conditions, excision of the prophage from the chromosome can also take place, the excised phage then switches to the lytic cycle. This is known as prophage induction and occurs when the host cells DNA is damaged; this can be due to UV irradiation or in the presence of DNA damaging drugs such as mitomycin C (Shinagawa and Ito, 1973). Prophage induction also occurs in the bacterial population spontaneously, however, it is at low levels (in around 0.005% of the host cells per generation, in *in vitro* conditions (Livny and Friedman, 2004)). Excision of the prophage from the chromosome is usually precise, however, at a low frequency; faulty excision of the prophage occurs, which results in the phage acquiring some

of the bacteria's chromosomal sequence and in its place it leaves some of its phage sequence. All of the resulting phage particles then carry the bacterial sequence. If any of the new phage particles then infect another cell, the bacterial sequence may be incorporated into the hosts' genome by recombination or by integration of the phage into the chromosome at its specific *att* site; this is known as specialised transduction. Unlike generalised transduction, in specialised transduction bacterial DNA is not packaged randomly; only sequences that are close to the original location of the prophage can be transduced into the second cell.

The above mechanisms also highlight the role that plasmids and bacteriophages play in the shaping and evolution of GIs and the association of plasmid and phage-like genes with GIs. Bacteriophage mediated HGT in particular plays a major role in the dissemination of GIs across bacteria; this is because unlike with transformation and conjugation, where the donor and recipient cells have to be in close proximity; transduction allows DNA to be transferred from one environment to another (Medini *et al.*, 2005) and the donor organism is not always present at the same place or time as the recipient bacteria. Also, as phages infect nearly all known bacterial species (Schmidt and Hensel, 2004), transduction therefore enables DNA to be exchanged between organisms that live in completely different ecological niches and even between different genera.

1.6 tRNA loci as integration sites for genomic islands

Nearly twenty years ago transfer RNA (tRNA) loci were reported to be frequent integration sites for mobile genetic elements in bacteria (Reiter *et al.*, 1989) and it is well known that lysogenic bacteriophages often insert by recombination at tRNA loci because their *att* sites frequently lie within tRNA genes (Campbell, 2003). As GIs often harbour prophage-like *int* genes (see Figure 1.1 above) and DRs that resemble phage *att* sites, they too frequently insert at tRNA loci and previous studies have shown that around 75% of PAIs discovered so far are inserted at tRNA genes (Hacker and Kaper, 2000). Also recent studies have shown that in

some cases a GI needs only a single DR to be able to insert into the bacterial chromosome, as other integrases already present on the chromosome can substitute for the activity of a GI harboured integrase gene (Muniesa *et al.*, 2006). The *attB* regions within tRNA sites vary in size, but are usually 20-30 bp long, and they often extend to and sometimes past the 3' terminus of the tRNA gene; this may be in order to prevent disruption of the tRNA after the integration event, so that it still functions (Williams, 2002). This may also explain why GIs are always found downstream of the 3' end of their associated tRNA genes (see Figure 1.1) and never upstream of the 5' end. This orientation would also promote expression of the GI, as it is colinear with the tRNA gene.

Why tRNA loci are frequent targets for the insertion of GIs is unknown, however, they have a number of characteristics that make them an attractive choice.

tRNA loci are well conserved across all bacteria and their average sequence divergence rate per base pair is 6-fold lower than for protein coding genes (Williams, 2002); they are therefore reliable targets for GIs to insert at in any potential host and also increase the chances of the maintenance and dissemination of GIs across a bacterial population (Schmidt and Hensel, 2004). tRNA genes are distinct to other genes and have a conserved secondary structure that may provide structural motifs that facilitate the integration and excision of GIs (Reiter *et al.*, 1989). Their small size (typically 70-80 bp) also reduces the amount of host DNA that must be recombined in order to restore the tRNA gene on integration of the GI (Williams, 2002). There are also numerous tRNA genes in bacterial chromosomes (*E. coli* K12 MG1655 has 86 tRNA genes); this improves the chances of the integration of GIs into the host chromosome. However, the collective results so far have shown that certain tRNA loci are favoured as insertion sites for GIs and prophages and are more frequently associated with foreign DNA; the reasons for this are unclear and there are two hypotheses as to why specific tRNA sites are preferred. The 'minor codon' hypothesis suggests that less common tRNA genes are favoured, as they can read rare codons present on the associated GI, therefore

improving the expression of the island genes. This hypothesis is supported by work on PAI II in *E. coli* 536 (UPEC), which is associated with the *leuX* tRNA and shows that expression of *leuX* is required for the production of virulence factors encoded on the island (Ritter *et al.*, 1997). This hypothesis is not supported by the fact that other genes that are not involved in virulence functions also depend on *leuX*, also that GIs are often associated with the *ssrA* locus, which is the transfer-messenger-RNA (tmRNA) and does not perform the same functions as tRNA genes (Hou, 1999).

The ‘major codon’ hypothesis suggests that tRNA genes with multiple copies are more frequently occupied because they provide multiple integration sites for GIs, therefore amplifying virulence factors (Schmidt and Hensel, 2004). This is not the case however for *selC* and *leuX*, which are present as single copies (Hou, 1999) and *selC* is one of the most frequently occupied tRNA loci (Hacker and Kaper, 2000).

The further discovery and characterisation of GIs and their tRNA gene preferences will hopefully shed more light on this phenomenon.

1.7 Signatures used to discover genomic islands *in silico*

Genomic islands often have a number of characteristics that make them distinct to the rest of the chromosome and indicate that they were acquired by HGT from a foreign source. These signatures have been utilised to discover GIs *in silico* across the chromosomes of fully sequenced bacterial strains. It was first recognised that GIs have different GC contents to their surrounding chromosomal regions by Jörg Hacker and colleagues (Hacker *et al.*, 1997); soon after, a number of software tools were developed that scan DNA sequence using a sliding window approach to detect changes in the GC content across the chromosome, with the results being presented in graphical format. The use of GC skew (Grigoriev, 1998) and GC wavelets (Lio and Vannucci, 2000) which pick up GC profile irregularities resulted in the discovery of genomic rearrangements in *E. coli* and two putative PAIs in *N. meningitidis* Serogroup B strain MC58 respectively. A more sensitive windowless approach was also

developed known as the cumulative GC profile or 'z' curve' that incorporates, purine/pyrimidine, amino/keto and weak/strong hydrogen bond distribution to calculate the GC content profile along a DNA sequence. The presence of a sharp uprising peak in the z' curve graph of a DNA sequence indicates the presence of a GI (Zhang and Zhang, 2004).

Samuel Karlin and colleagues found that; codon bias, amino acid bias and dinucleotide bias relative to the average for the chromosome could also be used to identify horizontally acquired DNA and developed a software tool that incorporated all four of the above compositional features to detect GIs in sequenced strains of eight bacterial species including *N. meningitidis*, *V. cholerae*, *Mycobacterium tuberculosis* and *E. coli* (Karlin, 2001, Karlin *et al.*, 1998a, Karlin *et al.*, 1998b).

Jeffrey Lawrence and Howard Ochman also used GC content, codon bias and the relative location of putative island-like sequences, to show that at least 18% of open reading frames (ORFs) in the *E. coli* K12 MG1655 genome were acquired by HGT (Lawrence and Ochman, 1998)

The use of differential base composition and codon bias to detect island DNA in sequenced strains was however criticised by Liisa Koski and colleagues. They reported that sequences acquired by HGT from closely related bacteria would have a similar GC content and codon usage and would therefore not be detected. They also claimed that as horizontally acquired DNA slowly ameliorates to match the base composition of the new host genome (Lawrence and Ochman, 1997), more ancient island DNA would also be missed using this approach (Koski *et al.*, 2001).

Since then, an elegant visualisation tool known as IslandPath has been developed which includes the above base compositional features and other annotational features such as proximity to tRNA genes and the presence of mobility genes to accurately detect GIs in sequenced bacterial chromosomes (Hsiao *et al.*, 2003).

Also, more recently, an algorithm that detects tRNA associated GIs that are flanked by DRs and harbour an intact integrase gene homolog was used to screen the 106 completed prokaryotic genomes available at GenBank in July 2003, to create a database of integrative elements called Islander (Mantri and Williams, 2004). Out of all of the bacterial groups screened, the *Escherichia/Shigella/Salmonella* group had the highest average number of GIs detected per strain (5.0), indicating that their respective GIs have a broad host range and that tRNA associated islands play a significant role in these bacterial species.

1.8 Genomic island discovery in unsequenced strains

As the number of fully sequenced bacterial genomes currently available represent only a tiny proportion of the number of strains in a bacterial population, it is therefore essential for us to probe multiple representatives from each species in order to improve our understanding of their diversity (Medini *et al.*, 2005). The ongoing discovery and characterisation of PAIs in particular has important clinical implications in both our overall knowledge of pathogens and virulence properties, also in revealing potentially new antimicrobial targets.

Whole-genome sequencing of hundreds of strains to discover GIs is costly and not always practical, especially when you consider that only up to 35% of sequences present in the genome of a given strain at any particular time is island DNA (Fukuya *et al.*, 2004). Therefore, other methods have also been developed to interrogate unsequenced strains.

DNA microarrays are a powerful tool which have been used to compare bacterial strains (genomotyping) and can therefore be used to rapidly determine the overall collection of GIs that are harboured by a particular bacterial group (Lucchini *et al.*, 2001), or to find out the prevalence of specific elements across multiple strains. Microarrays are therefore very useful in clinical diagnostics; an example of this is the ‘STEC-EPEC Oligonucleotide Microarray’ which is used to type Shiga toxin-producing *Escherichia coli* (STEC) and enteropathogenic *E.*

coli (EPEC) from human and animal infections based on the variation found in their respective locus of enterocyte effacement (LEE) PAIs (Garrido *et al.*, 2006).

Even though microarrays enable scientists to generate huge amounts of data by comparing hundreds of strains, they do not enable us to capture novel genetic elements as they are based on hybridisation of the test sample with known DNA probes.

The detection of novel DNA on a whole-genome scale can be performed using subtractive hybridisation (SH). This technique also uses DNA hybridisation between a test strain and a reference strain; however, DNA that is unique to the test organism which does not hybridise is then captured, amplified and sequenced. SH was first reported nearly twenty years ago and has since been used extensively to discover novel GIs across bacteria. For an excellent review on SH and its applications see (Winstanley, 2002).

Another technique developed by Chad Malloff and colleagues which uses two-dimensional gel displays to compare bacterial genomes was also used to identify strain differences between *M. tuberculosis*, *M. avium* and *M. bovis* and potential novel elements across strains of *Bordetella pertussis* (Malloff *et al.*, 2003, Malloff *et al.*, 2002, Malloff *et al.*, 2001).

SH and two-dimensional displays are extremely useful tools in the identification and capture of novel GIs found in closely related bacterial strains; however they sometimes miss novel elements; in SH this can be due to non-specific or background hybridisation, or sub-optimal hybridisation conditions. Also, when using two-dimensional displays, DNA can simply be run off the gel, or contained in areas of poor resolution (Malloff *et al.*, 2003, Winstanley, 2002).

The continual development of effective screening methods to discover both known and novel GIs in unsequenced bacterial strains is therefore important in furthering our knowledge of bacterial diversity, particularly in bacterial pathogens; as they enable us to recognise emergent pathogenic strains and may help us to identify potential drug targets.

I aim to address this in my study, by developing an effective, high-throughput method to detect tRNA associated GIs in *Shigella*, one of the causative agents of bacillary dysentery in humans.

1.9 *Shigella*

Shigella are Gram-negative, non-motile, nonsporulating, facultatively anaerobic bacilli that cause shigellosis (bacillary dysentery), with humans being their only natural host. The organism is highly infectious with a 50% infective dose (ID_{50}) of 500 (the number of organisms required to cause infection in at least half of the exposed hosts); also, ingestion of as few as 10 organisms can cause shigellosis (DuPont *et al.*, 1989). *Shigella* are transmitted mainly via the fecal-oral route or by the ingestion of contaminated food. *Shigella* causes dysentery through invasion of the colonic mucosa and subsequent multiplication followed by intracellular and intercellular spread and infection of neighbouring cells. Degeneration of the epithelium then occurs; this, along with the hosts' acute inflammatory response leads to colitis of the mucosa which results in the leakage of blood and mucus into the intestinal lumen. The symptoms of shigellosis can range from mild diarrhoea to severe dysentery with the passage of frequent bloody, mucoid stools; other symptoms include fever, intestinal cramps and convulsions. The disease is usually self-limiting, but it can lead to death, especially in infants (Hale, 1991).

Shigella were first described by Kiyoshi Shiga in 1898 (Shiga, 1898), after he isolated what he called *Bacillus dysenteriae* (now known to be *Shigella dysenteriae* serotype 1) from a patients' stool during a dysentery epidemic in Japan in 1897. Further studies resulted in the isolation of different biotypes and serotypes of the organism and in 1950, the genus *Shigella* comprising four species was announced; *S. dysenteriae*, *S. flexneri*, *S. boydii* and *S. sonnei* (Ewing, 1949, Hale, 1991, Niyogi, 2005).

It is now well known that *Shigella* are members of the Enterobacteriaceae family and they are pathogenic clones of *E. coli* (Lan and Reeves, 2002, Pupo *et al.*, 2000); however, the genus distinction of *Shigella* has been retained because of their medical significance.

Shigella can be distinguished from *E. coli* by their biochemical properties; unlike *E. coli*, *Shigella* do not decarboxylate lysine, and generally do not ferment lactose (*S. sonnei* exhibit 'slow' lactose fermentation when incubated for an extended time). The four species of *Shigella* are also subdivided by their O-antigens (lipopolysaccharide [LPS]) into different serotypes. There are currently 47 recognised *Shigella* serotypes; *S. dysenteriae* has 13, *S. flexneri* has 15 (including subserotypes that are due to phage conversion), *S. boydii* has 13 and *S. sonnei* has only one serotype, as it is derived from a single clone (Karaolis *et al.*, 1994, Niyogi, 2005). *Shigella* are also different to *E. coli* in that they lack both the flagellar H antigen (explaining why they are non-motile) and capsular K antigen that are also used to type *E. coli* strains (Pupo *et al.*, 2000).

1.10 Epidemiology

A review in 1999 on *Shigella* infection between 1966 and 1997 by Kotloff and colleagues showed that *Shigella* are responsible for an estimated 164.7 million cases of diarrhoea per year; 163.2 cases per year occurred in developing countries with 1.1 million deaths. Overcrowding, poor hygiene, contamination of water supplies with sewage and HIV – associated immunodeficiency are the main reasons for 99% of Shigellosis cases being in developing countries. 1.5 million cases per year occurred in industrialised countries. 69% of all infections and 61% of all deaths were in children under 5 years old. *S. dysenteriae* serotype 1 is the causative agent of epidemic Shigellosis in the Indian subcontinent, Southeast Asia and Central Africa, it is also responsible for 30% of *S. dysenteriae* cases in endemic regions; however more recent studies have shown that in some regions serotypes 2 to 12 are becoming predominant (Talukder *et al.*, 2003). Infection with *S. dysenteriae* serotype 1 is particularly severe because it also produces Shiga toxin (see section 1.11 also). *S. flexneri*

was found to be the predominant species in endemic regions (see Table 1.2), with serotype 2a being the major strain, however, a more recent study in Bangladesh showed that *S. flexneri* 2b is now the predominant serotype in this area (Talukder *et al.*, 2001).

Table 1.2. The median of *Shigella* isolates in developed and developing countries ^a.

	<i>S. flexneri</i>	<i>S. sonnei</i>	<i>S. boydii</i>	<i>S. dysenteriae</i>
Developing countries	60%	15%	6%	6%
Industrialised countries	16%	77%	2%	1%

^a Figures taken from (Kotloff *et al.*, 1999)

S. sonnei is the outbreak strain in developed countries and is responsible for sporadic, common-source outbreaks, transmitted by uncooked food. Homosexual men are also at high risk, with infections mainly being from *S. flexneri* (Tauxe *et al.*, 1988). *S. boydii* is the least common species and is generally only found in the Indian subcontinent (Niyogi, 2005), it is also the most diverse of the *Shigella* species (Feng *et al.*, 2005).

1.11 *Shigella* Pathogenesis

Work in the 1980s by Philippe Sansonetti and colleagues showed that the invasion ability of *Shigella* is due to the presence of a 210-230 kb virulence plasmid (pINV); as the transfer of pINV to *E. coli* K12 ‘genoconverted’ the *E. coli* into ‘*Shigella*’ as it enabled the recipient cells to invade HeLa cells (Sansonetti *et al.*, 1983). Further work also showed that a number of chromosomal loci also play a regulatory role in virulence. The virulence plasmid is present in all *Shigella* strains tested so far and all of the genes that are required for entry into host cells are found in a 32 kb region. This region includes the *ipa*, *mxi* and *spa* genes which encode a type III secretion system (T3SS). The T3SS acts like a needle complex that translocates effector proteins from the bacterial cytoplasm into the host cell; the effector proteins then mediate the formation of an actin-rich ruffle on the surface of the host cell, the

bacterium then enters the host cell through a mechanism that is similar to phagocytosis. The *mxi* (standing for ‘membrane excretion of Ipa’) and *spa* (standing for ‘surface presentation of Ipa’) genes encode the type III needle complex that in turn delivers the *ipa* gene products – Ipa proteins (standing for ‘invasion plasmid antigen’) into the host cell. There are also other genes encoded elsewhere on the virulence plasmid that regulate expression of the T3SS and others that are involved in intracellular and intercellular spread of the bacteria (Blocker *et al.*, 2001, Clerc and Sansonetti, 1987, Hale, 1991, Maurelli, 1989, Maurelli and Sansonetti, 1988, Parsot and Sansonetti, 1996, Sasakawa *et al.*, 1992). In some *Shigella* strains, in particular *S. sonnei*, the virulence plasmid also encodes an enterotoxin known as ShET2 (Nataro *et al.*, 1995, Vargas *et al.*, 1999).

A second enterotoxin known as ShET1 is well conserved across *S. flexneri* 2a strains, which is the most common strain in endemic regions. The genes that encode this enterotoxin (*setIA* and *setIB*) were found to be harboured on a PAI termed the *she* PAI by Kumar Rajakumar and colleagues (see Table 1.3 also). Their frequent presence across *S. flexneri* 2a strains compared with their limited presence across other *Shigella* strains indicates that this toxin plays an important role in the virulence of *S. flexneri* 2a strains and could be one of the factors that has led to it being the most prevalent endemic strain (Noriega *et al.*, 1995, Rajakumar *et al.*, 1997)

S. dysenteriae 1 is distinct to all of the other *Shigella* strains in that it causes the most severe dysentery as well as a rare sequela called haemolytic-uremic syndrome (Hale, 1991). This is because it also produces an exotoxin with enterotoxic, cytotoxic and neurotoxic effects known as Shiga toxin, which is similar in structure to the Shiga-like toxins of enterohaemorrhagic *E. coli* (EHEC) (Niyogi, 2005). Shiga toxin is encoded by a Shiga toxin-encoding bacteriophage (Stx-phage) that is incorporated into the *S. dysenteriae* 1 chromosome. Upon induction of the

phage, due to damage to the host cells DNA, the *stx* genes are expressed and Shiga toxin is released.

LPS (O-antigen) is another important virulence factor of *Shigella*, as it is an endotoxin that stimulates a strong inflammatory response in the host by activating immune cells, in particular macrophages, to release pro inflammatory cytokines. It also enables *Shigella* to resist host defence systems such as opsonisation and phagocytosis (Lindberg *et al.*, 1991).

As the immune response to *Shigella* is O-antigen specific, after infection with a particular *Shigella* strain, the host then has protective immunity to all strains of the same serotype, but not to infection with other serotypes; therefore, the ability to change serotypes would confer a selective advantage to the invading organism (Schmidt and Hensel, 2004).

Across *S. flexneri* there are currently 15 distinct serotypes, these are due to differences in the structure of their O-antigens. These modifications are due to glucosylation and O acetylation of the basic O-antigen side chain, which are mediated by proteins that are encoded by serotype converting bacteriophages that integrate into the *Shigella* chromosome. A number of these bacteriophages have been characterised, they integrate into the *thrW* tRNA locus and are a highly variable group of elements; however, they all harbour the genes responsible for O-antigen modification and are known collectively as ‘*Shigella* island-O’ (SHI-O, see Table 1.3).

Shigella also has a number of other GI encoded virulence factors, which contribute to the burden of the disease; these are described in section 1.12.

1.12 Genomic islands in *Shigella*

So far, across *Shigella*, only five chromosomal GIs have been well characterised; SHI-O, SHI-I (also known as the *she* PAI), SHI-2, SHI-3 and the SRL PAI, these are all associated with tRNA loci and are described in Table 1.3.

Table 1.3. tRNA associated genomic islands characterised in *Shigella*

GI name	Ass. tRNA	Size (kb)	GC (%)	Key encoded genes	Corresponding proteins produced and their functions	Distribution across <i>Shigella</i>	References
SHI-O	<i>thrW</i>	6.5-39.0	N/A ^a	<i>gtr</i> <i>oac</i>	Gtr – Specific glucosyltransferase, mediates specific O-antigen (LPS) modification Oac – O-acetyltransferase, O-antigen (LPS) modification	<i>S. flexneri</i>	(Adhikari <i>et al.</i> , 1999, Allison <i>et al.</i> , 2002, Allison and Verma, 2000, Casjens <i>et al.</i> , 2004, Clark <i>et al.</i> , 1991, Guan <i>et al.</i> , 1999, Ingersoll <i>et al.</i> , 2002, Mavris <i>et al.</i> , 1997, Verma <i>et al.</i> , 1991, Verma <i>et al.</i> , 1993)
SHI-I (<i>she</i> -PAI)	<i>pheV</i>	46.6	49.1	<i>int</i> <i>she (pic)</i> <i>sigA</i> <i>setIA & setIB</i> <i>sap</i>	Int – mediates integration of the GI into the bacterial chromosome Pic – serine protease, mucinase, haemagglutination, serum resistance (member of SPATE family) SigA – serine protease, cytopathic, enterotoxic, (member of SPATE family) ShET1 – enterotoxin May encode a protein involved in autoaggregation	<i>S. flexneri</i> 2a, 2b, 3c and structural variants of the PAI are found in all species of <i>Shigella</i>	(Al-Hasani <i>et al.</i> , 2001a, Al-Hasani <i>et al.</i> , 2000, Al-Hasani <i>et al.</i> , 2001b, Henderson <i>et al.</i> , 1999, Rajakumar <i>et al.</i> , 1997)

SHI-2	<i>selC</i>	23.8	48.6	<i>int</i>	Int – mediates integration of the GI into the bacterial chromosome	<i>S. flexneri</i> and <i>S. sonnei</i> (<i>S. sonnei</i> do not harbour <i>shiD</i>)	(Ingersoll <i>et al.</i> , 2003, Moss <i>et al.</i> , 1999, Vokes <i>et al.</i> , 1999)
				<i>iucA, B, C, D and iutA</i>	Aerobactin – Iron acquisition system (siderophore)		
				<i>shiA</i>	ShiA – Down-regulates the host inflammatory response		
				<i>shiD</i>	ShiD – Colicin V immunity		
SHI-3	<i>pheU</i>	21.0	51.9	<i>int</i> ^b	Int – mediates integration of the GI into the bacterial chromosome	<i>S. boydii</i>	(Purdy and Payne, 2001)
				<i>iucA, B, C, D and iutA</i>	Aerobactin – Iron acquisition system (siderophore)		
SRL PAI	<i>serX/serW</i>	66.3		<i>int</i>	Int – mediates integration of the GI into the bacterial chromosome	<i>S. flexneri</i> , <i>S. sonnei</i> , <i>S. boydii</i> and <i>S. dysenteriae</i>	(Luck <i>et al.</i> , 2004, Luck <i>et al.</i> , 2001, Turner <i>et al.</i> , 2001, Turner <i>et al.</i> , 2003, Turner <i>et al.</i> , 2004)
				<i>oxa-1, cat, aadA1 and tetA(B)</i>	Oxa-1, Cat, AadA1, TetA(B) - Confer Ap ^r , Cm ^r , Sm ^r and Tc ^r respectively		
				<i>fec operon</i>	Fec – ferric dicitrate uptake system (siderophore)		

Ass., associated; SHI, *Shigella* island; SPATE, serine protease autotransporters of *Enterobacteriaceae*; SRL, *Shigella* resistance locus.

^a The SHI-O islands are highly variable, therefore their GC contents vary; however, the *gtr* gene cluster has a GC content of 40%.

^b The SHI-3 integrase gene has nucleotide insertions or deletions that have generated frameshifts, preventing the translation of a functional integrase protein.

1.13 Treatment of *Shigella* infections

Oral rehydration and treatment with antibiotics has proved to be effective against *Shigella*, however, the range of effective antimicrobials is becoming limited due to the continuing global emergence of drug resistance across *Shigella*. Resistance to ampicillin, trimethoprim-sulfamethoxazole, tetracycline and sulfonamides exists all over the world and around half of *S. sonnei* and *S. flexneri* strains are resistant to ampicillin and trimethoprim-sulfamethoxazole. Currently fluoroquinolones such as ofloxacin, ciprofloxacin or norfloxacin are the drugs of choice if ampicillin treatment does not work, however fluoroquinolones are not usually given to pregnant women or children as there have been concerns that they may induce cartilage degeneration, but studies have shown that they are safe (Niyogi, 2005, Schaad and Wedgwood, 1992).

Unfortunately there have been some recent reports of fluoroquinolone resistant *S. dysenteriae* 1 strains (Sur *et al.*, 2003), also as there is no approved vaccine for *Shigella*, there is an urgent need to develop new treatments for *Shigella*.

1.14 The aims of this study

- To develop an effective, high-throughput screen to detect tRNA associated GIs in unsequenced *Shigella* and *E. coli* strains.
- To detect and identify both known and novel tRNA associated GIs amongst a diverse set of *Shigella* strains in order to gain further insights into the distribution of these entities throughout this group of pathogens and to characterise at least one novel island.
- To identify potential targets for novel *Shigella*-specific antimicrobial therapies.

2.0 Materials and methods

2.1 Bioinformatics

2.1.1 *E. coli* genomes analysed

The sequences of four complete *E. coli* and two complete *Shigella* genomes were downloaded from the NCBI (<ftp.ncbi.nih.gov/genomes/>). The strains analysed were *E. coli* K12 MG1655 (accession number NC_000913) (Blattner *et al.*, 1997), enterohaemorrhagic *E. coli* (EHEC) strains O157:H7 EDL933 (NC_002655) (Perna *et al.*, 2001) and O157:H7 RIMD 0509952 (Sakai outbreak isolate NC_002695) (Hayashi *et al.*, 2001), uropathogenic *E. coli* (UPEC) CFT073 (NC_004431) (Welch *et al.*, 2002), *S. flexneri* 2a Sf301 (NC_004337) (Jin *et al.*, 2002) and *S. flexneri* 2a Sf2457T (NC_004741) (Wei *et al.*, 2003). Genome sequences were visualised using Artemis Release 5 (<http://www.sanger.ac.uk/Software/Artemis>).

2.1.2 tRNA loci chosen to study

Following an initial *in silico* analysis of one *Shigella* (Sf301) and three *E. coli* genomes (MG1655, CFT073, and EDL933) using Blastn (Altschul *et al.*, 1997) (<http://www.ncbi.nlm.nih.gov/BLAST/>), fifteen tRNA loci and a tmRNA gene (collectively known as tRNA loci from this point onwards) were designated as GI integration hotspots, as each locus harboured a GI over 5.0 kb in at least one of the four genomes (see Table 2.1).

Table 2.1 The sixteen tRNA loci designated as hotspots for GI insertion after analysis of four of the completely sequenced *E. coli* genomes.

tRNA gene	<i>E. coli</i> K12 MG1655	<i>E. coli</i> UPEC CFT073	<i>E. coli</i> O157:H7 EDL933	<i>S. flexneri</i> 2a Sf301
<i>aspV</i>	- ^a	+	+	+
<i>thrW</i>	+ ^b	+	+	+
<i>serW</i>	-	-	+	-
<i>serT</i>	-	-	+	-
<i>serX</i>	-	+	+	-
<i>serU</i>	-	+	+	+
<i>asnT</i>	+	+	+	-
<i>argW</i>	+	+	+	+
<i>metV</i>	-	+	-	-
<i>glyU</i>	+	-	+	+
<i>pheV</i>	+	+	+	+
<i>selC</i>	-	+	+	+
<i>pheU</i>	-	+	-	-
<i>leuX</i>	+	+	+	+
<i>ssrA</i>	+	+	+	-
<i>asnV</i>	-	+	-	-

^a Indicates that the tRNA locus is either ‘empty’ or is occupied by an element less than 5.0 kb in length.

^b Indicates that the tRNA locus is occupied by a GI over 5.0 kb long.

2.1.3 tRNA core flanking regions

At each of the sixteen tRNA loci, across the four sequenced *E. coli* genomes, the core DNA sequences immediately flanking the empty tRNA loci and/or corresponding GI occupied loci were found to be highly conserved between strains (see Figure 2.1).

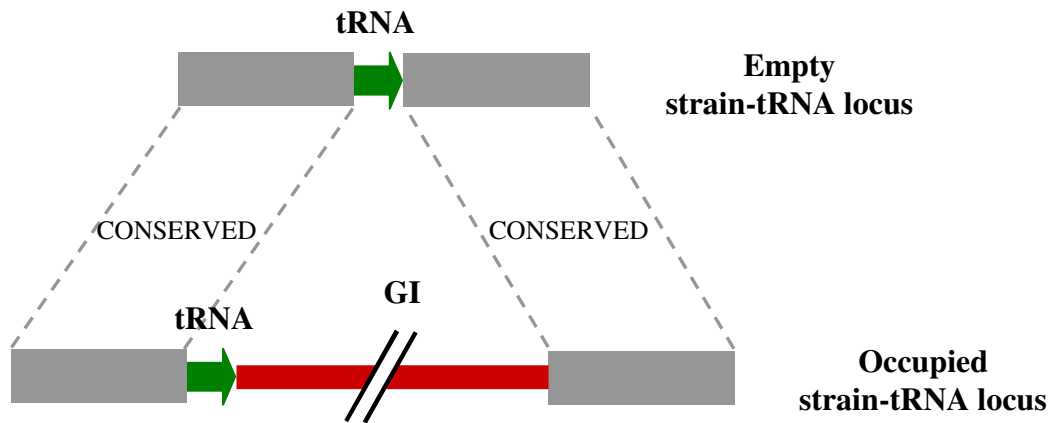


Figure 2.1. The cognate tRNA loci in a strain that harbours no GI and a strain harbouring a GI.

The grey boxes indicate the core flanking regions that are conserved from strain to strain.

Figure is not to scale.

2.2 tRNA site Interrogation for PAIs, prophages and other GIs (tRIP)

2.2.1 The principle of tRIP

As the core regions flanking the tRNA sites described in Table 2.1 were found to be well conserved across the sequenced *E. coli* genomes, it was hypothesised that they were also well conserved in unsequenced strains. Therefore, a high throughput PCR screen termed tRNA site Interrogation for PAIs, prophages and other GIs (tRIP) was designed to discover tRNA associated GIs across a panel of uncharacterised strains representative of the four ‘species’ of *Shigella* (see Table 2.3). tRIP is a negative-based PCR strategy, where a negative PCR result indicates the presence of island DNA between the 3’ terminus of the tRNA locus and the 5’ end of the corresponding downstream flanking region (Figure 2.2).

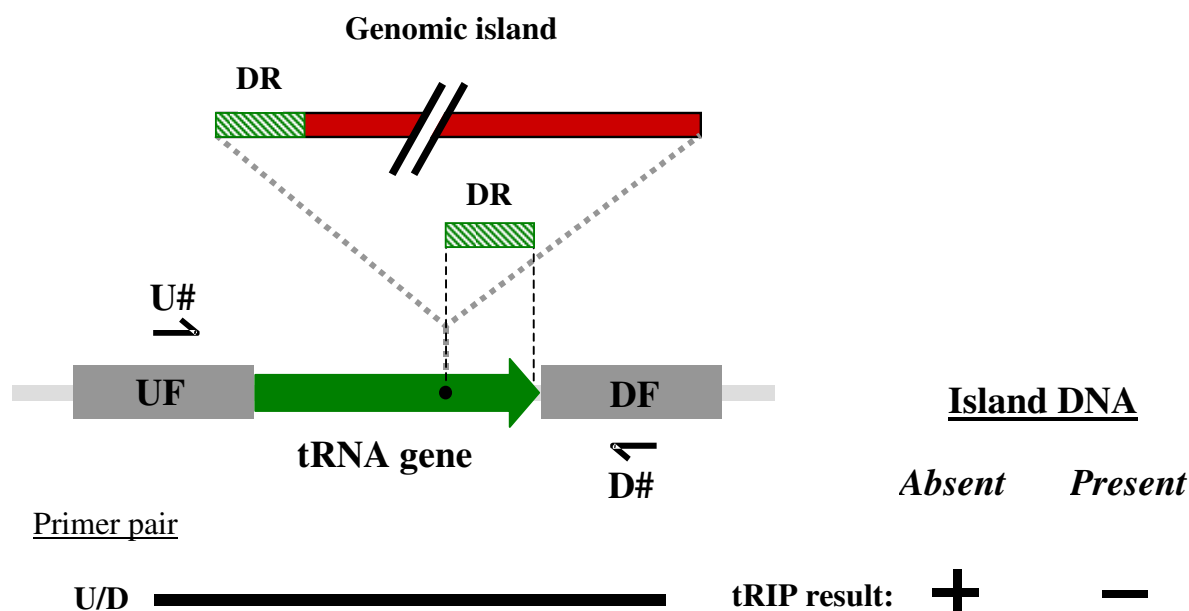


Figure 2.2 tRNA site Interrogation for PAIs and other GIs (tRIP).

Integration of a genomic island, bearing a single copy of a direct repeat (DR) corresponding to the sequence of the 3' terminus of the target tRNA gene, results in regeneration of the tRNA gene, but displaces the original DR and DF. Hence the negative PCR result with primer pair U/D, when a large integrated element is present. Plus and minus symbols indicate the generation or not, respectively, of PCR amplicons spanning regions defined by the lower most thick black line, depending on whether a GI is absent or present within the tRNA integration site (•). UF and DF stand for upstream flanking region and downstream flanking region respectively. U# and D# are specific oligonucleotide primers.

2.2.2 Multiple sequence alignments and primer design

Multiple sequence alignments (MSAs) of the 2 kb upstream (UF) and downstream flanking regions (DF) from each of the sequenced strains at each tRNA locus were created (Figure 2.3) using the ClustalW program (Chenna *et al.*, 2003) in the MegAlign software package (DNASTarTM). Conserved upstream (U#) and downstream (D#) oligonucleotide primer pairs for each locus were then designed using PrimerSelect software (DNASTarTM) which also checks primers for hairpin and primer dimer formation. Primers were checked with Blastn

with a word size of 7 in order to select primers that anneal the most specifically to the sequenced *E. coli* and *Shigella* genomes. Details of the primers are in Table A2. 1.

Sequence Name	< Pos = 157	< Pos = 1157
-  +		
<input checked="" type="checkbox"/> Consensus	TC AAGT GACG AGTTT GCG AGC AAA ACG ATG ATT AAG TGG CCG CTG GAA AGT ACA AGA ATC AGC ACA	AA TTT TTT G TTT TTT C
6 Sequences	150 170 180 190 200 210 220	1150 1170
UF <i>serU</i> K12 MG1655	TC AAGT GACG AGTTT GCG AGC AAA ACG ATG ATT AAG TGG CCG CTG GAA AGT ACA AGA ATC AGC ACA	AA ATT CT AAA AAT TC
UF <i>serU</i> EDL933	TC AAGT GACG AGTTT GCG AGC AAA ACG ATG ATT AAG TGG CCG CTG GAA AGT ACA AGA ATC AGC ACA	AA ATT CT AAA AAT TC
UF <i>serU</i> CIT073	TC AAGT GACG AGTTT GCG AGC AAA ACG ATG ATT AAG TGG CCG CTG GAA AGT ACA AGA ATC AGC ACA	AA TTT CT GTTT CT
UF <i>serU</i> Sf301	TC AAGT GACG AGTTT GCG AGC AAA ACG ATG ATT AAG TGG CCG CTG GAA AGT ACA AGA ATC AGC ACA	AA TTT CT GTTT CA
UF <i>serU</i> Sf2457T	TC AAGT GACG AGTTT GCG AGC AAA ACG ATG ATT AAG TGG CCG CTG GAA AGT ACA AGA ATC AGC ACA	AA TTT CT GTTT CA
UF <i>serU</i> Sakai	TC AAGT GACG AGTTT GCG AGC AAA ACG ATG ATT AAG TGG CCG CTG GAA AGT ACA AGA ATC AGC ACA	AA ATT CT AAA AAT TC

Figure 2.3 Multiple sequence alignment (MSA) of the conserved *serU* upstream flanking regions (UFs) in 6 sequenced genomes.

The shaded region indicates the selected conserved U primer. The position numbers indicate the location of the sequences relative to the 3' end of the cognate *serU* tRNA gene, with this terminus being labelled as position '1'.

2.2.3 tRIP Primer locations

The U and D Primers were selected to lie in close proximity to the putative GI, so that when potential amplicons were sequenced from the U or D primer the maximum amount of putative island DNA was read. However, to ensure that each amplicon could be confirmed as specific to the U or D flanking region, as a general rule at each tRNA locus the U# was located at least 100 bp from the 3' terminus of the associated tRNA gene (most tRNA genes are around 75 - 85 bp, the *ssrA* tmRNA gene is 363 bp long so the *ssrA* U# was placed further upstream) and the D# at least 100 bp from the 5' end of the DF (see Figure 2.4 and Table 2.2).

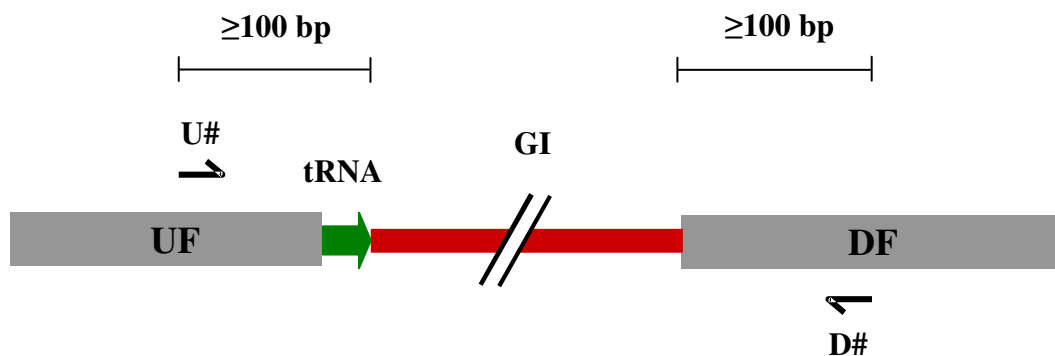


Figure 2.4. Schematic showing the minimum distance from the GI end of the conserved flanking regions that the U and D primers were located at each tRNA locus.

This ensured that sequencing from the U/D primer confirmed association of the putative GI with the corresponding flanking region. Figure is not to scale.

Table 2.2. Locations of the conserved U and D tRIP primers relative to the putative GI boundary.

tRNA gene	U# – tRNA 3' (bp)	D# – DF 5' (bp)
<i>aspV</i>	550 ^a	159
<i>thrW</i>	144	200
<i>serW</i>	676	746
<i>serT</i>	170	136
<i>serX</i>	266	637
<i>serU</i>	202	148
<i>asnT</i>	460	58 ^b
<i>argW</i>	326	587
<i>metV</i>	481	898
<i>glyU</i>	153	685
<i>pheV</i>	288	715
<i>selC</i>	379	566
<i>pheU</i>	548	201
<i>leuX</i>	931	828
<i>ssrA</i>	806	182
<i>asnV</i>	1374	604

^a Values correspond to the *E. coli* K12MG1655 genome.

^b The *asnT* D# was the only exception to the general rule.

2.3 *Shigella*, *E. coli* strains and plasmids used in this study

See Table 2.3 for details.

2.3.1 *Shigella* strains chosen for the tRIP screen

Twenty *Shigella* strains were screened in this study; those chosen represented strains across the four ‘species’ of *Shigella*. Each strain was given a code for identification.

Table 2.3. *Shigella*, *E. coli* strains and plasmids used in this study

Strain code	Original	Plasmids	Genotype/phenotype	Reference
<i>Shigella</i> strains used in tRIP screen				
S101	SBA 1394		<i>S. dysenteriae</i> 3 Ap ^r , Cm ^r , Sm ^r , Tc ^r	This study
S102	SBA 1395		<i>S. dysenteriae</i> 9	This study
S103	SBA 1396		<i>S. dysenteriae</i> 6	This study
S104	NCTC 0003		<i>S. flexneri</i> 1a	This study
S105	SBA 1173		<i>S. flexneri</i> 1b	This study
S106	SBA 1321		<i>S. flexneri</i> 2a	This study
S107	SBA 1300		<i>S. flexneri</i> 2b Ap ^r , Cm ^r , Sm ^r , Tc ^r	This study
S108	SBA 1407		<i>S. flexneri</i> 3a	This study
S109	SBA 1391		<i>S. flexneri</i> 4a	This study
S110	AL 124		<i>S. flexneri</i> 6 Ap ^r , Cm ^r , Sm ^r , Tc ^r	This study
S111	AL 132		<i>S. flexneri</i> X	This study
S112	AL 133		<i>S. flexneri</i> Y	This study
S113	SBA 1376		<i>S. sonnei</i>	This study
S114	SBA 1400		<i>S. sonnei</i> bio a	This study
S115	SBA 1399		<i>S. sonnei</i> bio g	This study
S116	SBA 1381		<i>S. boydii</i> 1	This study
S117	SBA 1382		<i>S. boydii</i> 2	This study
S118	SBA 1383		<i>S. boydii</i> 3	This study
S119	SBA 1384		<i>S. boydii</i> 4	This study
S120	SBA 1385		<i>S. boydii</i> 7	This study
Other strains				
			<i>E. coli</i> K12 MG1655	(Blattner <i>et al.</i> , 1997)
			<i>E. coli</i> CFT073	(Welch <i>et al.</i> , 2002)
			<i>E. coli</i> O157:H7 EDL933	(Perna <i>et al.</i> , 2001)
			<i>S. flexneri</i> 2a Sf301	(Jin <i>et al.</i> , 2002)
	DH5α TM		K12 derivative, <i>supE44</i> , <i>lacU169</i> , Δ 80 <i>lacZ</i> , Δ M15, <i>hsdR17</i> , <i>recA1</i> , <i>endA1</i> , <i>gyrA96</i> , <i>thi-1</i> , <i>relA1</i>	Bethesda Research Laboratories, 1986
KR144	CC118λ <i>pir</i>		K12 derivative, F- Δ (<i>ara-leu</i>)7697, <i>araD139</i> Δ □(<i>lac</i>)X74, <i>phoAd20</i> , <i>galE</i> , <i>galK</i> , <i>thi</i> , <i>rpsE</i> , <i>rpoB</i> , <i>argE</i> (Am), <i>recA1</i>	(Manoil and Beckwith, 1985)
PBA 1018	SM10λ <i>pir</i>		K12 derivative, <i>thi1</i> , <i>thr1</i> , <i>leuB6</i> , <i>supE44</i> , <i>tonA21</i> , <i>lacY1</i> ,	(Miller and Mekalanos,

			<i>recA</i> ::RP4-2-Tc::Mu Km ^r λpir	1988)
KR135	DH5α TM	pBluescript KS II (+)	Ap ^r	Stratagene
SBA 447	DH5α TM	pWSK29	Ap ^r	(Wang and Kushner, 1991)
KR143	SM10λpir	pRT733	Ap ^r , Km ^r	(Taylor <i>et al.</i> , 1989)
KR178	CC118λpir	pDS132	Cm ^r	Philippe <i>et al.</i> , 2004
KR147	DH5α TM	pJL1	pBluescript with 1040 bp of the <i>leuX</i> UF cloned into <i>Xba</i> I site. <i>Nsi</i> I site at 484 bp in the UF region. Ap ^r	This study
KR177	DH5α TM	pJL2	pJL1 with Km ^r cassette cloned into <i>Nsi</i> I site in <i>leuX</i> UF region, cassette is colinear with UF DNA. Ap ^r , Km ^r	This study
KR206	CC118λpir	pJL3	pDS132 with <i>leuX</i> construct from pJL2 cloned into <i>Xba</i> I site. Km ^r cassette reads in opposite direction to <i>sacB</i> . Cm ^r , Km ^r	This study
KR218	CC118λpir	pJL4	pDS132 with <i>leuX</i> construct from pJL2 cloned into <i>Xba</i> I site. Km ^r cassette is colinear with <i>sacB</i> . Cm ^r , Km ^r	This study
KR238	SM10λpir	pJL3	pDS132 with <i>leuX</i> construct from pJL2 cloned into <i>Xba</i> I site. Km ^r cassette reads in opposite direction to <i>sacB</i> . Cm ^r , Km ^r	This study
KR239	SM10λpir	pJL4	pDS132 with <i>leuX</i> construct from pJL2 cloned into <i>Xba</i> I site. Km ^r cassette is colinear with <i>sacB</i> . Cm ^r , Km ^r	This study
KR221	DH5α TM	pJL5	pBluescript with 959 bp of <i>argW</i> UF cloned into <i>Xba</i> I site. <i>Nsi</i> I site at 492 bp in the <i>argW</i> UF region. Ap ^r	This study
KR236	DH5α TM	pJL6	pJL5 with Km ^r cassette inserted into <i>Nsi</i> I site in <i>argW</i> UF region, cassette is colinear with UF DNA. Ap ^r , Km ^r	This study
KR237	DH5α TM	pJL7	pJL5 with Km ^r cassette cloned into <i>Nsi</i> I site in <i>argW</i> UF region, cassette reads in opposite direction to UF DNA. Ap ^r , Km ^r	This study
KR240	CC118λpir	pJL8	pDS132 with <i>argW</i> construct from pJL6 cloned into <i>Xba</i> I site. Km ^r cassette reads in opposite direction to <i>sacB</i> . Cm ^r , Km ^r	This study
KR241	CC118λpir	pJL9	pDS132 with <i>argW</i> construct from pJL6 cloned into <i>Xba</i> I site. Km ^r cassette is colinear with <i>sacB</i> .	This study

			Cm ^r , Km ^r pDS132 with <i>argW</i> construct from pJL6 cloned into <i>Xba</i> I site. Km ^r cassette reads in opposite direction to <i>sacB</i> . Cm ^r , Km ^r	This study
KR242	SM10 λ <i>pir</i>	pJL8		
KR243	SM10 λ <i>pir</i>	pJL9	pDS132 with <i>argW</i> construct from pJL6 cloned into <i>Xba</i> I site. Km ^r cassette is colinear with <i>sacB</i> . Cm ^r , Km ^r	This study
KR244	DH5 α TM	pJL10	pWSK29/ <i>Bam</i> HI::X101/ <i>Bam</i> HI marker rescue clone. 24.6 kb insert Ap ^r , Km ^r	This study
KR245	DH5 α TM	pJL11	pWSK29/ <i>Hind</i> III::X101/ <i>Hind</i> III marker rescue clone. 7.0 kb insert Ap ^r , Km ^r	This study
KR254	DH5 α TM	pJL17	pWSK29/ <i>Eco</i> RI::X106/ <i>Eco</i> RI marker rescue clone. 9.3 kb insert Ap ^r , Km ^r	This study
KR255	DH5 α TM	pJL18	pWSK29/ <i>Hind</i> III::X106/ <i>Hind</i> III marker rescue clone. 5.5 kb insert Ap ^r , Km ^r	This study
KR304	DH5 α TM	pJL19	pWSK29/ <i>Sac</i> I::X106/ <i>Sac</i> I marker rescue clone. 11.1 kb insert Ap ^r , Km ^r	This study
KR219	S101		S101 merodiploid with pJL3 inserted into <i>leuX</i> UF, position of cassette relative to MCS of pDS132 not confirmed	This study
KR220	S101		S101 merodiploid with pJL4 inserted into <i>leuX</i> UF, position of cassette relative to MCS of pDS132 not confirmed	This study
X101	KR220		S101 with Km ^r resistance cassette inserted into <i>leuX</i> UF, Ap ^r , Cm ^r , Sm ^r , Tc ^r , Km ^r	This study
X102	S116	pWSK29	Ap ^r	This study
X103	S116	pBluescript KS II (+)	Ap ^r	This study
X104	X102	pWSK29	X102 merodiploid with pJL9 inserted into <i>argW</i> U flank, <i>ori</i> , <i>cat</i> , <i>sacB</i> and MCS of pDS132 are between the <i>argW</i> tRNA and Km ^r resistance cassette, Ap ^r , Km ^r	This study
X105	X102	pWSK29	X102 merodiploid with pJL9 inserted into <i>argW</i> U flank, Km ^r resistance cassette is between the <i>ori</i> , <i>cat</i> , <i>sacB</i> and MCS of pDS132 and the <i>argW</i> tRNA. Ap ^r , Km ^r	This study
X106	X104	pWSK29	X102 with Km ^r resistance cassette inserted into <i>argW</i> UF, Ap ^r , Km ^r	This study

Frozen stocks of each strain were stored at -20°C and -80°C in Brain Heart Infusion (BHI) (Oxoid) containing 30% (v/v) glycerol. Strains were streaked from -20°C stocks onto Luria agar (LA), (see Appendix 1) and grown overnight at 37°C to single colonies. Plates were stored at $+4^{\circ}\text{C}$ for a maximum of 4 weeks.

2.4 Genomic DNA extraction

Modified phenol: chloroform extraction, miniprep method

(After Ausubel *et al.*, 1987)

A single colony was inoculated into 5 ml of Luria broth (LB) (see Appendix 1) (plus appropriate antibiotics) and incubated at 37°C overnight shaking at 200 rpm. 3 ml of the culture was harvested by centrifugation and the cells resuspended in 567 μl of TE buffer (Appendix 1), 30 μl of 10% (w/v) SDS (Appendix 1) and 3 μl of 20 mg/ml proteinase K (Sigma) were then added. The solution was mixed by inversion and incubated at 37°C for 45 min. After this, 2 μl of 10 mg/ml RNase (Sigma) was added, the solution mixed and incubated at 65°C for 45 min. Following this, another 2 μl of 20 mg/ml proteinase K was added and the solution was incubated at 37°C for 30 min. 100 μl of 5 M NaCl was added, the solution was mixed and 80 μl of preheated CTAB/NaCl solution was added (Appendix 1). The solution was mixed and incubated at 65°C for 10 min. DNA was extracted by adding an equal volume of chloroform:isoamylalcohol, followed by phenol:chloroform:isoamylalcohol (25:24:1) and a final chloroform:isoamylalcohol extraction. All of these steps require thorough mixing of the two phases, separation by centrifugation at $15700 \times g$ for 10 min and then the top aqueous phase is removed to a fresh tube and used in the next extraction. After the final extraction, the DNA was precipitated by the addition of 0.7 volumes of isopropanol and centrifugation at 13000 rpm for 30 min. The pellet was then washed twice with 500 μl of 70% (v/v in nH_2O) ethanol, air dried and resuspended in 100 μl of nH_2O . All genomic DNA was stored at -20°C .

2.5 Plasmid extraction

2.5.1 Modified alkaline lysis method

(After Morelle, 1989)

A single colony was inoculated into 5 ml of LB (plus appropriate antibiotics) and incubated at 37°C overnight shaking at 200 rpm. 3 ml of the culture was harvested by centrifugation and resuspended in 200 µl of 50 mM glucose, 25 mM Tris-HCL, 10 mM EDTA, 0.8 mg of lysozyme (Sigma) and incubated at room temperature for 10 min. Next 400 µl of fresh 0.2 M NaOH, 1% (w/v) SDS was added, the solution mixed by gentle inversion and incubated on ice for 5 min. Then 300 µl of 7.5 M ammonium acetate (Sigma) solution was added, the solution mixed and incubated on ice for 10 min. The mixture was then centrifuged at 13000 rpm for 10 min, 800 µl of the supernatant removed to a fresh tube and the plasmid DNA precipitated by adding 480 µl of isopropanol, incubated at room temperature for 10 min, then centrifuged at 13000 rpm for 30 min. The pellet was washed twice with 500 µl of ice cold 70% (v/v) ethanol, dried and resuspended in 50-100 µl nH₂O for high copy number plasmids and 10-20 µl for low copy number plasmids. Following this 2 µl of 10 mg/ml DNase free RNase was added and the tube left in the fridge overnight at +4°C. The plasmid was then finally stored at -20°C.

2.5.2 GenElute™ miniprep kit (Sigma)

This was used to obtain high yields of pure plasmid DNA for applications such as sequencing and sub-cloning. The kit uses a modified alkaline lysis protocol, after lysis the supernatant containing the plasmid DNA was spun through a column containing a silica-based membrane which the DNA binds to. The DNA was then washed to remove any contaminants and finally eluted using nH₂O into a clean eppendorf ready for further manipulation.

2.6 Restriction endonuclease digestion

Digests were performed in 20-100 μ l volumes depending on the amount and quality of DNA being digested. Between 2 and 10 units of enzyme were used per digest with the recommended reaction buffer, the volume of enzyme was never more than 10% of the total reaction volume in order to minimise star activity. Reactions were incubated at 37°C for 3 hr, then heat inactivated at the appropriate temperature. Enzymes were purchased from Roche, Promega or New England Biolabs. Prior to ligation, all plasmid digests were vacuum dried in a DNA-mini (Heto), washed twice with an equal volume of 70% (v/v) ethanol, vacuum dried and resuspended to 10 ng/ μ l in nH₂O. All other digests were cleaned as described and resuspended in 5-10 μ l of nH₂O. All digested DNA was stored at -20°C.

2.7 DNA ligation

T4 DNA ligase catalyses the joining of DNA molecules with compatible cohesive ends or blunt ends as long as one of the molecules has a 5'phosphate terminus. The reaction requires ATP and Mg²⁺ provided in the ligase reaction buffer.

Ligations to produce restriction libraries for SGSP-PCR (see 4.1) were performed in 20 μ l volumes as follows:

10 μ l (1.5 – 2 μ g) digested genomic DNA, cleaned

2 μ l (20 ng) digested pBluescript KS II (+), cleaned

2 μ l 10 x ligase buffer (Promega)

1 μ l (3 units) T4 DNA ligase (Promega)

5 μ l nH₂O

Reactions were incubated at 10°C for a maximum of 16 hr, then inactivated for 10 min at 70°C, vacuum dried, washed twice with 50 μ l of 70% (v/v) ethanol, vacuum dried and finally resuspended in 20 μ l of nH₂O. All other ligations were performed in 10-20 μ l volumes with an insert to vector ratio of at least 3:1, reaction conditions were the same as described and after cleanup, the DNA was resuspended in 5 μ l of nH₂O. All ligations were stored at -20°C.

2.8 Polymerase Chain reaction (PCR)

PCR is the *in vitro* amplification of the DNA sequence that lies between two oligonucleotide primers. PCR occurs in three steps, denaturation of the double stranded template DNA, followed by annealing of the single stranded DNA primers (usually around 20-30 nucleotides long), then extension of the 3' end of the primer sequences by incorporation of dNTPs to produce a copy of the target DNA. These steps are then cycled 25 to 30 times and as both strands of the DNA are amplified in each cycle, exponential amplification of the target DNA occurs. The extension step in PCR requires a thermostable DNA polymerase (*Taq*) that is the product of the *pol* gene from the thermophilic bacterium *Thermus aquaticus*. *Taq* polymerase, MgCl₂ and the 10 x reaction buffer were purchased from ABgene. dNTPs were purchased from Promega. Primers were synthesised by VHBio and MWG.

PCR reactions were performed in 20 µl volumes in a DNA Engine DYAD™ Peltier Thermal Cycler (MJ Research); reagents for each reaction were as follows:

MgCl₂: 1.2 µl of a 25 mM stock solution to give a final concentration of 1.5 mM.

10 x reaction buffer: 2 µl.

dNTPs: 1.6 µl of a 2.5 mM stock solution to give a final concentration of 0.2 mM.

Primer 1: 2 µl of a 10 µM stock solution.

Primer 2: 2 µl of a 10 µM stock solution.

nH₂O: 9.8 µl.

Taq: 0.4 µl (2 units).

Template DNA: 1 µl.

2.8.1 Standard PCR

1. 95°C for 5 min
2. 94°C for 30 sec
3. χ °C for 30 sec (annealing temperature, varied depending on the primer pair used, see Table A2. 1)
4. 72°C for γ min, 1 kb per min (extension time, varied depending on experiment, see results section and Table A2. 1)
5. Cycle to step 2 for 29 more times
6. 72°C for 10 min
7. 15°C forever

2.8.2 Hot-start, 'touchdown' PCR

The touchdown protocol used was a modification of the method used by Don *et al.*, 1991 (see 4.3.1).

1. 95°C for 15 min (activates the hot-start *Taq*)
2. 94°C for 30 sec
3. χ °C for 30 sec (annealing temperature, varied depending on the primer pair used, see Table A2. 1), starts at 10°C above the T_m of the primer with the lowest T_m of the pair.

Drop annealing temperature by 1 °C per cycle

4. 72°C for 2.5 min
5. Cycle to step 2 for 9 more times (reach T_m of primer after the 10 cycles)
6. 94°C for 30 sec
7. T_m °C for 30 sec
8. 72°C for 2.5 min
9. Cycle to step 6 for 19 more times
10. 72°C for 10 min
11. 15°C forever

2.8.3 Band-stab PCR

A modification of the technique by (Bjourson and Cooper, 1992) was used. After agarose gel electrophoresis, the gel surface was blotted dry using tissue paper, and the selected band stabbed using a pipette tip, which was then washed off into 20 µl of nH₂O in an eppendorf. The band-stab mixture was heated at 65°C for 10 min to melt any agarose and 1 µl used in a second round PCR.

2.8.4 Colony PCR

A single bacterial colony was resuspended in 20 µl of nH₂O, then heated in a boiling water bath for 10 min, spun at 13000 rpm for 5 min and 1 µl of the supernate used in a PCR.

2.8.5 Splicing by overlap extension (SOE) PCR

(Modification of the method used by (Horton *et al.*, 1989))

This technique was used to join by PCR two first round PCR products with homologous ends incorporated by primers in two separate initial standard PCRs. The purified primary PCR amplicons were then added as the template in a second round PCR that was initially run at low stringency (40°C) for 3 cycles to allow the denatured primary PCR products to anneal at their homologous ends. The fusion product then acts as template for future amplification cycles that are run at high stringency.

1. 95°C for 5 min
2. 94°C for 30 sec
3. 40°C for 30 sec (low stringency conditions)
4. 72°C for 1 min
5. Cycle to step 2 for 2 more times
6. 94°C for 30 sec
7. 64°C for 30 sec (stringent conditions, to amplify specifically the fusion product)
8. 72°C for 1 min

9. Cycle to step 6 for 24 more times
10. 72°C for 10 min
11. 15°C forever

2.9 Agarose gel electrophoresis

Agarose gels were prepared using 1x TAE buffer (Appendix 1) and 0.5 µg/ml ethidium bromide (Sigma) was added prior to casting the gel. In most cases 0.8% (w/v) agarose gels were used, if DNA fragments of less than 1 kb were to be well resolved, 1.5% (w/v) agarose was used. Before loading the DNA, 6x loading dye (Appendix 1) was added to each sample to a final concentration of 1x. This ensured that the samples sat in their individual wells and so that migration of the bands could be checked. Gel electrophoresis was performed in in-house made electrophoresis tanks with 1x Tris-acetate-EDTA (TAE) buffer (Appendix 1). Gels were viewed under UV illumination and photographed using the ID version 3.5.0 gel documentation package (Kodak). Agarose was purchased from Bioline.

2.10 Gel extraction

DNA was extracted from agarose gels using a clean scalpel, excess agarose was removed to minimise the gel volume and the DNA was then cleaned up immediately or stored in an eppendorf tube at -20°C for future cleanup.

2.11 DNA cleanup from agarose gel

DNA embedded in agarose was cleaned up using commercially available kits from either Qiagen or Yorkshire Biosciences, these use a column containing a silica-membrane which the DNA adsorbs to, any agarose, nucleotides and enzymes are washed out of the column and the pure DNA is then eluted using nH₂O into a clean eppendorf tube ready for future manipulation.

2.12 DNA sequencing

PCR products to be sequenced were gel extracted and cleaned (see above). Plasmids to be sequenced were extracted from the host strain using the Sigma GenElute™ kit (see 2.5.2). The DNA was then quantified and prepared according to the instructions stipulated by the sequencing company and sent to either MWG (www.mwg-biotech.com/html/all/index.php) or AGOWA (www.agowa.de). Both of the sequencing services used provided up to 1000 bp of quality sequence from a single sequence run, for further details on sequencing see section 4.7.

2.13 Preparation of eletrocompetent cells

2.13.1 Large-scale preparation of *E. coli* cells

A single colony was inoculated into 5 ml of LB and incubated at 37°C overnight shaking at 200 rpm. The 5 ml was inoculated into 500 ml of pre-warmed LB and grown at 37°C, 200 rpm to an OD_{600nm} of 0.5-0.6. The cells were then chilled on ice for 20 min and harvested by centrifugation at 4000 x g, at 4°C for 15 min. The supernate was discarded, the cells resuspended in 500 ml of sterile, ice cold, 10% (v/v) glycerol and centrifuged. The cells were resuspended to 250 ml and centrifuged, then resuspended to 20 ml and centrifuged. The cells were finally resuspended in 2 ml then separated into 100 µl aliquots and snap frozen in a dry ice/100% methanol bath and stored at -80°C ready for future use.

2.13.2 Small-scale preparation of *E. coli* and *Shigella* cells

A single colony was inoculated into 5 ml of LB and incubated at 37°C overnight shaking at 200 rpm. 1 ml was inoculated into 100 ml of pre-warmed LB and grown at 37°C, 200 rpm to an OD_{600nm} of 0.5-0.6. The cells were then chilled on ice for 5 min and harvested by centrifugation at 4000 x g, at 4°C for 10 min. The supernate was discarded, the cells resuspended in 25 ml of ice cold nH₂O and centrifuged. The cells were washed 3 more times

as above, then after the fourth wash were resuspended in 200 μ l of ice cold nH_2O and stored on ice ready for electroporation.

2.14 Electroporation

Up to 3 μ l of DNA to be transformed was mixed with a 50 μ l aliquot of electrocompetent cells in an ice cold eppendorf. The mixture was transferred to an ice cold, clean electroporation cuvette with an electrode gap of 2 mm (BioRad) and the cells electroporated at 1.5 kV with 1000 Ω resistance and 25 μ F capacitance using the Gene pulser II system (BioRad). A time constant of less than 18 indicated trace salt contamination. 1 ml of pre-warmed SOC (Appendix 1) was then added to the cells and the mixture transferred to a sterile universal and incubated at 37°C, shaking at 200 rpm for 60-90 min. The cells were then plated onto the appropriate selective medium and incubated overnight at the desired temperature.

2.15 Southern hybridization

2.15.1 Restriction digestion and gel electrophoresis

2 μ g of genomic DNA from each strain was digested in a total volume of 50 μ l with 10 units of *Hind*III for 16 h to allow for complete digestion, this was then vacuum dried, washed twice with 50 μ l of 70% (v/v) ethanol, vacuum dried and finally resuspended in 20 μ l of nH_2O . 10 μ l (1 μ g) of the DNA was electrophoresed in a 0.8% agarose gel using the protocol described in section 2.9, the DNA was run at 30 v for 6 h in order to get maximum resolution of the digested fragments then viewed under UV and photographed.

2.15.2 Southern transfer

The gel was depurinated by submerging in 250 mM HCl for 10 min, with gentle shaking, and then rinsed with nH_2O . The gel was washed in denaturation solution (0.5 M NaOH, 1.5 M NaCl) twice for 15 min, with gentle shaking and then rinsed with nH_2O . The gel was then washed in neutralisation solution (0.5 M Tris-HCl, 3 M NaCl, pH 7.5) twice for 15 min, with

gentle shaking. While this was happening the membrane (positively charged nylon membrane from Roche) was cut to be slightly larger than the area of the gel with the top left corner cut off also for orientation on the gel. The gel was then soaked in 20 x SSC (Appendix 1) for 3 minutes to equilibrate it. The DNA from the gel was then blotted to the membrane by capillary transfer using 20 x SSC overnight to ensure efficient transfer of the DNA (method taken from Ausubel *et al.*, 1987). The DNA was then fixed to the wet membrane by 1 x crosslink in a UV Stratalinker® 1800 (Stratagene). The membrane was then rinsed in nH_2O and air dried on a piece of blotting paper. At this stage the membrane can be stored flat between two pieces of blotting paper at 4 °C for future use, or used straight away for prehybridisation.

2.15.3 Labelling of the probe

The PCR generated probe was Digoxigenin (DIG)-labelled using DIG-High Prime® (Roche) which incorporates DIG-11-dUTP into the denatured template by using random-prime labelling. The PCR product was gel extracted, cleaned and quantified by UV spectroscopy to be 80 ng/ μl . Following the DIG-High Prime® instructions, ~ 1 μg of the DNA (enough for 2 Southern hybridisations) was used in a labelling reaction as follows:

13 μl template PCR product – 1040 ng

3 μl nH_2O

This was heated at 100°C for 10 min to denature the DNA, then chilled straight away on ice, 4 μl of DIG-High Prime® was added to the mixture, the tube centrifuged briefly and incubated at 37°C for 3 h.

2 μl of EDTA was added to stop the reaction and the mixture was then stored at – 20°C ready for use in the hybridisation.

2.15.4 Prehybridisation and hybridisation

The membrane was soaked in prehybridisation solution (Appendix 1) in a glass tube, rotating in a hybridisation oven at 65°C for 2 h. The DIG-labelled probe was then diluted in 12 ml prehybridisation solution (Roche DIG manual advises 20 ml per cm² of membrane, membrane was 60 cm²) to a concentration of 25 ng/ml. The prehybridisation solution was discarded and the hybridisation solution containing the DIG-labelled probe was added to the tube, and the probe hybridised to the DNA on the membrane, rotating in the hybridisation oven at 65°C for 12 h. The hybridisation solution was poured out of the tube and the membrane was then washed with 2 x wash solution (Appendix 1) twice for 5 min at room temperature. It was then washed with 0.5 x wash solution (Appendix 1) for 25 min at 65°C. The membrane was then ready for the detection step.

2.15.5 Detection

To detect the probe bound to the membrane, an alkaline phosphatase (AP) conjugated anti-DIG antibody (Anti-DIG-AP) which binds to the DIG-labelled probe was incubated with the membrane, then the chemiluminescent substrate CDP-Star® (Roche) was added. The substrate is dephosphorylated by the AP and decomposes causing it to emit light which was detected by X-ray film (Roche). For all of the following steps, powder free gloves were worn and the membrane was only manipulated by holding the very edges with sterile blunt ended forceps. First the membrane was equilibrated in washing buffer (Appendix 1) for 1 min, then in a fresh container the membrane was submerged in blocking solution (Appendix 1) with gentle shaking for 60 min. The membrane was placed in a freshly washed dish and submerged in blocking solution (Appendix 1) for 60 min with gentle shaking. Towards the end of this period the CDP-Star®, ready-to-use substrate was brought to room temperature and the antibody solution was prepared as recommended by Roche:

The Anti-DIG-AP was diluted 1:20,000 in blocking solution (0.5 µl of Antibody was added to 10 ml of blocking solution as this is enough to cover the surface of a minigel membrane).

The blocking solution was discarded and the antibody solution was poured over the surface of the membrane and incubated at room temperature for 30 min. The antibody solution was poured off and the membrane was submerged in washing buffer with gentle shaking, twice for 15 min. This was poured off and the membrane was equilibrated in detection buffer (Appendix 1) for 2 min. The membrane was then placed DNA side up on an A4 size piece of cling film and ~ 20 drops of CDP-Star®, ready-to-use substrate were poured all over the membrane (0.5 ml per 100 cm² of membrane). The membrane was then covered with another piece of cling film, creating a liquid seal, and then wiped gently with a damp tissue to remove any air bubbles from the surface of the membrane, which would affect the exposure step if left. This was incubated for 5 min at room temperature and then exposed to Lumi-Film (Roche) for 2 minutes. The film was then developed.

2.16 Allelic exchange

Allelic exchange was used to replace part of an original tRNA UF with a homologous region disrupted with an antibiotic resistance cassette. The tagged UF could then be used as a marker to clone larger regions of DNA spanning the associated GI. The homologous recombination step was performed using the suicide vector pDS132 (Philippe *et al.*, 2004) to introduce the disrupted UF into the chromosome of the *Shigella* test strain. Suicide vectors have a number of useful properties. One is a counterselectable marker, which causes the death of the host cell when expressed under specific conditions (Reyrat *et al.*, 1998). pDS132 contains the *sacB* gene from *Bacillus subtilis* that encodes levansucrase and converts sucrose to levans which is toxic to Gram-negative bacteria. Therefore when grown in the presence of sucrose, only cells that have lost the suicide plasmid from their chromosome will survive. It also contains a controllable origin of replication, the *oriR6K* that will only replicate in strains that synthesise the Pi protein which is encoded by the *pir* gene, present on the lysogenic phage λ *pir*. Therefore the plasmid and its derivatives can only be propagated in *E. coli* strains that carry the λ *pir* prophage. pDS132 also contains the *mob* region of RP4 so that the plasmid

can be mobilised into recipient cells during conjugation. However for this to occur extra transfer genes are also required, these are encoded by the RP4-2 derivative on the chromosome of *E. coli* SM10 λ pir, so the conjugation step must be performed using this strain as the donor. The vector also contains the *cat* gene so that cells harbouring pDS132 can be selected for by using chloramphenicol resistance.

2.16.1 Conjugation by filter mating

A single colony of the donor and recipient strains were inoculated separately into 5 ml LB with appropriate antibiotic selection and grown at 37°C, 200 rpm overnight. The cultures were diluted 100 fold into LB (with no antibiotics) and grown to an OD_{600 nm} of 0.8. 1 ml of each the donor and recipient were centrifuged at 8000 rpm for 2 min, the supernate discarded and donor pellet resuspended in 100 μ l of LB, then transferred to the recipient tube and the pellet resuspended. The mixture was spotted onto a sterile 0.45 μ m pore size nitrocellulose filter (Millipore) on an LA plate, and then incubated for 6 hr at 37°C. The filter was washed with 1 ml of LB into a falcon tube; serial dilutions to 10⁻⁷ were then spread onto LA + appropriate antibiotics to select for the *Shigella* strain and the resistance cassette. Plates were incubated at 37°C overnight. Transconjugant colonies were streaked to purity on the same medium.

2.16.2 Sucrose selection.

This was used to select for clones that had undergone a double homologous recombination event and lost the suicide vector from the chromosome. A single colony of a potential transconjugant was grown for 12 hr at 37°C, 200 rpm, in 5 ml LB + appropriate antibiotics. 1 ml of the culture was diluted to 10⁻⁵ to 10⁻⁷ – fold in LB and 100 μ l aliquots spread onto 6% (w/v) sucrose agar plates and incubated overnight at 30°C (Blomfield *et al.*, 1991). Sucrose resistant colonies were streaked to purity on LA + appropriate antibiotics and incubated overnight at 37°C. Sucrose resistant, antibiotic resistant clones were verified as true transconjugants by PCR spanning the entire region, to prove that they had lost the suicide

vector through a double-crossover event and retained the antibiotic resistance cassette in the UF.

2.17 Bacterial identification using the API[®] 20 E test system (bioMeriuex.Inc)

Some strains were identified using an API[®] 20 E test strip that is commonly used to identify members of the Enterobacteriaceae based on their core biochemical properties. The test strip contains of a number of plastic tubes that contain dehydrated substrates which become rehydrated after adding a suspension of the bacteria of interest.

A single fresh colony (up to 24 hr old) off an LA plate was emulsified into 5 ml sterile 0.85% (w/v) saline and this solution was used to fill each of the wells of the API[®] 20 E test strip according to the manufacturer's instructions. The strip was then covered with the supplied plastic lid and incubated at 35°C for 24 hr in a non-CO₂ incubator.

The reaction from each of the tubes was recorded onto the API[®] 20 E analytical profile index and this is compared against a database to provide the bacterial identification.

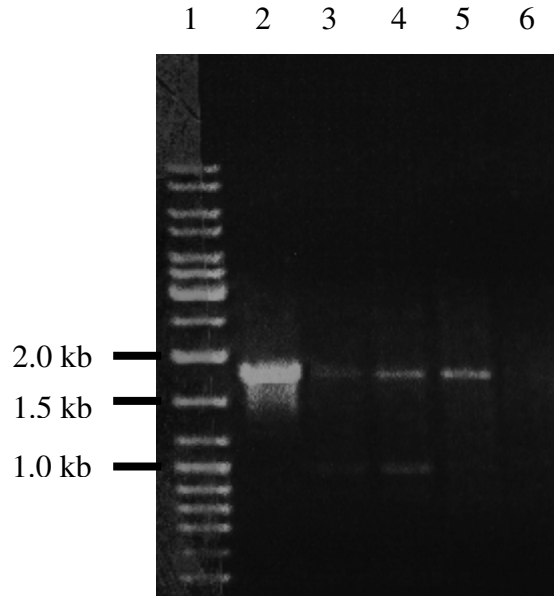
3.0 Development and optimisation of the tRIP screen

3.1 tRIP

tRIP PCR amplifications were performed using standard PCR (see 2.8.1) methodology. As tRIP is a negative-based PCR screen, used to detect the presence or absence of a tRNA borne GI rather than to simply amplify DNA, where appropriate both positive and negative PCR controls were performed at each tRNA locus. At most tRNA loci, the *in silico* tRIP analysis indicated that K12 MG1655 harboured either no island DNA or an islet under 5.0 kb. Therefore, under standard PCR conditions, a tRIP amplicon should be produced, so this strain could be used as the positive control for tRIP (see Table 3.1) to verify that the PCR conditions were optimised. CFT073 was found to harbour a GI over 5 kb at over half of the tRNA loci, so this was used as a negative control at these tRNA loci to show that the presence of a GI over 5 kb yields no PCR products. At sites where there was no positive control strain the tRIP PCR extension time was set to 3 min.

3.1.1 tRIP sensitivity and optimisation

Initially, 1 µl of the stock genomic DNA from each strain was used as the template in each PCR (at least 200 ng of genomic DNA); however, some PCRs yielded faint products and in some experiments, a faint tRIP amplicon was detected in the negative control strain. For example, *in silico* analysis using Artemis indicated that *E. coli* CFT073 has a 23.2 kb GI at the *serU* locus so therefore no amplicon should be produced by tRIP. However the *serU* tRIP screen produced a faint band that was the same size as the tRIP-positive control K12 MG1655 (see Figure 3.1). These products were unlikely to be non-specific amplicons as they were the same size as other strong tRIP-positive amplicons.



Lane number:

1. GeneRuler™ 1 kb ladder (Fermentas)
2. K12 MG1655 (positive control, expected an amplicon of 1.8 kb)
3. CFT073 (negative control, harbours a 23.2 kb GI at this locus so expected no amplicon)
4. S104
5. S116
6. S117

Figure 3.1. Agarose gel of an initial *serU* tRIP screen using over 200 ng of genomic DNA as template.

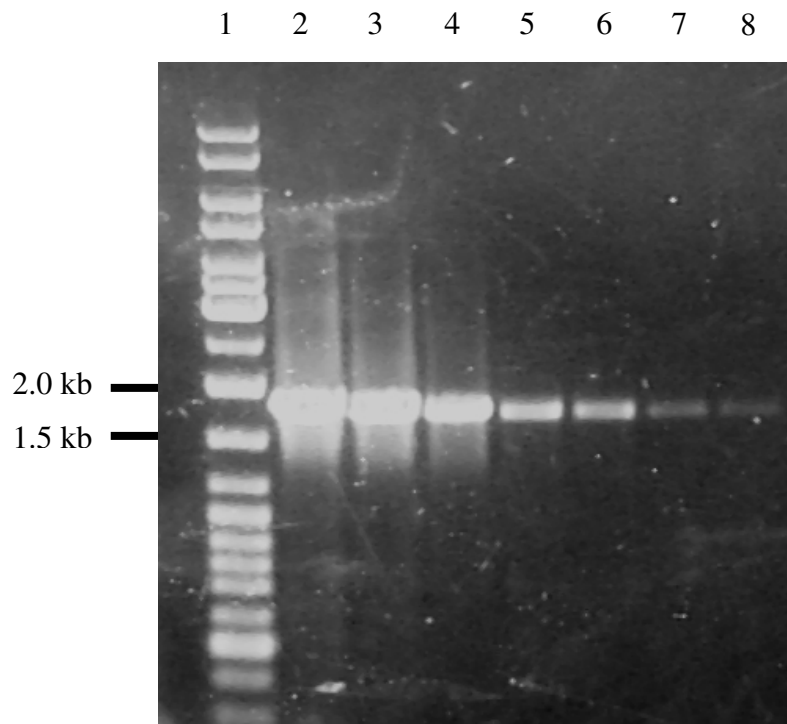
Lanes 3, 4 and 5 yielded faint products.

Each of the faint bands was verified by resolving a second aliquot of the original PCR products on a separate gel, to prove that the faint bands had not resulted from DNA carry-over from adjacent lanes.

The faint bands were likely to be false-positive tRIP amplicons and there were two possible reasons for their presence. One, is they were due to cross contamination of the stock genomic DNA with stock DNA from other strains that are truly tRIP positive at the same locus; possibly because the tubes were stored in close proximity in the same freezer box and were

opened near each other when adding the template to each PCR. Or secondly, in a very small proportion of the cells in the genomic DNA preparation, the GI had excised from the chromosome, to leave only a small islet between the U and D flanks as is present in K12 MG1655. GIs often delete at a high frequency as compared to core DNA mutation rates; *E. coli* core DNA has a spontaneous random mutation rate of 1 in 10^8 (Luria and Delbruck, 1943) , whereas entire GIs have been reported to delete at frequencies as high as 1 in 10^5 (Rajakumar *et al.*, 1997). Events of this nature would therefore be detected by the tRIP screen and seen as a faint band on the gel when a high concentration of template was used. This notion is reinforced further by the presence of a smaller faint tRIP band of 1.0 kb also found present in some of the lanes (see Figure 3.1); suggesting that the GI may also excise to leave a slightly smaller islet between the U and D flanks. This phenomenon was also observed with some of the initial *aspV* and *asnV* tRIP screens.

These results indicated that the tRIP screen is a highly sensitive assay; therefore the regime was optimised by testing the detection limit of tRIP by performing a titration experiment. The genomic DNA from *E. coli* K12 MG1655, which is known to be tRIP-positive at the *serU* locus, was diluted, and then the *serU* tRIP screen was run with the different concentrations of template (see Figure 3.2).



Lane number:

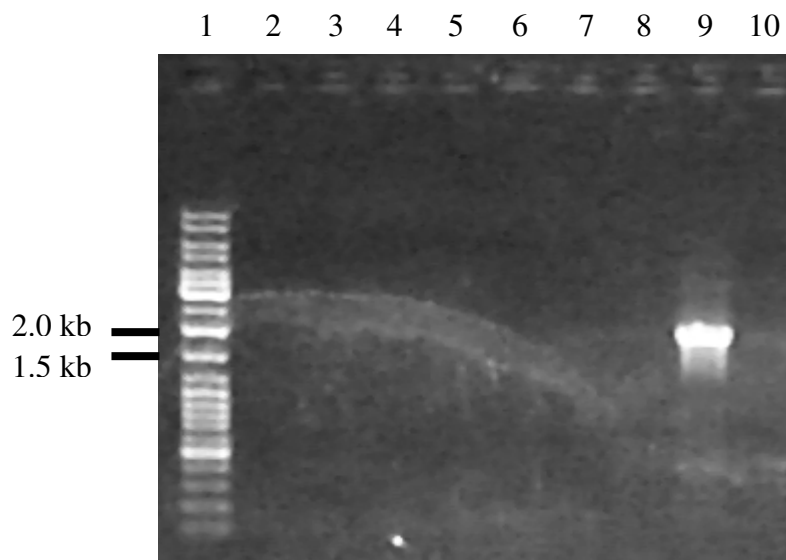
1. GeneRuler™ 1 kb ladder
2. 100 ng template
3. 10 ng template
4. 1 ng template
5. 100 pg template
6. 50 pg template
7. 10 pg template
8. 5 pg template

Figure 3.2. Agarose gel showing the *serU* tRIP sensitivity assay using different concentrations of *E. coli* K12 MG1655 genomic DNA as template.

The results from Figure 3.2 show that even with only 5 pg of template, a tRIP product was detected and with 1 ng, a strong tRIP amplicon was produced. This indicates how sensitive the tRIP screen is (only ~ 1000 copies of the genome are required to produce a visible product when viewed on an agarose gel, as the weight of 1 genome of K12 MG1655 is ~5 fg). This experiment was also performed at the *aspV* locus (data not shown); a product was not detected at concentrations lower than 1 ng, possibly because the MG1655 *aspV* tRIP amplicon was 3120 bp, 1.3 kb larger than the *serU* tRIP product.

Therefore, to standardise the tRIP screen, 10-15 ng of genomic DNA was used as the amount of template in all future tRIP PCRs. The stock DNA from each strain was serially diluted to 10-15 ng/ μ l and these 'tRIP DNA' tubes were stored in a separate freezer box to all other DNA.

This strategy resulted in eliminating faint bands from most tRIP PCRs (see Figure 3.3), so that truly GI-occupied loci yielded no amplicons. This is because at the tRIP-negative strain-tRNA loci, the specific DNA template that amplified the faint bands was in the minority and was undetectable by PCR after dilution, whereas truly tRIP positive strain-tRNA loci still produced a strong amplicon because there was sufficient template



Lane number:

1. GeneRuler™ 1 kb ladder
2. CFT073 (negative control)
3. S104.....-ve
4. S116.....-ve
5. S117.....-ve
6. S106.....-ve
7. S107.....-ve
8. S118.....-ve
9. S114.....+ve
10. S105.....-ve

Figure 3.3. Agarose gel of the *serU* tRIP PCR screen after dilution of the template DNA to 10-15 ng/ μ l.

However, very rarely, in some PCRs, faint bands of the same size as other tRIP-positive strain-tRNA loci were detected, these were regarded as tRIP-negative and later were all confirmed as harbouring GIs by SGSP-PCR (see Table 3.2 and Table 5.1). These results may highlight the variable degree of instability that GIs exhibit and that some delete at higher frequencies.

At some strain-tRNA loci, tRIP amplicons of an unexpected size were sometimes produced (see Table A2. 5). Some of the larger amplicons (>4.0 kb) were fainter than other, smaller tRIP amplicons generated from different strains at the same locus; however, this was more likely due to the efficiency of the PCR regime and not because they were non-specific or false-positives. For this reason, they were regarded as tRIP-positive. Of the strain-tRNA loci that yielded larger tRIP amplicons, those that were characterised through sequencing of the tRIP amplicon from either the cognate U or D primer or by SGSP-PCR and subsequent sequencing, were all found to harbour island DNA.

3.2 tRIP results

3.2.1 *In Silico* tRIP results

The U and D primers were used in an *in silico* tRIP PCR with the four fully sequenced *E. coli* and two *Shigella* genomes available across the sixteen tRNA loci (see Table 3.1). Using Blastn, the coordinates of each primer was located, and the size of the putative tRIP amplicon calculated for each primer pair. This gave an insight into the size of each of the associated GIs, the relative occupancy of each strain-tRNA locus, and indicated which strains could be used as positive controls for the tRIP screen at each locus. It also highlighted any anomalous sites (see Figure 3.4).

Table 3.1. *In silico* tRNA site interrogation for PAIs (tRIP) of the published complete *E. coli* and *Shigella* genomes, across the sixteen tRNA loci designated as hotspots for GI insertion.

tRNA gene	Empty tRNA site	<i>E. coli</i> K12 MG1655	<i>E. coli</i> UPEC CFT073	<i>E. coli</i> O157:H7 EDL933	<i>E. coli</i> O157:H7 Sakai	<i>S. flexneri</i> 2a Sf301	<i>S. flexneri</i> 2a 2457T
<i>aspV</i>	0.7 ^a	3.1 ^b	100.7	37.6	37.6	58.4	61.8
<i>thrW</i>	0.3	40.2	8.0	35.5	35.5	DF (-)	DF (-)
<i>serW</i>	1.1	1.4	1.4	89.0	1.4	1.5	1.5
<i>serT</i>	0.3	0.3	0.3	45.6	50.0	0.3	0.3
<i>serX</i>	0.4	0.9	DF (-) ^c	87.9	86.6	0.4	0.4
<i>serU</i>	0.4	1.8	23.5	47.0	45.8	22.7	21.3
<i>asnT</i>	0.5	10.6	37.8	11.6	11.6	5.0	5.0
<i>argW</i>	0.9	13.7	15.7	15.3	15.3	6.5	6.5
<i>metV</i>	1.4	1.4	34.0	1.4	1.4	1.3	1.3
<i>glyU</i>	0.7	12.4	0.8	28.4	28.4	10.7	9.5
<i>pheV</i>	1.0	10.1	128.9	DF (-)	DF (-)	56.1	52.6
<i>selC</i>	1.0	2.9	69.6	44.6	44.6	DF (-)	DF (-)
<i>pheU</i>	0.6	0.6	52.8	0.6	0.6	0.6	0.6
<i>leuX</i>	1.8	41.9	17.7	46.2	46.2	9.3	10.4
<i>ssrA</i>	1.0	30.6	D (-)	30.2	32.8	4.6	4.6
<i>asnV</i>	2.0	2.0	UF (inv) ^d	2.0	2.0	2.0	2.0

^a Values are to the nearest 0.1 kb

^b Values highlighted in bold indicate that the corresponding strain was used as the positive control for the tRIP PCR screen at the cognate tRNA locus

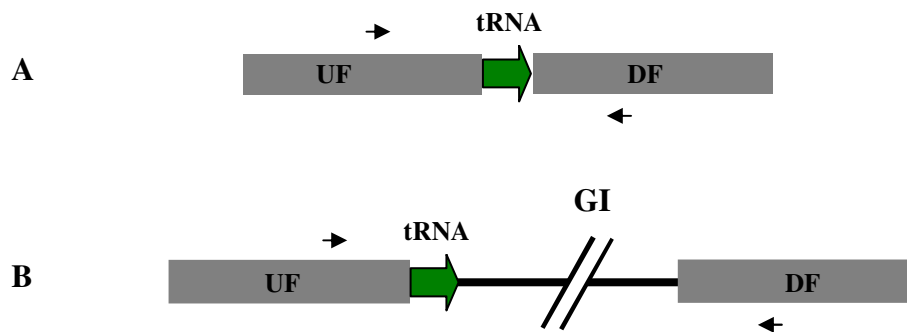
^c *In silico* tRIP PCR results marked with the codes listed in Fig. 3.4 indicate that there is no amplicon.

^d See section A2.7 for details of the *asnV* UF inversion.

3.2.2 Orientation of flanking regions

The *in silico* tRIP screen highlighted a number of anomalous sites where the upstream flanking region (UF) or downstream flanking region (DF) was found in different conformations to what was expected, these are described in Figure 3.4.

Normal orientation of the conserved flanking regions:



Other orientations discovered:

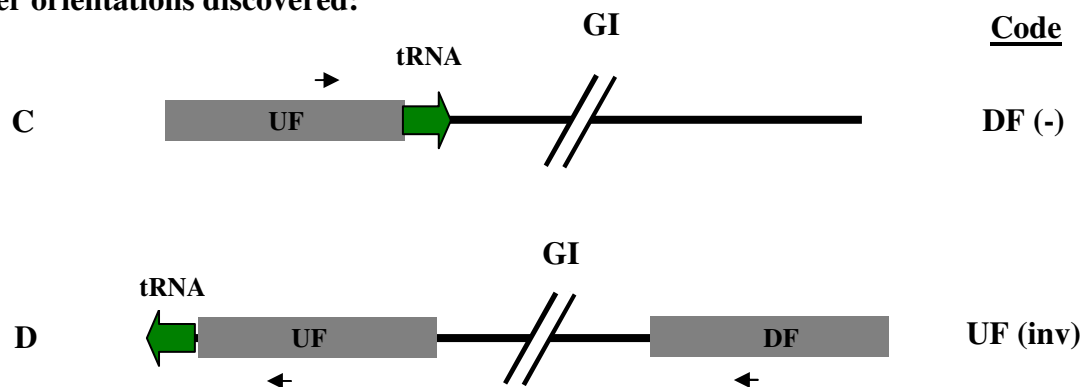


Figure 3.4. Alternative arrangements of the conserved upstream flanking region (UF) and conserved downstream flanking region (DF) relative to each other, found by the *in silico* tRIP screen of the six *E. coli* and *Shigella* genomes.

The small arrows indicate primer orientations. A: empty tRNA locus, B: occupied tRNA locus, C: downstream conserved flanking region is deleted, D: upstream conserved flanking region is inverted. Figure is not to scale.

This raised the question as to whether in some instances a negative tRIP could be a false-negative due to the absence, or inversion of one of the primers relative to the other at an empty tRNA locus. However, further *in silico* studies showed that at all tRNA loci where these events have occurred, there is always island DNA present, therefore the negative tRIP is still representative of the occupied status of the locus.

3.2.3 tRIP screen results across the *Shigella* strains

The results of the tRIP screen are shown in Table 3.2 and Table 3.3.

Table 3.2. tRNA site interrogation for PAIs (tRIP) screen across the 16 tRNA loci designated as GI insertion hotspots in twenty unsequenced *Shigella* strains representative of the four ‘species’ and two sequenced *E. coli* control strains.

Strain	code	tRNA locus																tRIP profile ^c
		<i>aspV</i>	<i>thrW</i>	<i>serW</i>	<i>serT</i>	<i>serX</i>	<i>serU</i>	<i>asnT</i>	<i>argW</i>	<i>metV</i>	<i>glyU</i>	<i>pheV</i>	<i>selC</i>	<i>pheU</i>	<i>leuX</i>	<i>ssrA</i>	<i>asnV</i>	
<i>E. coli</i> K12 MG1655		+	-	+	+	+	+	-	-	+	-	-	+	+	-	-	+	8
<i>E. coli</i> CFT073		-	-	+	+	-	-	-	-	-	+	-	-	-	-	-	-	13
<i>S. dysenteriae</i> 3	S101	+ ^a	- ^b	-	+	-	-	-	-	-	-	-	+	-	-	-	+	1
<i>S. dysenteriae</i> 9	S102	+	-	+	+	-	-	-	+	-	-	-	+	-	-	-	+	2
<i>S. dysenteriae</i> 6	S103	+	-	+	+	-	-	-	-	-	-	-	+	-	-	-	+	3
<i>S. flexneri</i> 1a	S104	-	-	+	+	+	-	-	-	+	-	-	-	+	-	+	+	4
<i>S. flexneri</i> 1b	S105	-	-	+	+	+	-	-	-	+	-	-	-	+	+	+	+	5
<i>S. flexneri</i> 2a	S106	-	-	+	+	+	-	-	-	+	-	-	-	+	-	+	+	4
<i>S. flexneri</i> 2b	S107	-	-	+	+	-	-	-	-	+	-	-	-	+	-	+	+	6
<i>S. flexneri</i> 3a ^d	S108	-	-	+	+	-	+	-	-	+	-	-	-	+	-	-	+	7
<i>S. flexneri</i> 4a ^e	S109	+	-	+	+	+	+	-	-	+	-	-	+	+	-	-	+	8
<i>S. flexneri</i> 6	S110	+	-	+	+	-	-	-	+	-	+	-	+	-	+	-	+	9
<i>S. flexneri</i> X	S111	-	-	+	+	+	-	-	-	+	+	-	-	+	-	+	+	10
<i>S. flexneri</i> Y	S112	-	-	+	+	+	-	-	-	+	+	-	-	+	+	+	+	11
<i>S. sonnei</i>	S113	-	-	+	+	-	+	-	-	+	-	-	-	+	-	-	+	7
<i>S. sonnei</i> bio a	S114	-	-	+	+	-	+	-	-	+	-	-	-	+	-	-	+	7
<i>S. sonnei</i> bio g	S115	-	-	+	+	-	+	-	-	+	-	-	-	+	-	-	+	7
<i>S. boydii</i> 1	S116	+	-	+	+	-	-	-	-	-	-	-	+	-	-	-	+	3
<i>S. boydii</i> 2	S117	+	-	+	+	-	-	-	+	-	-	-	+	-	-	-	+	2
<i>S. boydii</i> 3	S118	+	-	+	+	-	-	-	+	-	-	-	+	-	-	-	+	2
<i>S. boydii</i> 4	S119	+	-	+	+	-	-	-	+	-	-	-	+	-	-	-	+	2
<i>S. boydii</i> 7	S120	-	-	+	+	-	-	-	-	-	+	-	+	-	-	-	-	12

^a Positive tRIP PCR, with a GI of no greater than 5 kb at this strain-tRNA locus, see table 3.3 for full details

^b Negative tRIP PCR, indicating the presence of a putative GI at this strain-tRNA locus

^c Each strain was assigned a tRIP profile referring to the pattern of occupancy of its tRNA loci

^d tRIP profile suggested that S108 was *S. sonnei*-like. After characterisation with SGSP-PCR (see section 4.1) and an API[®] 20 E test strip, S108 was re-classified as a *S. sonnei* (see section 5.7 for more details)

^e tRIP profile suggested that S109 was *E. coli* K12 MG1655-like. After characterisation with SGSP-PCR (see section 4.1) and an API[®] 20 E test strip, S109 was re-classified as an atypical *E. coli* and omitted from the study (see section 5.6 for more details)

Table 3.3. tRIP screen across the 16 tRNA loci designated as GI insertion hotspots in twenty *Shigella* strains representative of the four ‘species’, with the tRIP amplicon sizes indicated.

Strain	code	tRNA locus																tRIP profile
		<i>aspV</i>	<i>thrW</i>	<i>serW</i>	<i>serT</i>	<i>serX</i>	<i>serU</i>	<i>asnT</i>	<i>argW</i>	<i>metV</i>	<i>glyU</i>	<i>pheV</i>	<i>selC</i>	<i>pheU</i>	<i>leuX</i>	<i>ssrA</i> (tmRNA)	<i>asnV</i>	
<i>E. coli</i> K12 MG1655		3120 ^a	-	1422	306	903	1790	-	-	1379	-	-	2850	749	-	-	1978 # set1, 564 # set 2 ^c	8
<i>E. coli</i> CFT073		- ^b	-	1422	306	-	-	-	-	-	838	-	-	-	-	-	-	13
<i>S. flexneri</i> 2a Sf301 ^d		-	-	1481	306	369	-	-	-	1270	-	-	-	749	-	4640	1978	4
<i>S. dysenteriae</i> 3	S101	+ 1.7 kb ^f	-	-	+	-	-	-	-	-	-	-	+ 3.3 kb	-	-	-	+ 2.0 kb	1
<i>S. dysenteriae</i> 9	S102	+ 1.7 kb	-	+ ^e	+	-	-	-	+ 2.6 kb	-	-	-	+ 3.3 kb	-	-	-	+ 2.0 kb	2
<i>S. dysenteriae</i> 6	S103	+ 1.7 kb	-	+	+	-	-	-	-	-	-	-	+ 3.3 kb	-	-	-	+ 2.0 kb	3
<i>S. flexneri</i> 1a	S104	-	-	+	+	+ 0.4 kb	-	-	-	+ 1.3 kb	-	-	-	+	-	+ 4.8 kb	+ 2.0 kb	4
<i>S. flexneri</i> 1b	S105	-	-	+	+	+ 0.4 kb	-	-	-	+ 1.3 kb	-	-	-	+	+ 6.0 kb	+ 4.8 kb	+ 2.0 kb	5
<i>S. flexneri</i> 2a	S106	-	-	+	+	+ 0.4 kb	-	-	-	+ 1.3 kb	-	-	-	+	-	+ 4.8 kb	+ 2.0 kb	4
<i>S. flexneri</i> 2b	S107	-	-	+	+	-	-	-	-	+ 1.3 kb	-	-	-	+	-	+ 4.8 kb	+ 2.0 kb	6
<i>S. flexneri</i> 3a	S108	-	-	+	+	-	+	-	-	+	-	-	-	+	-	-	+ 0.6 kb	7
<i>S. flexneri</i> 4a	S109	+	-	+	+	+	+	-	-	+	-	-	+	+	-	-	+ 2.0 kb	8
<i>S. flexneri</i> 6	S110	+ 1.7 kb	-	+	+	-	-	-	+ 1.6 kb	-	+ 1.2 kb	-	+ 4.8 kb	-	+ 4.5 kb	-	+ 0.6 kb	9
<i>S. flexneri</i> X	S111	-	-	+	+	+ 0.4 kb	-	-	-	+ 1.3 kb	+ 3.4 kb	-	-	+	-	+ 4.8 kb	+ 0.6 kb	10
<i>S. flexneri</i> Y	S112	-	-	+	+	+ 0.4 kb	-	-	-	+ 1.3 kb	+ 3.4 kb	-	-	+	+ 6.0 kb	+ 4.8 kb	+ 0.6 kb	11
<i>S. sonnei</i>	S113	-	-	+ 1.5 kb	+	-	+	-	-	+	-	-	-	+	-	-	+ 0.6 kb	7
<i>S. sonnei</i> bio a	S114	-	-	+ 1.5 kb	+	-	+	-	-	+	-	-	-	+	-	-	+ 0.6 kb	7
<i>S. sonnei</i> bio g	S115	-	-	+ 1.5 kb	+	-	+	-	-	+	-	-	-	+	-	-	+ 0.6 kb	7
<i>S. boydii</i> 1	S116	+ 1.7 kb	-	+	+	-	-	-	-	-	-	-	+ 4.8 kb	-	-	-	+ 0.6 kb	3
<i>S. boydii</i> 2	S117	+ 1.7 kb	-	+	+	-	-	-	+ 1.6 kb	-	-	-	+ 4.8 kb	-	-	-	+ 0.6 kb	2
<i>S. boydii</i> 3	S118	+ 1.7 kb	-	+	+	-	-	-	+ 2.5 kb	-	-	-	+ 5.0 kb	-	-	-	+ 2.0 kb	2
<i>S. boydii</i> 4	S119	+ 1.7 kb	-	+	+	-	-	-	+ 1.6 kb	-	-	-	+ 4.8 kb	-	-	-	+ 0.6 kb	2
<i>S. boydii</i> 7	S120	-	-	+	+	-	-	-	-	-	+ 4.4 kb	-	+	-	-	-	-	12

^a The exact *in silico* tRIP amplicon lengths in bp are shown in the three control strains

^b Control Strain-tRNA loci with GIs over 5 kb, were marked as -ve

^c Two different primer sets were used to account for possible inversion of part of the UF in the test strains, see section A2.7 for further details

^d The Sf301 *in silico* values were added to the table to help with analysis of different size tRIP products

^e Strain-tRNA loci yielding tRIP amplicons the same length as K12 MG1655 to within 0.1 kb, were marked as +ve

^f Strain-tRNA loci yielding tRIP amplicons of different lengths to K12 MG1655 were marked as +ve, with the length of the amplicon indicated to the nearest 0.1 kb. For more details on these amplicons see table A2.5.

4.0 Characterisation of the extremities of tRNA associated islands identified by tRIP

4.1 Interrogation of tRNA borne GIs using SGSP-PCR

Putatively GI occupied (tRIP-negative) strain-tRNA loci (meaning a specific tRNA locus in a specific strain, for example S101-*serU*) were interrogated further using Single Genome Specific-Primers-PCR (SGSP-PCR) (Figure 4.1). The technique uses individual restriction endonuclease libraries of each test strain as the template for each PCR. Ideally, each library would contain a cloned fragment bearing the U or D flanking region and a portion of the associated putative GI. PCR with the U or D primer and a vector primer was then used to produce an amplicon that ‘walked’ into the U or D ‘arm’ of the inserted GI (U-arm or D-arm meaning the part of the GI that is adjacent to the UF or DF respectively). Subsequent sequencing of the amplicon provided a snapshot of its features. For each strain, five different enzyme libraries were generated to increase the chances of producing specific amplicons that extended beyond the flank and into putative island DNA. A further two libraries were used to investigate the remaining uncharacterised strain-tRNA loci.

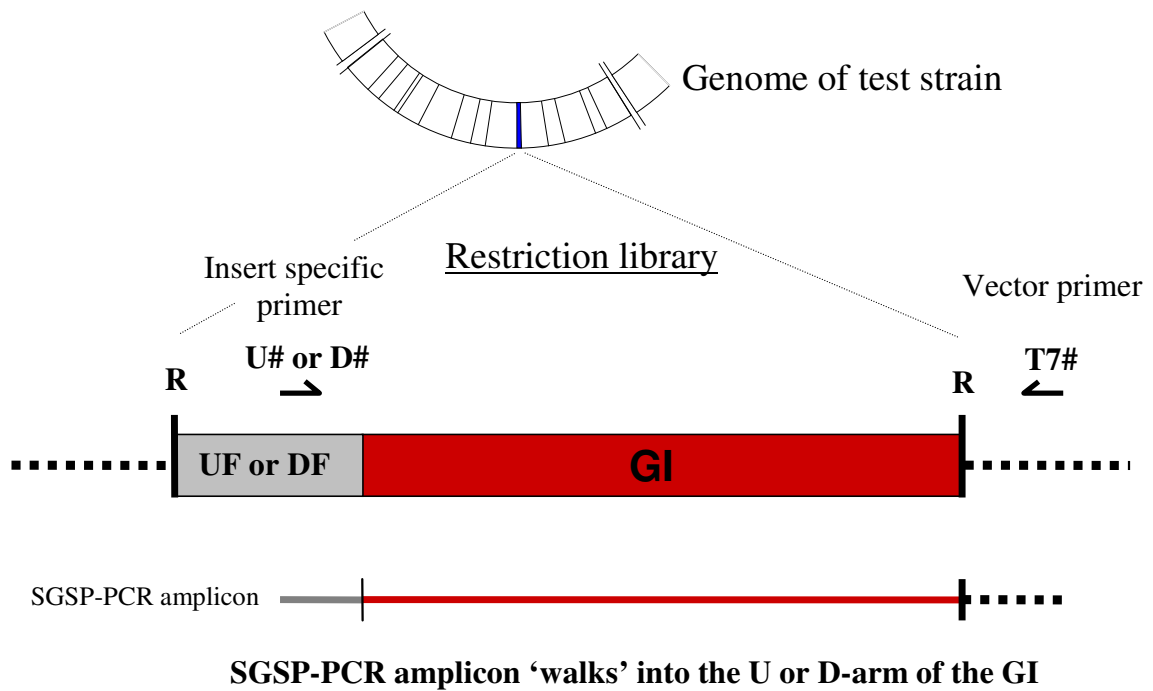


Figure 4.1. Single Genome-Specific Primer-PCR (SGSP-PCR).

R, restriction endonuclease cut sites; thick broken lines, vector DNA. Figure is not to scale.

The SGSP-PCR protocol was optimised to ensure it was a high-throughput technique, by using the strategies described below.

4.1.1 Enzyme choice

It was important when selecting the restriction endonucleases to generate the genomic libraries used as the template for the SGSP-PCRs, to look at the number and length of the restriction fragments produced by the enzymes, as if they produced too many small fragments, the SGSP-PCR amplicons produced would not walk far enough into the GI to provide valuable information. Whereas fragments that were too few and large (over 45 kb) would be less likely to lead to successful SGSP-PCR amplification using standard PCR, as the chances of the insert-specific primer binding too far from the corresponding downstream restriction site become greater as the fragment size increased. In addition, the chance of the fragment being sheared increases as the size increases. Therefore, a study of the potential candidates was carried out to check their suitability and it was found that 6-base cutting enzymes would

be the most useful for SGSP-PCR, after a study of the frequency that *Hind*III cuts the K12 MG1655 chromosome and the fragment sizes it produces (see Figure 4.2).

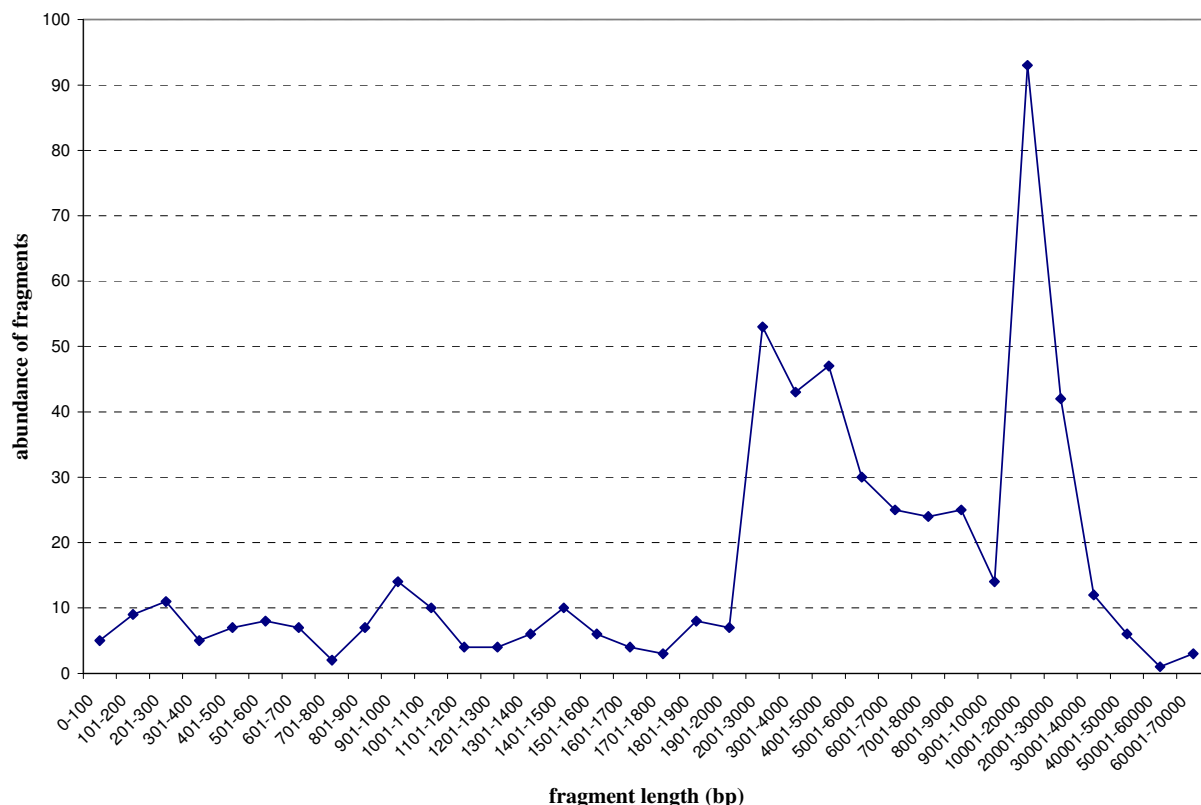


Figure 4.2. Distribution of *Hind*III fragments across the *E. coli* K12 MG1655 chromosome.

Figure 4.2 shows that over 70% of the *Hind*III fragments are between 2 kb and 30 kb, which is suitable for SGSP-PCR as the majority are below the maximum clonable fragment size for pBluscript KS II (+), but large enough to produce amplicons that walk beyond the flank and into potential island DNA. Therefore, four more enzymes that cut with a similar frequency and distribution to *Hind*III were also chosen to generate the restriction libraries for SGSP-PCR (Table 4.1). All five of the enzymes are readily available, cost effective, 6-base cutters.

Table 4.1. The five enzymes used to generate genomic libraries for SGSP-PCR

Restriction enzyme	<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> II
Recognition sequence (5'-3')	G'GATCC	G'AATCC	A'AGCTT	CTGCA'G	G'TCGAC
No of sites in K12 MG1655	495	645	556	958	544

It should be noted that this data was generated using the distribution of restriction sites across the *E. coli* K12 MG1655 chromosome only and that the enzyme choices are based on the assumption that all of the test *E. coli* strains will have a similar distribution.

4.2 *In silico* SGSP-PCR

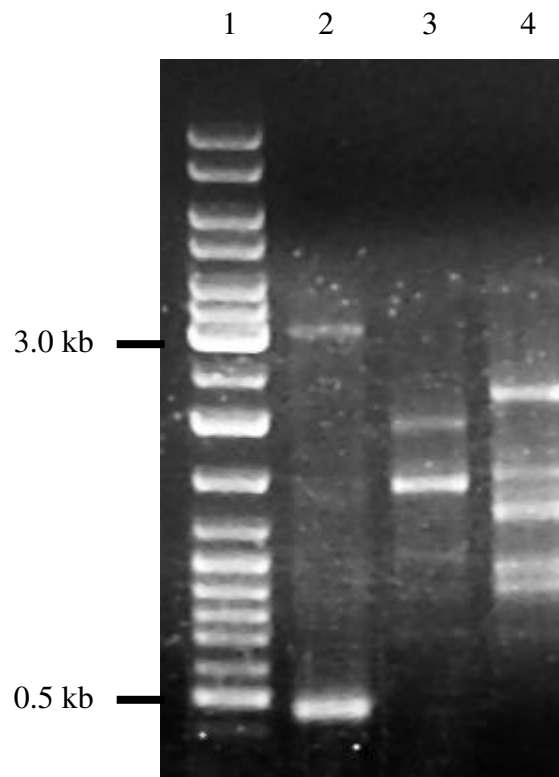
In silico SGSP-PCR was performed using *E. coli* K12 MG1655 as the template genome, so that the corresponding K12 MG1655 libraries could be selected to use as positive controls when performing SGSP-PCR with the *Shigella* strains to prove that the PCR mix was viable (see Table A2. 3 and

Table A2. 4).

4.3 SGSP-PCR regime

Standard PCR was used initially for the SGSP-PCRs, with an extension time of 3 min (as early experiments showed that even with an extension time of 6 min, the maximum SGSP-PCR amplicon length obtained was 3.5 kb) and the annealing temperature set at 1°C below the T_m of the primer with the lowest T_m of the pair. M13 F was first used as the vector specific primer. However, under these conditions, multiple products were obtained (see

Figure 4.3), even when the T_m of both the vector and insert primer were within 1°C of each other, making it impossible to tell which band was the true SGSP-PCR amplicon.



Lane number:

1. GeneRuler™ 1 kb ladder
2. K12 MG1655/*Hind*III genomic library (positive control)
3. S104
4. S117

Figure 4.3. *serU* U# SGSP-PCR, showing multiple amplicons.

The second lane is an SGSP-PCR with the positive control library, which from the *in silico* study was expected to produce an SGSP-PCR amplicon of 0.5 kb; this was successful, however there was also a faint non-specific band of 3.0 kb. The third and fourth lanes are *Shigella* strains, their respective specific SGSP-PCR amplicons (if any), could not be selected due to the number of non-specific PCR products formed.

This result was due to one or both of the primers binding non-specifically to the template in many locations, leading to the generation of multiple amplicons. As SGSP-PCR requires specific binding of both the insert- and vector-associated primers, a more stringent PCR

regime was required to obtain a single amplicon. To achieve this, three approaches were employed.

4.3.1 Touchdown PCR

A modification of the technique by (Don *et al.*, 1991) was used. The authors' method worked by starting the reaction with the annealing temperature 10°C above the usual annealing temperature for a standard PCR, then every second cycle the annealing temperature was decreased by 1°C until it reached the original annealing temperature, *i.e.* after twenty cycles. The PCR was then run for 10 more cycles at this temperature. This regime reduced mispriming as it ensured that the primers bind specifically in the early cycles of the reaction, producing specific amplicons which act as template for the future cycles run at lower stringency to increase the final product yield. I modified this technique slightly and as a rule for each SGSP-PCR, set the initial annealing temperature at 10°C above the T_m of the primer with the lowest T_m of the pair. Then every cycle this was decreased by 1°C, until after 10 cycles, the T_m of the primer was reached and the reaction was run for 20 more cycles at this temperature. Therefore, there were still thirty PCR cycles, but the touchdown step was 'steeper', this helped to maximise the amount of product without losing any of the specificity. The extension time was also reduced from 3 min to 2.5 min as the efficiency of the *Taq* polymerase was found to be higher than initially thought, so there was no need to have such lengthy extension times. See 2.8.2 for the detailed protocol.

4.3.2 Optimal vector primer choice

As pBluescript KS II (+) has 6 primers flanking the multiple cloning site which could be used for SGSP-PCR, it was important to choose the primer which anneals the least to the genome of *E. coli*, so that the chances of mispriming were minimised further. Each primer was checked against the chromosome of the control strain K12 MG1655 using Artemis, to see

how many times it annealed from the 3' end, from 3 oligomers (mers) all the way to the entire length of the primer. The results of the screen are shown in Table 4.2.

Table 4.2. pBluescript KS II (+) primer matches across the K12 MG1655 chromosome.

Number of hits to the K12 MG1655 chromosome							
bp from 3' end	3	6	9	12	15	18	entire #
Vector #							
M13 F (17 mer)	>200	>200	40	2	1	N/A	1
M13 R (19 mer)	>200	>200	73	3	1	1	1
T7 (22 mer)	>200	>200	4	0	0	0	0
T3 (20 mer)	>200	>200	7	0	0	0	0
SK (20 mer)	>200	>200	8	0	0	0	0
KS (17 mer)	>200	>200	51	1	0	N/A	0

From this, it was clear that both of the M13 primers hit to the K12 MG1655 chromosome once across their entire length, plus they have many partial hits; so these would be bad choices as vector primers for SGSP-PCR as they could produce many non-specific amplicons. T7 had the least hits to K12 MG1655 and was therefore selected as the vector primer for SGSP-PCR. This was convenient as corresponding SGSP-PCR amplicons could then be sequenced from SK which is positioned internal to T7 on the multiple cloning site (MCS) and is closer to the putative island DNA; therefore providing the maximum amount of 'island' sequence data from any sequence run. Also using a heminested sequence approach was likely to be more robust than sequencing from the primer used to generate the PCR amplicon (K. Rajakumar personal communication).

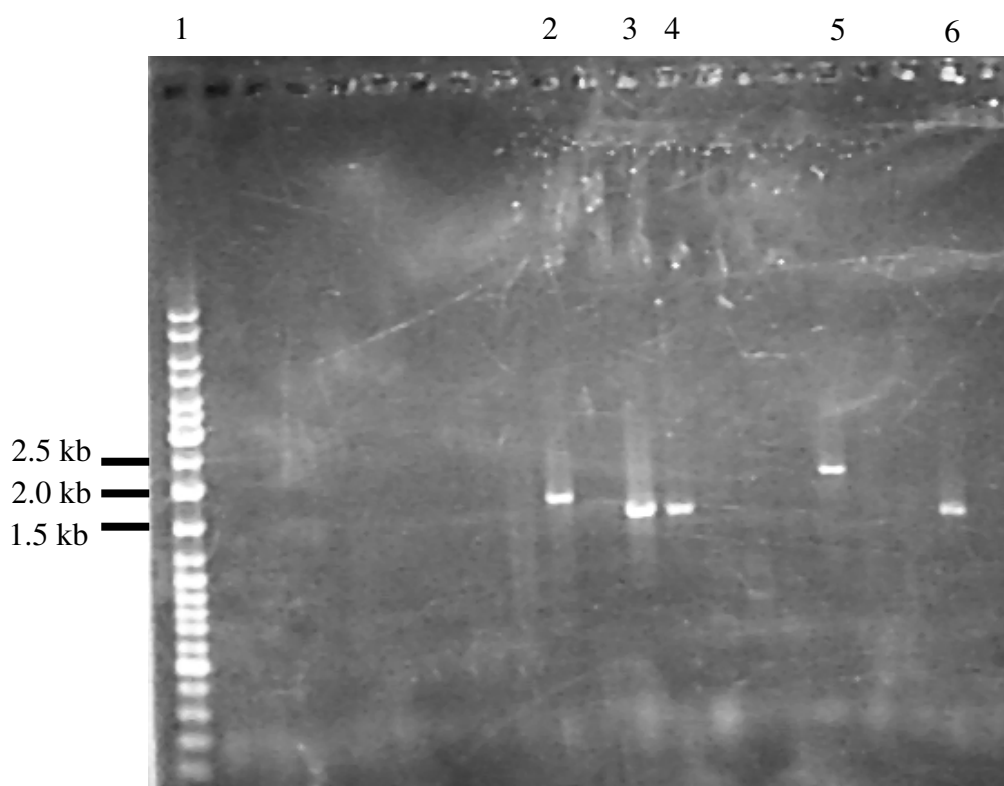
4.3.3 Use of hot-start *Taq* polymerase

Thermo-Start® DNA Polymerase, (ABgene) was used because this *Taq* is only active after an initial heating step of 95°C for 15 min, therefore eradicating the formation of non-specific products during the stage from the bench to the thermocycler. Reactions could then be prepared at room temperature, which was helpful when performing large numbers of PCRs.

In addition, the 10 x reaction buffer supplied with the hot-start *Taq* is a high performance buffer that gives a better yield of product with ‘problematic’ templates. Island DNA is known to be difficult to amplify due to the presence of repeats, and GC profile irregularities. Also specifically with SGSP-PCR, in many cases there may be a sudden increase or drop in the GC profile of the template DNA when the *Taq* moves from the core flanking DNA to island DNA; as a feature of island DNA is that its GC content is often different from the core DNA due to it being horizontally acquired. The optimised buffer was therefore a useful addition.

4.4 SGSP-PCR after optimisation

Figure 4.4 shows a typical SGSP-PCR using T7 as the vector primer, hot-start *Taq* polymerase and the touchdown PCR regime as described in sections 4.3.1 to 4.3.3.



Lane number:

1. GeneRuler™ 1 kb ladder
2. S103-*Pst*I library generated amplicon
3. S107-*Hind*III library generated amplicon
4. S111-*Hind*III library generated amplicon
5. S113-*Hind*III library generated amplicon
6. S117-*Eco*RI library generated amplicon

Figure 4.4. An optimised *leuX* U#-T7# SGSP-PCR indicating the generation of specific amplicons.

4.5 Band-stab PCR

Some SGSP-PCRs yielded only small amounts of product and these were reamplified in order to send enough for sequencing. In these cases, a modification of the band-stab technique (Bjourson and Cooper, 1992) was used and is described in section 2.8.3. In the second round

PCR the heminested vector primer (SK) was used, rather than T7, with the original insert specific primer in order to maximise the product yield.

4.6 Negative SGSP-PCRs

At strain-tRNA loci where SGSP-PCR from both U and D primers yielded no results from all five of the restriction enzyme libraries, three approaches were used to help characterise these problematic sites.

4.6.1 Additional libraries

Two more restriction endonucleases were used to generate genomic libraries for SGSP-PCR. *EcoRV* and *HincII* were chosen as they cut the genomes of *E. coli* and *Shigella* more frequently (in K12 MG1655 there are 2040 and 4071 cut sites respectively) than the original five enzymes. I hypothesised that SGSP-PCR with these libraries may therefore increase the chance of generating an SGSP-PCR amplicon; however, the respective amplicon may be shorter than with the initial 5 libraries and may provide less data on the associated putative GI.

4.6.2 *Int*-PCR

This approach was used at the *serW* tRNA locus; see section A2.6.1 for the details.

4.6.3 Southern hybridisation

This approach was used at the *aspV* tRNA locus; see section 7.9.1 for the details.

4.7 Sequencing and analysis

See section 2.12 for details on the preparation of DNA for sequencing and details of the sequencing companies used.

4.7.1 tRIP PCR amplicon sequencing

Representative tRIP amplicons of unexpected size were sequenced from either the U or D primer (see Table A2. 5). Those too large to sequence were further characterised by SGSP-PCR followed by sequencing and analysis.

4.7.2 SGSP-PCR amplicon sequencing

Representative SGSP-PCR amplicons were sequenced firstly from the vector primer so the sequence data acquired was from within the putative GI. Usually this sequence data was sufficient to verify that the amplicon was ‘specific’ i.e. generated from the U#/D# annealing specifically to the UF/DF region on the chromosome of the *Shigella* strain, this was because the DNA sequence obtained was in most cases known to be associated with the corresponding tRNA UF/DF in previously sequenced strains. If there was any doubt, due to the sequence being completely novel or mosaic, the amplicon was sequenced from the U#/D# to check that the island DNA was truly specific to the UF/DF (Figure 4.5).

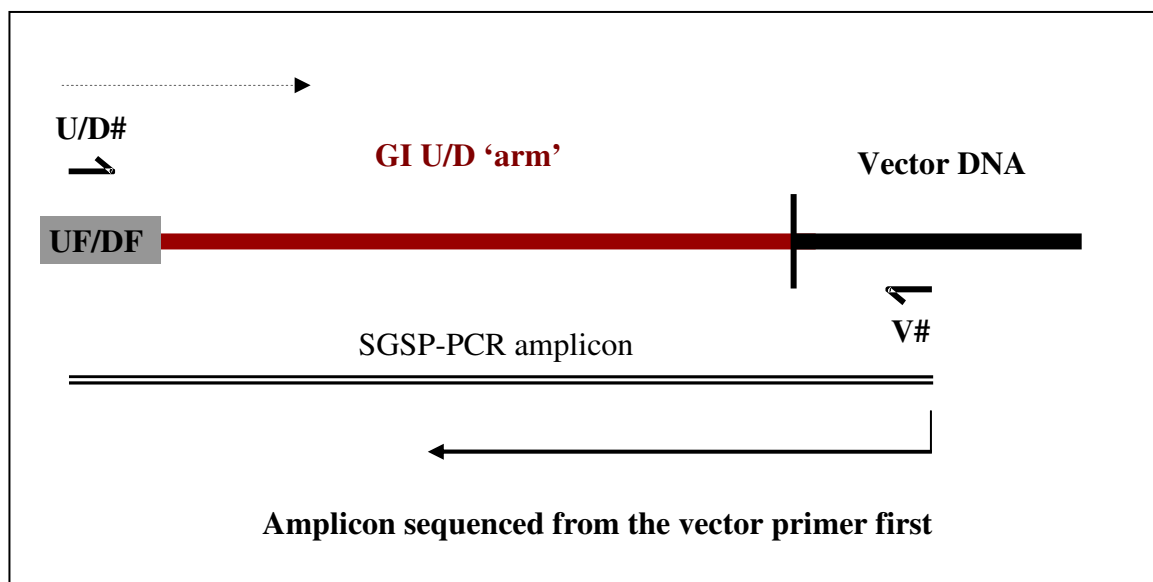


Figure 4.5. The system for sequencing of SGSP-PCR amplicons.

If required, the amplicon was sequenced from the insert specific U#/D# primer to confirm its association with the UF/DF.

4.7.3 Sequence runs

Sequence files were retrieved from the corresponding sequencing company websites (see 2.12), the length of the quality nucleotide sequence was recorded and the sequence data saved. The quality (clipped) sequences were saved in fasta format as EditSeq files (DNASTar™ software) for future analysis and manipulation. All of the sequences used in this study are stored in EditSeq format on the CD enclosed on the inside back cover of the thesis (see Addendum 1 also).

Both of the sequencing companies use *phred* (Ewing and Green, 1998) to check the quality of their sequence runs. *Phred* is a base-calling program that analyses the dye peaks on the raw DNA sequence chromatogram and ‘calls’ each of the peaks (either A, T, C or G) and assigns a quality score to each of the base calls, which represents the probability that the base call is wrong, see Table 4.3 for *phred* scores.

Table 4.3. *Phred* scores and the corresponding base call error probabilities.

<i>Phred</i> score	Probability that the base call is wrong
10	1 in 10
20	1 in 100
30	1 in 1000
40	1 in 10000

The higher the score the higher the quality of the base call.

The sequencing companies use a *phred* score of 20 or more to select the quality bases, known as the PHRED 20 standard, i.e. bases that have at least a 99% probability of being correct. This standard is commonly used to assess the quality of a sequence run. From the start of the sequence run, once the bases have a *phred* score of 20 or more, this is classed as the quality or ‘clipped’ sequence. Once the sequences’ *phred* quality drops below 20 for a significant

number of bases, the clipped sequence ends. With the sequencing services used in this study, the clipped sequence acquired from an individual sequence run was a maximum of 1000 bp. If the sequence run was of low quality, then no clipped sequence would be obtained and the reaction was classed as 'failed'. The entire sequence run is known as the 'unclipped' sequence. Therefore in most cases the clipped sequence was used in the sequence analysis, however if the sequence run failed, the unclipped sequence sometimes provided valuable information.

4.8 Blast searches using the NCBI database

Nucleotide sequences were analysed using Blastn (<http://www.ncbi.nlm.nih.gov/BLAST/>), this program aligns the input 'query' sequence with known nucleotide sequences that are available in the database, so it is a nucleotide-nucleotide match. It does this by breaking the query sequence down into smaller pieces called 'words', finds exact matches to these words and then extends the words to create an alignment; the program also creates gaps to allow for mismatches in the sequence. The default word size is 11, and was used most of the time, however it can be reduced to a minimum word size of 7, to make the search more accurate, this is especially useful if the query sequence is very short (7-20 bases), and was used to check primer sequences.

Blastn was usually sufficient to provide detailed information on the nature and location of the query sequences relative to the known microbial sequences available on the NCBI database; however, sometimes the query sequence or a part of the query sequence had no significant matches to the database. In these cases the sequence was analysed using Blastx and tBlastx; both programs translate the query nucleotide sequence in all six reading frames and check it against a protein database or translated nucleotide database respectively, therefore accounting for any errors or differences in the query sequence that may affect the reading frame of a potential protein-coding region. These programs then find proteins that are homologous to

the translated nucleotide query sequence and are very useful when analysing potentially novel sequences as they help assign a putative function to the nucleotide sequence.

5.0 Overall results of the characterisation of island DNA across *Shigella*

5.1 Matrix of results across the *Shigella* strains

Table 5.1 shows the overall island characterisation across nineteen *Shigella* strains at the sixteen tRNA loci that are known to be hotspots for GI insertion in *E. coli* and *Shigella*. tRIP-positive strain-tRNA-loci are represented by boxes containing ‘+’ symbols. Each tRIP-negative strain-tRNA locus is represented by a 3-tier box, showing the associated GIs U-arm classification, D-arm classification (meaning the island sequence associated with the upstream and downstream flanking regions respectively) and its overall ‘island family’ designation.


Each island family (IF) is represented by a number, the designation of GIs into families was to help indicate the presence of similar islands from strain to strain at each tRNA locus. These families are described in more detail in each of the separate tRNA locus results sections. ‘U’ indicates that the GI family is unclassifiable due to the nature of the sequence obtained, ‘–’ indicates that the putative GI present at this locus has not been characterised yet.


GI U- and D-arm assignments were based on the most significant Blastn hit across the full length of the sequence obtained. Split assignments were given if the Blastn hit was the same across more than one sequence.

Table 5.1. Overall results of the characterisation of island DNA in nineteen *Shigella* strains across 16 tRNA loci that are known hotspots for GI insertion.

Strain	Strain code	tRNA locus															
		<i>aspV</i>	<i>thrW</i>	<i>serW</i>	<i>serT</i>	<i>serX</i>	<i>serU</i>	<i>asnT</i>	<i>argW</i>	<i>metV</i>	<i>glyU</i>	<i>pheV</i>	<i>selC</i>	<i>pheU</i>	<i>leuX</i>	<i>ssrA</i>	<i>asnV</i>
<i>S. dysenteriae</i> 3	S101	+	-	SRL(1) ^a	+	1	1	1	5	1	-	1	+	1	1	1	+
<i>S. dysenteriae</i> 9	S102	+	2	+	+	1	1	2	+	1	1	1	+	1	1	1	+
<i>S. dysenteriae</i> 6	S103	+	-	+	+	1	-	1	U	1	1	-	+	1	1	1	+
<i>S. flexneri</i> 1a	S104	1	1	+	+	+	2	3	1	+	4 ^c	2	SHI-2 (1)	+	2	+	+
<i>S. flexneri</i> 1b	S105	1	1	+	+	+	2	3	1	+	2	2	SHI-2 (1)	+	+	+	+
<i>S. flexneri</i> 2a	S106	1	1	+	+	+	2	3	1	+	2	she (3)	SHI-2 (1)	+	2	+	+
<i>S. flexneri</i> 2b	S107	1	4	+	+	SRL (2) ^a	2	3	1	+	2	she (3) ^b	SHI-2 (1)	+	2	+	+
<i>S. sonnei</i>	S108	2	2	+	+	1	+	-	2	+	3	she (3) ^b	SHI-2 (1)	+	3	-	+
<i>S. flexneri</i> 6	S110	+	2	+	+	SRL (2) ^a	-	-	+	1	+	1	+	1	+	1	+
<i>S. flexneri</i> X	S111	1	U	+	+	+	2	3	1	+	+	2	SHI-2 (1)	+	2	+	+
<i>S. flexneri</i> Y	S112	1	1	+	+	+	2	3	1	+	+	-	SHI-2 (1)	+	+	+	+
<i>S. sonnei</i>	S113	-	3	+	+	1	+	-	2	+	3	she (3) ^b	SHI-2 (1)	+	3	2	+
<i>S. sonnei</i> bio a	S114	-	-	+	+	1	+	-	2	+	3	she (3) ^b	SHI-2 (1)	+	3	-	+
<i>S. sonnei</i> bio g	S115	2	2	+	+	1	+	-	2	+	3	she (3) ^b	SHI-2 (1)	+	3	-	+
<i>S. boydii</i> 1	S116	+	2	+	+	1	1	-	3	1	-	1	+	SHI-3 (1)	4	1	+
<i>S. boydii</i> 2	S117	+	2	+	+	1	1	-	+	1	-	-	+	SHI-3 (1)	4	1	+
<i>S. boydii</i> 3	S118	+	2	+	+	1	-	-	+	1	-	-	+	1	1	1	+
<i>S. boydii</i> 4	S119	+	2	+	+	1	1	-	+	1	-	1	+	1	4	1	+
<i>S. boydii</i> 7	S120	-	3	+	+	-	1	1	4	-	+	-	+	1	1	-	-

GI U-arm	GI U- and D-arm assignation	
Island family	<i>E. coli</i> K12 MG1655	
GI D-arm	<i>E. coli</i> UPEC CFT073	
	<i>E. coli</i> O157:H7 strains EDL933 & Sakai	
	<i>S. flexneri</i> 2a strains Sf301 & Sf2457T	
	<i>S. sonnei</i> 046 ^d	
	<i>S. boydii</i> 4 Sb227 ^e	
	Unclassifiable	
	Novel	
	Yet to be characterised	

 Amplicon sequence hits to GI(s) represented in the Islander database^f

 Integrase found but the GI is not represented in Islander^f

- ^a The GI sequence corresponds to the SRL PAI in *S. flexneri* 2a YSH6000 (GenBank accession number AF326777)
- ^b The strain was already characterised as harbouring *she* PAI-like elements at this locus by (Al-Hasani *et al.*, 2001a)
- ^c The U-arm associated sequence was found associated with another *gly* tRNA gene in *S. flexneri* 2a SF2457T only (see section A2.6.3)
- ^d GenBank accession no. NC_007384
- ^e GenBank accession no. NC_007613
- ^f See section 5.3 for more details.

5.2 Occupancy of tRNA loci across *Shigella*

Figure 5.1 shows the total number of tRIP-negative and tRIP-positive tRNA loci across the *Shigella* strains.

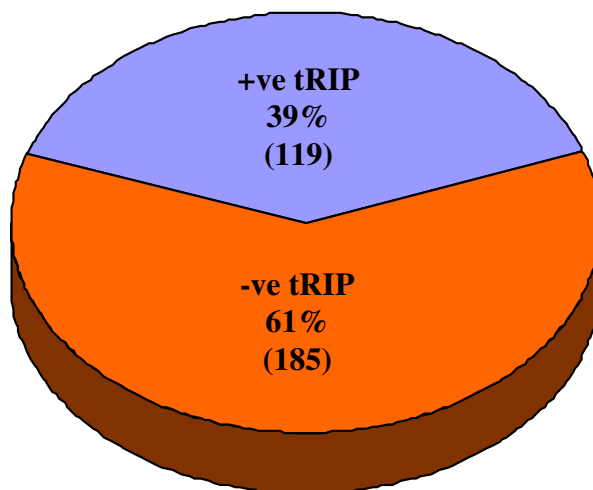


Figure 5.1. The proportion of empty and putatively GI occupied tRNA loci as indicated by the tRIP screen at 16 tRNA loci across nineteen *Shigella* strains.

Total numbers are shown in parentheses.

5.3 Breakdown of putatively GI occupied tRNA loci

Figure 5.2 shows the overall classification of each of the tRIP-negative strain-tRNA loci after characterisation of their putative GI U and/or D-arms.

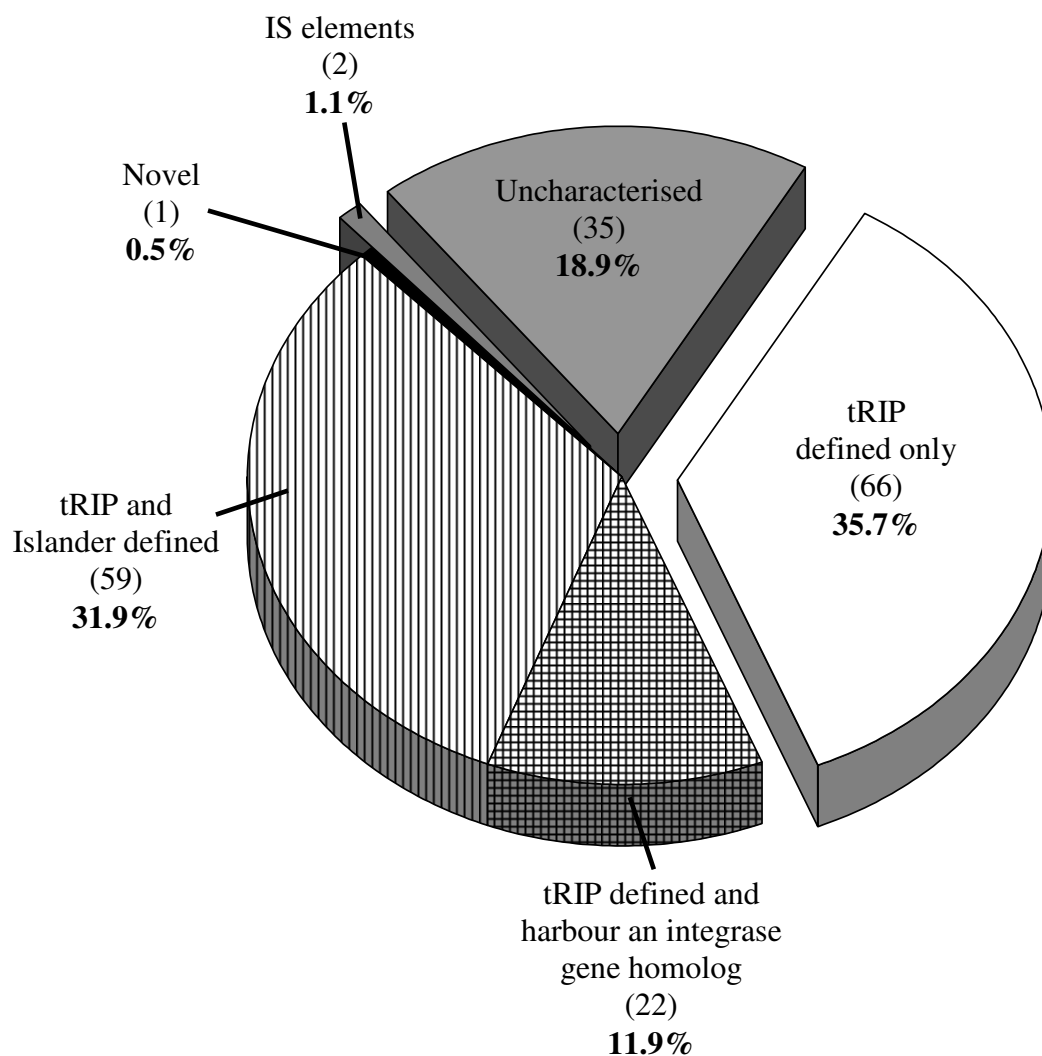


Figure 5.2. The proportions and final classification of each tRIP-negative strain-tRNA locus.

Total numbers are shown in parentheses, percentages are shown in bold.

Fifty nine (32%) of the elements identified in this study resembled GIs defined by Islander (<http://kementari.bioinformatics.vt.edu/cgi-bin/islander.cgi>), an algorithm that scans bacterial genomes to detect tRNA-borne archetypal integrative elements that are bounded by short direct repeats and that contain an intact integrase gene homologue (Mantri and Williams,

2004) (see section 1.7 also). A further twenty two identified GIs were found to contain an integrase gene, but these were similar to islands not represented in the Islander database.

Only one novel GI was detected, this was found in S116 (*S. boydii* 1 strain) at the *argW* locus, its characterisation is described in section 6.2.

At least eighty one (54%) of the elements identified harbour phage-like integrase genes, indicating that many of these elements arose following acquisition of horizontally transferred integrative GIs.

5.4 Island families observed across *Shigella*

A total of thirty five distinct elements (island families) were observed across the *Shigella* strains tested in this study, Table 5.2 shows the site by site breakdown.

Fifteen are Islander defined elements and twenty are defined by tRIP only, this indicates that many of the GIs found across *Shigella* may be more ancient, mosaic elements that are likely to be locked into the chromosome. However, certain sites stand out as hotspots for prophage activity and diversity; these are *thrW*, *argW* and *leuX*.

Table 5.2. Island families found across nineteen *Shigella* strains at sixteen tRNA loci that are known to be hotspots for GI insertion in *E. coli*.

tRNA locus	Islander defined GI families	tRIP defined GI families
<i>aspV</i>	0	2
<i>thrW</i>	4	0
<i>serW</i>	1	0
<i>serT</i>	0	0
<i>serX</i>	1	1
<i>serU</i>	2	0
<i>asnT</i>	0	3
<i>argW</i>	4	1
<i>metV</i>	0	1
<i>glyU</i>	0	4
<i>pheV</i>	1	2
<i>selC</i>	0	1
<i>pheU</i>	1	0
<i>leuX</i>	1	3
<i>ssrA</i>	0	2
<i>asnV</i>	0	0
TOTAL	15	20

5.5 Summary of island characterisation across *Shigella*

Table 5.3 shows the breakdown of the tRIP and island characterisation results across the sixteen tRNA loci probed in this study.

Table 5.3. Site by site breakdown of the tRIP screen results and island characterisation across nineteen *Shigella* strains at sixteen tRNA loci that are known to be hotspots for GI insertion in *E. coli*.

tRNA gene	No. tested by tRIP	No. tRIP-negative	Percentage occupancy	No. bearing GIs confirmed by sequencing	No. bearing GIs confirmed by RP ^a	No. of strain-tRNA loci characterised	No. of strain-tRNA loci awaiting characterisation
<i>aspV</i>	19	11	58	5	3	8	3
<i>thrW</i>	19	19	100	10	6	16	3
<i>serW</i>	19	1	5	1	0	1	0
<i>serT</i>	19	0	0	0	0	0	0
<i>serX</i>	19	14	74	5	8	13	1
<i>serU</i>	19	15	79	6	6	12	3
<i>asnT</i>	19	19	100	5	5	10	9
<i>argW</i>	19	14	74	8	6	14	0
<i>metV</i>	19	9	47	3	5	8	1
<i>glyU</i>	19	15	79	5	5	10	5
<i>pheV</i>	19	19	100	8	6	14	5
<i>selC</i>	19	10	53	1	9	10	0
<i>pheU</i>	19	9	47	3	6	9	0
<i>leuX</i>	19	16	84	9	7	16	0
<i>ssrA</i>	19	13	68	5	4	9	4
<i>asnV</i>	19	1	5	0	0	0	1
Total	304	185	N/A	74	76	150	35

^a The SGSP-PCR amplicon(s) generated indicated that this strain had the same restriction pattern (RP) in the GI U/D-arm to a sequenced representative, and therefore was very likely to harbour the same sequence.

5.6 S109 (*S. flexneri* 4a) re-classification to an *E. coli*

The tRIP screen results show that S109 (*S. flexneri* 4a strain) had the same tRIP profile as *E. coli* K12 MG1655 and the same size tRIP amplicons (see Table 3.2 and Table 3.3). This indicated that in terms of S109's tRNA occupancy, it was more like *E. coli* than *Shigella*, straight away flagging this strain as a potential outlier to the rest of the strains. SGSP-PCR at the putatively GI occupied tRNA loci *glyU*, *argW*, *asnT*, *ssrA*, and *thrW* and subsequent sequence and/or RP analysis of the corresponding amplicons indicated that all of the S109 associated island DNA corresponded to K12 MG1655. Therefore, to determine its core biochemical phenotypes, S109 was plated on Xylose-Lysine-Desoxycholate (XLD) agar (Oxoid), a selective and differential medium used to isolate and identify Gram-negative enteric pathogens. It contains xylose, lysine and sodium thiosulfate and is mainly used to differentiate between *Shigella* and *Salmonella* and non-pathogens. The presence of the indicator phenol red resolves the xylose reaction of a respective organism. The pH of XLD agar is around 7.4, so to start with the agar appears red. Most enteric bacteria including *E. coli* ferment xylose, this produces acid, which reduces the pH of the medium to below 7.0, causing the phenol red in the medium to go yellow. However, *Shigella*, *Providencia* and *Edwardsiella* do not ferment xylose, so the medium stays red. *Salmonella* ferment xylose, but also decarboxylate lysine, therefore making the medium alkaline, so the medium stays red like the *Shigella* reaction. Therefore, most *Salmonella* and *Edwardsiella* are differentiated from *Shigella* by the presence of thiosulfate in the medium as they metabolise this to produce hydrogen sulphide, which causes the colonies to form black centres due to the presence of a hydrogen sulphide indicator in the medium.

When streaked to single colonies on XLD agar, after 24 hr growth at 37°C, the medium surrounding the S109 colonies went yellow, but to less of an extent as the *E. coli* control strain K12MG1655, indicating that S109 was phenotypically an *E. coli*, but exhibited slow xylose fermentation. Therefore, S109 was then further characterised using an API® 20 E test strip (see 2.17 for methodology). This identified S109 as being 81.5% the same as *E. coli*,

indicating that it is an 'atypical' *E. coli*. The only difference in its API® 20 E profile to a typical *E. coli* was that the ONPG reaction was 'slow', being only 90% positive after 24 hr, so it was designated as negative on the scoring sheet (Typical *E. coli* give a strong ONPG positive reaction). ONPG (Ortho-nitrophenyl- β -galactoside) is an artificial substrate used to measure the activity of the enzyme β -galactosidase, which all *E. coli* produce, whereas most *Shigella* apart from *S. dysenteriae* 1 and *S. sonnei* do not produce β -galactosidase (Ito *et al.*, 1991). This result suggested that S109 is either deficient in the production of β -galactosidase, or the activity of the enzyme is attenuated. Overall, both the core biochemical properties and the GI profile of S109 provided strong evidence to show that it is more *E. coli*-like than *Shigella*-like and it was therefore excluded from the study.

5.7 S108 (*S. flexneri* 3a) re-classification to a *S. sonnei*

The tRIP screen results show that S108 (*S. flexneri* 3a strain) had the same tRIP profile as the *S. sonnei* strains screened (see Table 3.2 and Table 3.3). Subsequent SGSP-PCR at the tRIP-negative tRNA loci also showed that S108's associated island DNA was *S. sonnei*-like (see Table 5.1). When tested using a slide agglutination test against specific antisera to each of the four *Shigella* species, S108 exhibited cross-reactivity. Therefore I tested S108 with an API® 20 E strip; it was found to be 98.3% the same as *S. sonnei*, with the next closest hit being 1.5% to *E. coli*. The rhamnose reaction was 'slow' being only 75% positive after 24 hr and the ONPG reaction was positive at 24 hr, this indicated that S108 was likely to be *S. sonnei* biotype g (Mehrabian and Tohidpour, 2005). Based on these collective results, S108 was therefore re-classified as a *S. sonnei* strain.

6.0 Island characterisation at the *argW* locus

A novel GI was discovered at the *argW* locus; see sections 6.2 to 6.4 for the details of the initial characterisation and subsequent probing of the element. The overall results of the characterisation of the *argW* associated island DNA across the other *Shigella* strains are also described in section 6.1.

6.1 *argW* overall results

Table 6.1. SGSP-PCR results of the *argW* tRIP negative strain-tRNA loci

<i>argW</i> U# - T7# SGSP-PCR results									
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)							Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	<i>Eco</i> RV	<i>Hinc</i> II	
K12 MG1655				~1.8					
<i>S. dysenteriae</i> 3	S101	~0.9 F, ~1.4 F	~0.8 F, ~1.3 F	~1.8 F	~0.4, 0.7 F	N	~ 0.8	~0.6	679 [SK#]
<i>S. dysenteriae</i> 6	S103	~0.9 F, ~1.4 F	~0.8 F, ~1.3 F	N ^a	~0.4	N			
<i>S. flexneri</i> 1a	S104	~0.9 F, ~1.4 F	~0.8 F, ~1.3 F ^c	~ 1.2 ^b	~0.4	N			750 [SK#]
<i>S. flexneri</i> 1b	S105			~1.2					
<i>S. flexneri</i> 2a	S106			~1.2					
<i>S. flexneri</i> 2b	S107	N	~0.8 F, ~1.3 F	~1.2	~0.4	N			
<i>S. sonnei</i>	S108	~0.9 F, ~1.4 F	~0.85 F, ~1.6	N	N	~3.5 F			
<i>S. flexneri</i> X	S111	N	~0.8 F	~ 1.2	~0.4	N			750 [SK#]
<i>S. flexneri</i> Y	S112			~1.2					
<i>S. sonnei</i>	S113			~ 2.3					750 [SK#]
<i>S. sonnei</i> bio a	S114	~0.9 F, ~1.4 F	~0.85 F, ~ 1.6	N	N	~3.5 F			2670 UC ^d [U#]
<i>S. sonnei</i> bio g	S115	N	~0.85 F, ~1.6	N	~0.4 F	N			
<i>S. boydii</i> 1	S116	N	N	N	~0.4	N		~ 1.4	556 [U#], 808 [SK#]
<i>S. boydii</i> 7	S120	N	~0.85 F, ~1.6	~ 0.5	~0.4	N			378 [U#], 353 [SK#]

<i>argW</i> D# - T7# SGSP-PCR results							
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)					Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sa</i> I	
K12 MG1655				~1.6			
<i>S. dysenteriae</i> 3	S101	~ 1.4	N	N	~ 2.0	N	1.4 kb 640 [SK#], 2.0 kb 611 [SK#] NS
<i>S. dysenteriae</i> 6	S103	~1.4	N	N	~ 2.0	~1.9	524 [SK#] NS
<i>S. flexneri</i> 1a	S104						
<i>S. flexneri</i> 1b	S105						
<i>S. flexneri</i> 2a	S106						
<i>S. flexneri</i> 2b	S107						
<i>S. sonnei</i>	S108	~0.5 F, ~ 2.1 F	N	~2.8 F	N	~1.2 F	451 [SK#]
<i>S. flexneri</i> X	S111						
<i>S. flexneri</i> Y	S112						
<i>S. sonnei</i>	S113						
<i>S. sonnei</i> bio a	S114	N	~0.9 F	~0.5 F, ~ 2.1 , ~2.8F	N	~1.2 F	674 [SK#]
<i>S. sonnei</i> bio g	S115	N	N	~ 2.8	N	N	1117 UC [SK#]
<i>S. boydii</i> 1	S116	N	~ 1.6	~1	~1.6	N	736 [SK#]
<i>S. boydii</i> 7	S120	~0.5	N	~0.6 F, ~1.2 F	~ 0.9	N	822 [SK#]

^a Indicates that no amplicon was generated

^b Text highlighted in bold indicates that the amplicon was sequenced

^c The addition of 'F' after the text indicates that the amplicon was faint

^d The addition of 'UC' after the text indicates that the unclipped sequence was used in the analysis. The sequence run failed by the Phred20 (Ewing and Green, 1998) standards used by the sequencing company; however the low quality sequence still provided some meaningful information

6.1.1 *S. flexneri*

The *argW* U# results show that the typical *S. flexneri* strains all have the same sequence as is present in the U-arm of the Sf301 *argW* associated 5.6 kb islet (*argW*-IF1). The sequence and restriction profile results indicate the presence of a truncated *S. flexneri* Sf6-like prophage integrase gene which is disrupted by an IS911 (see Figure 6.1). The product of the integrase gene will therefore be non-functional and because of this, the entire element may be 'locked' into the genome of *S. flexneri*. The small size of the islet suggests that much of the original *argW* borne element has been deleted from the genome of Sf301, possibly due to pathoadaptation (Maurelli, 2007) and/or IS activity, so that the entity now present is the remnants of a previously inserted Sf6-like prophage. This notion is supported by the observation that in other *E. coli* and *Shigella* strains, larger, similar prophage-like islands, with a homologous integrase gene and containing phage genes are found associated with *argW* (see section 6.1.2 below and Figure 6.2).

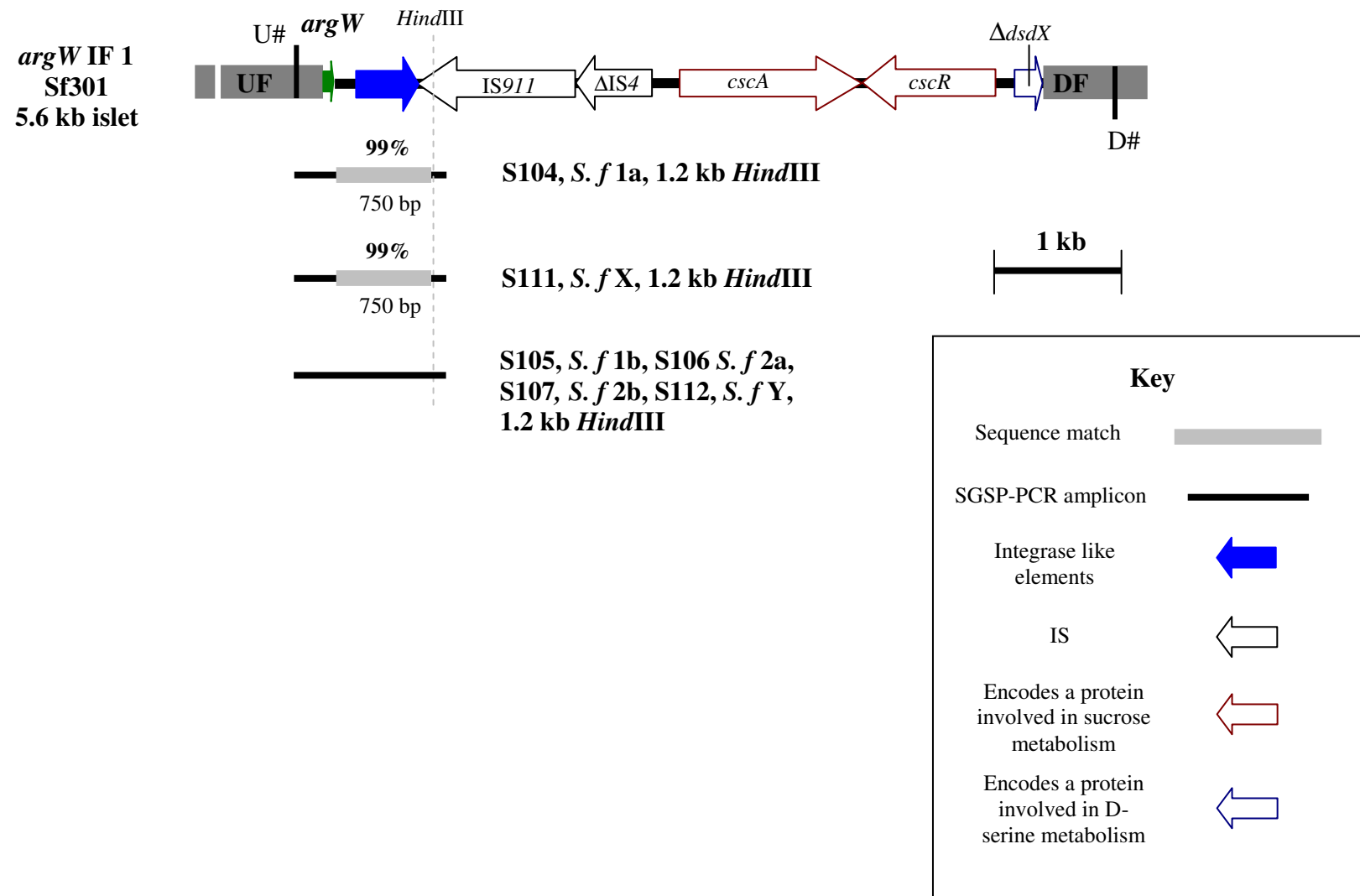


Figure 6.1. *argW* U-arm SGSP-PCR results for the *S. flexneri* strains

6.1.2 *S. sonnei*

The *argW* U and D-arm results show that three of the *S. sonnei* strains (S108, S114 and S115) harbour the same elements as are present in the corresponding U and D-arms of the *argW* 51.9 kb prophage-like GI (*argW*-IF2) in the genome of the fully sequenced *S. sonnei* strain Ss046. S113 (*S. sonnei* strain) however, was not characterised from the D# and has some variation in the U-arm, as there is an IS2 element present just downstream of the integrase which does not correspond to the Ss046 sequence (see Figure 6.2).

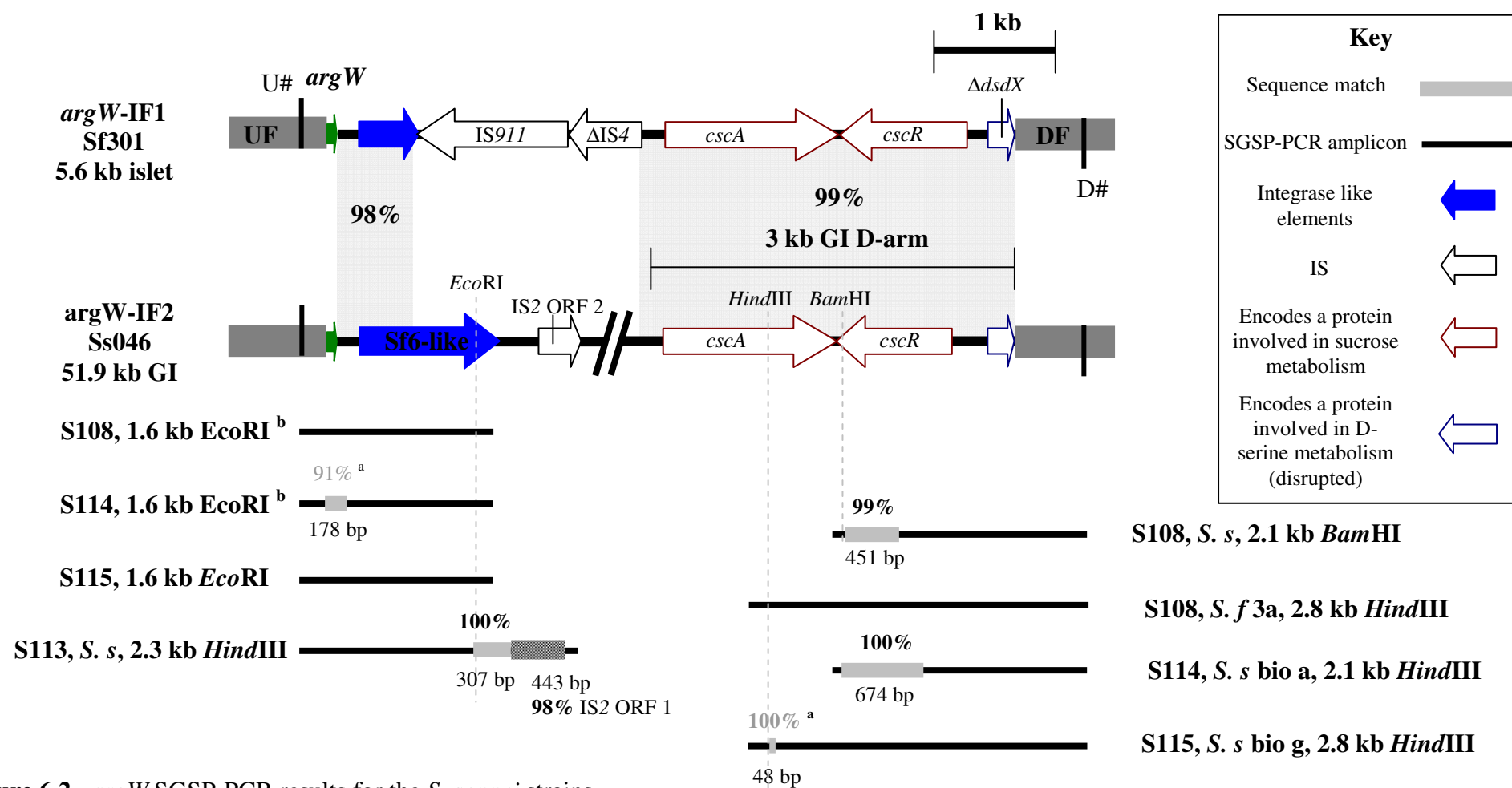


Figure 6.2. *argW* SGSP-PCR results for the *S. sonnei* strains

^a Sequence runs failed, however the unclipped (low quality) sequences hit to the regions indicated

^b S108 and S114 also yielded 3.5 kb *SalI* U# SGSP-PCR amplicons (not shown in this figure, see table 6.1), these also correspond to the *in silico* U# SGSP-PCR for Ss046.

After Blastn analysis, the *argW* Ss046 GI was found to harbour 17.5 kb of genes with 96% nucleotide identity to genes found on bacteriophage lambda and has a full length integrase gene which has 96% nucleotide identity to the *S. flexneri* serotype converting bacteriophage Sf6 (GenBank accession number AF547987) integrase gene. Sf6 is also a mosaic lambdoid-like phage that harbours O-antigen modification genes (Casjens *et al.*, 2004). The phage infects *S. flexneri* serotypes X and Y and after lysogenisation at the *thrW* tRNA locus, converts them into serotypes 3a and 3b respectively; Sf6 is interesting in that it is a lambdoid-like phage, but its integrase gene is homologous to the P4 phage integrase (Clark *et al.*, 1991). This is reflected by the orientation of the Sf6 integrase gene relative to the *argW* tRNA gene, Figure 6.2 shows how the Sf6-like integrase genes are P4-like as they are colinear with the *argW* tRNA gene. Previous studies have shown that lambdoid phages insert at the tRNA anticodon loop and the orientation of the integrase gene is opposite to that of the tRNA gene (one exception is the HK620 phage), whereas P4 phages insert at the T ψ C loop nearer the 3' terminus of the tRNA gene and the orientation of the integrase gene is in the same orientation as the tRNA gene (Campbell, 2003). In addition, the Sf6 integrase gene also has 96% nucleotide identity to the KpLE1 prophage integrase gene. The KpLE1 prophage is associated with the *argW* tRNA locus in *E. coli* and the K12 MG1655 KpLE1 element is present as an intact prophage, with a full length integrase gene and flanked by DRs and is therefore included in the Islander database. Downstream of the integrase gene, the KpLE1 prophage harbours O-antigen serotype conversion genes that have similarities to *Shigella* O-antigen modification genes suggesting that this prophage may be involved in O-antigen serotype conversion in *E. coli*. This observation has also been reported previously in a study on the *S. flexneri* SfV serotype converting bacteriophage (Allison *et al.*, 2002) and further emphasises the authors proposal of the possible coevolution of O-antigen modification genes and serotype converting phages in *E. coli* and *Shigella*.

The Ss046 *argW* GI therefore has features of a mosaic lambdoid-like prophage and has an integrase gene homologous to those found present on the above mentioned serotype converting prophages; however it harbours no serotype conversion genes. Interestingly *S. sonnei* harbours no chromosomal O-antigen synthesis genes either (unlike *E. coli* and the other *Shigella* species). The genes that encode the only *S. sonnei* serotype (form I O polysaccharide) are encoded on the large virulence plasmid and are believed to have been acquired by HGT from *Plesiomonas shigelloides*; as these genes conferred a strong selective advantage in the adaptation to being a human pathogen, most of the original chromosomal O-antigen synthesis genes were then later deleted, as there are remnants of these genes present on the *S. sonnei* chromosome (Lai *et al.*, 1998). On further analysis of the Ss046 *argW* associated GI, it is likely that O-serotype conversion genes were at some time harboured on the island but have also since been deleted, possibly due to IS activity in this region. These results suggest that IS may play an important role in negative selection, by inactivating/deleting horizontally acquired genes that are/become redundant to the host organism.

6.1.3 The *argW* associated sucrose metabolism and D-serine metabolism genes

The *argW* island DNA present in the D-arms of the *S. sonnei* strains shown in Figure 6.2 harbour two sucrose metabolism genes *cscA* and *cscR* that encode a sucrose hydrolase and a sucrose operon repressor respectively; they are also present in Sf301 and EDL933 (and others, see Table A2. 6). These genes form an operon and make up half of a regulon (*cscRAKB*) that comprises two more genes that should be present upstream of *cscA*, however these are deleted in Sf301 and displaced by an IS element in Ss046. The regulon will therefore in these strains be non-functional. The sucrose metabolism regulon has been studied previously in *E. coli* and it confers the ability to use sucrose as a carbon source; it is present in less than 50% of strains and is believed to have been acquired by HGT around the time when different enterics

diverged (Jahreis *et al.*, 2002), which would explain their presence in some *Shigella* strains. Their location, downstream of the prophage-like elements associated with *argW*, also indicates that they were acquired at an earlier time, and they therefore constitute a distal flanking islet that increases the fitness of the host organism. More recent studies have shown that the *argW* associated sucrose metabolism regulon is present in 95% of diarrhoeal pathogens and it substitutes the D-serine metabolism genes (the *dsdCXA* locus), disrupting the *dsdX* gene (see Figure 6.2). Conversely in CFT073, K12 MG1655 and most extraintestinal pathogenic *E. coli* (ExPEC) strains, the *dsdCXA* locus is intact and they are able to use D-serine as the sole carbon source, but unable to use sucrose (Moritz and Welch, 2006). The authors speculate that there is a strong selective pressure for intestinal pathogens to harbour the sucrose utilisation genes because of the sucrose rich environment of human and animal intestines, whereas the D-serine metabolism genes would be more useful to pathogens that infect a broader range of Extraintestinal sites. It is possible that loss/inactivation of some of the sucrose metabolism genes in the above *Shigella* strains has occurred due to pathoadaptation, as *Shigella* evolved to become a more specialised human pathogen.

6.1.4 The S120 *argW* D-arm is CFT073-like

The S120 (*S. boydii* 7 strain) *argW* SGSP-PCR results were interesting as the D-amplicon sequence and restriction fragment data indicated the presence of island DNA that is involved in D-serine metabolism, as is present in the D-arm of CFT073 (an intact *dsdX* gene, see Figure 6.3). As mentioned above in section 6.1.3 an intact version of this gene is also found present in K12 MG1655 and in many ExPEC *E. coli* strains, but not in EDL933 nor any of the *Shigella* genomes available on the NCBI, as in the latter strains the *dsdX* gene is disrupted and replaced with sucrose utilization genes (the *csc* operon).

The D-arm Sequence had 99% ID to the corresponding region in K12 MG1655 and 94% to CFT073, but the RPLP is the same as CFT073 not K12 MG1655, so for this reason the D-arm

was classified as being the most similar to CFT073. The S120 *argW* GI was therefore assigned to a distinct island family: *argW*-IF4.

The U# generated amplicon walks 0.1 kb into the U-arm of the corresponding prophage-like sequence that is found associated with *argW* in *E. coli* and *Shigella*, but as the island sequence data obtained was so short, it was not possible to assign it to a specific family, for this reason the U-arm was designated as ‘unclassified’ (see Table 5.1).

These results highlight the diversity of *S. boydii*, indicating that some strains the associated island DNA is sometimes ExPEC-like rather than *Shigella*-like and further confirms that *Shigella* has multiple independent origins.

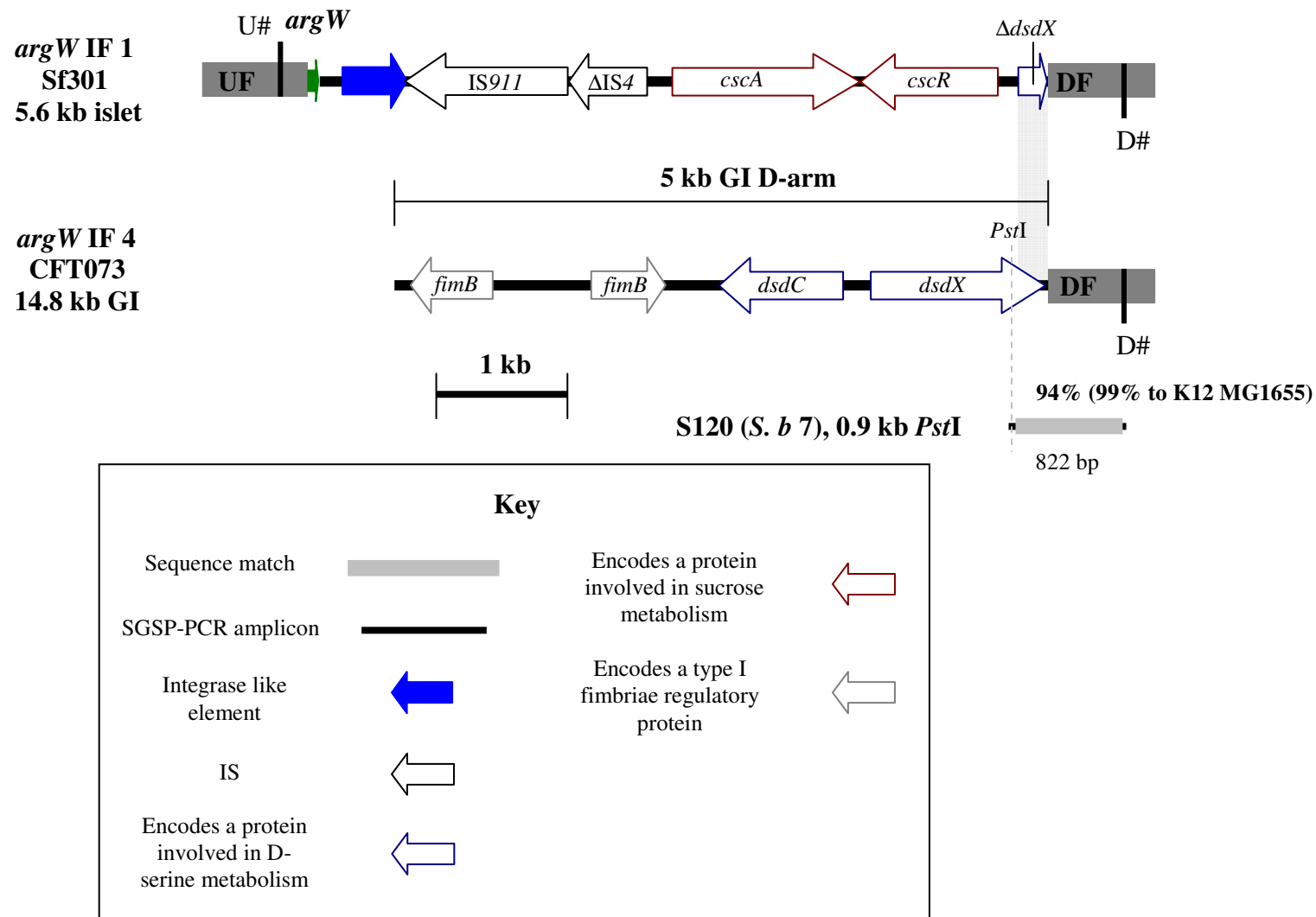


Figure 6.3. *argW* D-arm SGSP-PCR results for S120 (*S. boydii* 7 strain)

6.1.5 *S. dysenteriae*

The S101 (*S. dysenteriae* 3) *argW* GI has been assigned to a unique island family, *argW*-IF5, after sequencing analysis of its SGSP-PCR amplicons. The U# amplicon walks 160 bp into the U-arm of the corresponding Islander defined prophage-like sequences that are found associated with *argW* in *E. coli* and *Shigella* (the highest nucleotide identity being 99% to the UF and island DNA in the 1.0 kb *argW* Sb227 islet), but then this is disrupted by some sequence which has the highest nucleotide identity (98%) to the exact start of a 2.5 kb ISSbo6 (see Figure 6.4); a recently identified IS element which is found seven times in the chromosome of Sb227, and only once partially in the chromosome of Ss046 and not at all in any of the other *Shigella* genomes. ISSbo6 is similar to ISEc8 (a member of the IS66 family that is found next to the LEE PAI in EHEC) and is interestingly found present mostly in the SHI-1, SHI-2 and *ipaH* islands in Sb227 (Yang *et al.*, 2005).

The D# results indicate the presence of mosaic IS adjacent to the DF (see Figure 6.4), comprising a partial IS911 and the start of an ISEc8, which is inverted with respect to the ISSbo6 mentioned above.

This suggests that in this GI in S101, the U-arm associated ISSbo6 and D-arm associated ISEc8 elements have played/are playing a role in the rearrangement of the island DNA that lies/lay between them.

The S103 (*S. dysenteriae* 6) *argW* GI has been designated as ‘unclassifiable’ because no U-arm sequence was generated and it produced the same size D# SGSP-PCR amplicon as S101 with the *Bam*HI library (see Table 6.1), which was found to walk into IS elements only.

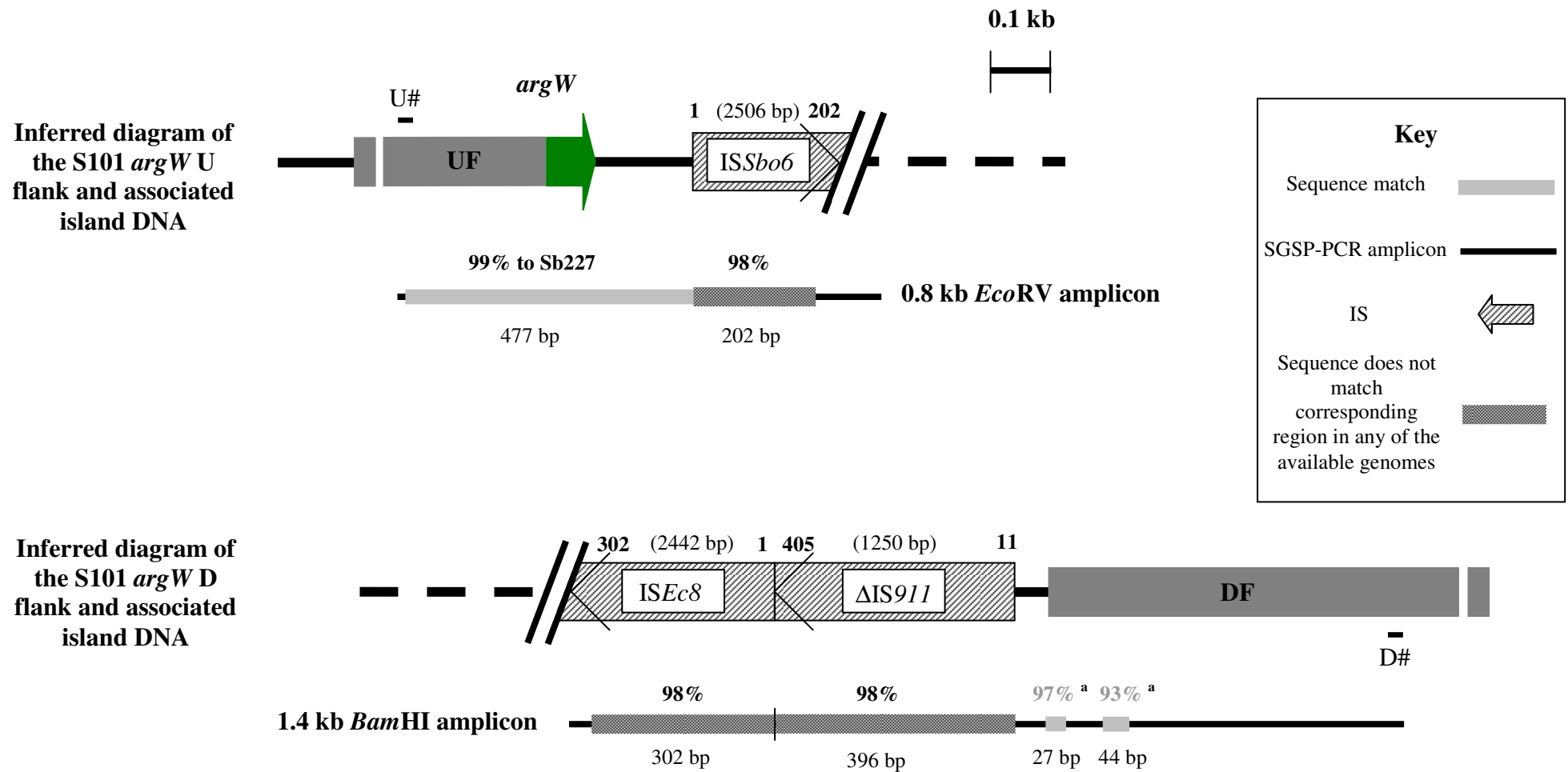


Figure 6.4. *argW* SGSP-PCR results for S101 (*S. dysenteriae* 3 strain)

^a Sequence runs failed, however the unclipped (low quality) sequences hit to the regions indicated.

6.2 The S116 (*S. boydii* 1 strain) *argW* novel prophage-like GI

6.2.1 SGSP-PCR Results

The *argW* S116 GI was initially characterised from the D#, as SGSP-PCR from the U# yielded no amplicons that walked beyond the *argW* tRNA and into the putative island DNA with the original five restriction enzyme libraries (see Table 6.1). The D# sequence data obtained had the highest nucleotide identity to the Sb227 *argW* DF, the Sb227 1.0 kb *argW* associated islet and also walked into the 3' terminus of the *argW* tRNA. However on closer inspection of the sequence I noticed that the first 13 bp of the quality sequence did not match to the *argW* tRNA, also that the SGSP-PCR amplicon was generated from the *EcoRI* genomic library, but the *argW* tRNA sequence in Sb227 does not contain an *EcoRI* site, so the restriction profile of this sequence was different to the entire *argW* tRNA. This suggested that the sequence obtained comprised 13 bp of novel DNA, a 24 bp DR of the 3' end of the *argW* tRNA then the corresponding downstream sequence and DF present in Sb227. The D# amplicon therefore, walked 13 bp into the distal end of a 'classical' integrative element found associated with *argW* in S116 (see Figure 6.5). It was clearly important to further characterise this site and U# SGSP-PCR with the two extra restriction libraries produced a 1.4 kb *HincII* amplicon that walked into the element. Blastn sequence analysis of this amplicon indicated the presence of the *argW* UF and an intact *argW* tRNA with 100% nucleotide identity to Sb227; however, directly downstream of the tRNA was 888 bp of sequence that was completely novel at the nucleotide level (see Figure 6.5).

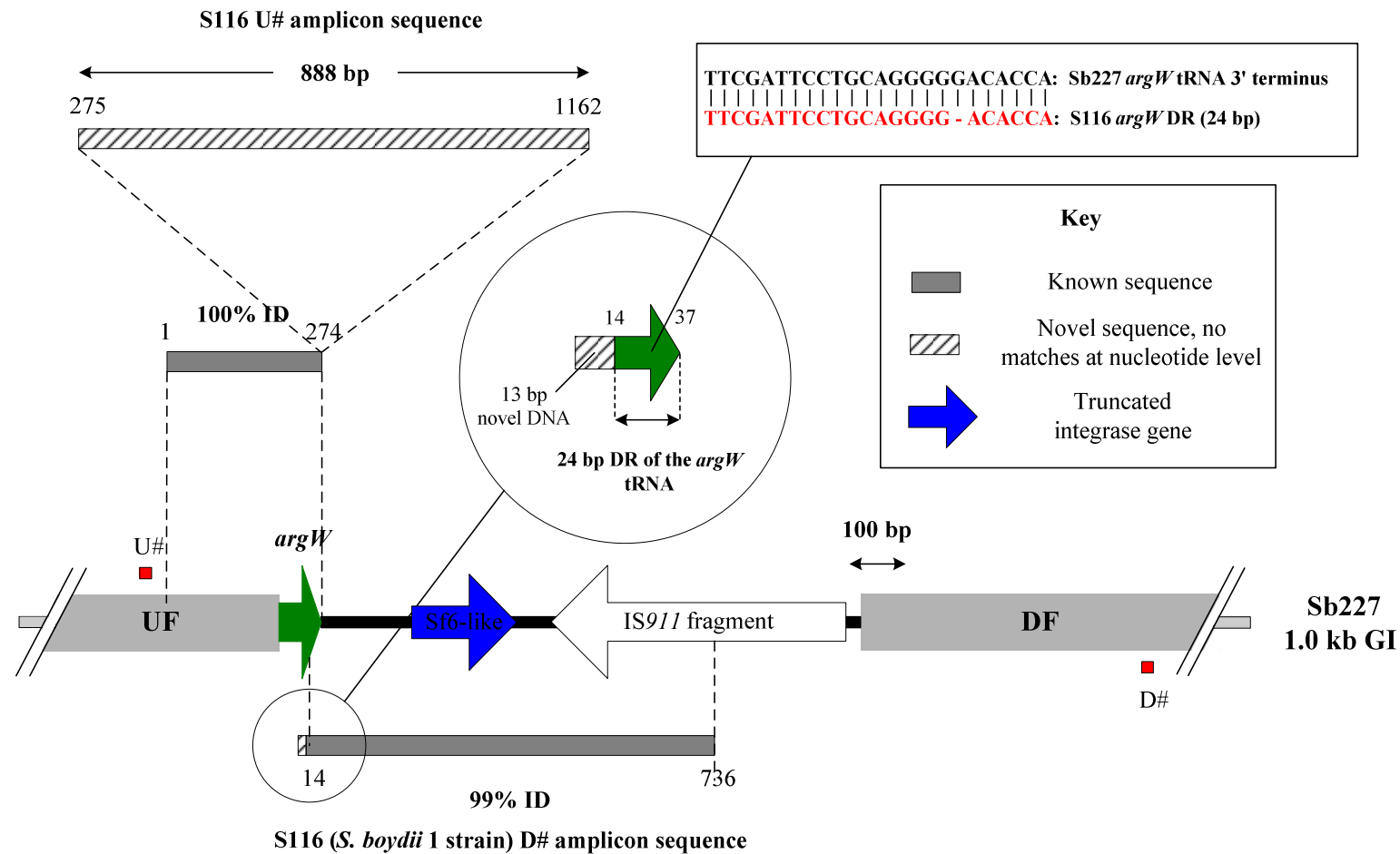


Figure 6.5. Schematic of the S116 (*S. boydii* 1) derived sequences from the *argW* U and D primer SGSP-PCRs compared with the Sb227 chromosome.

The hatched lines represent S116 derived sequence that is novel at the nucleotide level, indicating the presence of additional island DNA associated with *argW* in S116 that has displaced the original tRNA 3' terminus resulting in the formation of a 24 bp DR.

The U# derived 888 bp of novel sequence was analysed using Blastx and tBlastx, 1-184 had no significant hits and was therefore completely novel sequence. 185-886 (701 bp) was found to have amino acid similarity to the 5' end of P4-like integrase genes found in *Salmonella*, *E. coli*, *Shigella* and *Yersinia*, (see Figure 6.6) with the highest similarity being 58% to 4-705 of a 1260 bp P4-like integrase gene found in *Salmonella enterica* subsp. *enterica* serovar Paratyphi A str. ATCC 9150 (GenBank accession number AAV80021) which in turn is associated with the *leuX* GI in this strain.

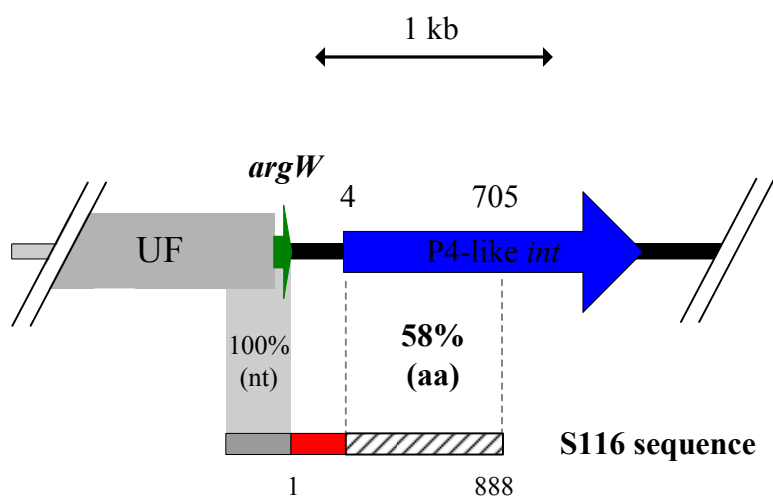


Figure 6.6. Schematic showing the start of the S116 novel GI.

Inferred from the U# sequence data as indicated by the segmented line below. The grey segment indicates the corresponding S116 sequence that has 100% nucleotide (nt) identity to the Sb227 UF and tRNA. The red segment is the sequence that is completely novel and the hatched segment is the sequence that is novel at the nucleotide level, but when translated has up to 58% amino acid (aa) similarity to P4-like integrase genes found across the Enterobacteriaceae.

These results indicated that S116 harboured a novel GI that contains a prophage-like integrase directly downstream of the *argW* locus, it was therefore worth interrogating this element further, this was carried out using a technique termed ‘island probing’ described below.

6.3 S116 *argW* island probing

6.3.1 Tagging of the *argW* UF

In order to further characterise the S116 *argW* associated novel island, I used a similar strategy to the island probing technique used by Rajakumar *et al.*, 1997, which was used to characterise the *S. flexneri she* PAI. This involved tagging the *she* gene by allelic exchange with a counterselectable marker, then capturing clones which harboured the marker and their subsequent sequencing to get snapshots of the islands content and to assess its size. The authors used a counterselectable marker rather than just a selectable marker in this case, to also measure the frequency at which the island was lost from the chromosome. In this study, I could not tag the island DNA as the only sequence data that was available from the GI was for the novel integrase gene described above, and this would not be a good candidate gene to tag as it could be unstable and it was very likely that there would be other integrase-like elements containing sequence similar to this in the genome of S116, which may cause problems when attempting to perform allelic exchange in a specific region. I therefore selected the *argW* UF as the region to tag. This was a useful choice because its sequence is known, it is unique, the genes present were found to be non-essential to *E. coli* (Gerdes *et al.*, 2003), so disruption of this region should not have any negative effects on the host and most importantly, it is conserved across many strains of *E. coli* and *Shigella*, so it very likely to be present in S116. In addition the same construct used to introduce a marker into the S116 *argW* UF could be used with other strains in the future.

6.3.2 Choice of marker

I decided to tag the S116 *argW* UF with a selectable marker; a counterselectable marker would be no extra use in this case because the UFs of tRNA loci are stable, core DNA. I selected a kanamycin resistance cassette, because S116 was found to be kanamycin sensitive, transconjugants could then be selected on kanamycin containing media; also none of the vectors chosen to manipulate the allelic exchange construct confer kanamycin resistance, this

helped with the cloning steps. In addition, *Shigella* are rarely naturally resistant to kanamycin; this further increased the chances of being able to utilise the same *argW* construct for island probing in many strains. The kanamycin resistance cassette used (from now on termed Km^r cassette) was derived from Tn5 (GenBank accession number U00004) which was harboured on pRT733 (see Table 2.3); primers were designed to PCR amplify the promoter sequence, kanamycin/neomycin resistance gene and bleomycin resistance gene. These primers were designated CF# and CR# (see Table A2. 2) and were designed to also incorporate *Nsi*I restriction sites into the ends of the PCR product, so that the cassette could be ligated into the cloned *argW* UF region to be used for allelic exchange (see section 6.3.3 below).

6.3.3 *argW* UF region chosen for allelic exchange

In order to insert the Km^r cassette into the *argW* UF of S116, firstly a mutant allele of part of the UF was constructed which contained the Km^r cassette, this construct was ligated into the suicide vector (pDS132), and the suicide construct was introduced by conjugation into S116 so that homologous recombination between the mutant UF region and the S116 chromosomal UF region could occur (see 2.16 for methodology). As I was to use standard allelic exchange, at least 300 bp of UF sequence either side of the cassette was required for the homologous recombination procedure to be specific and efficient. Also the length of sequence on either side of the cassette had to be ‘balanced’ as large discrepancies between each side can result in drastically reduced efficiency of recombination (R. Haigh personal communication). Inspection of the *argW* UFs in the sequenced genomes showed that the Sf301 *argW* UF contained an *Nsi*I restriction site, this is useful as none of the commonly used cloning vectors or the suicide vector pDS132 contain this site, so I could use this as a unique site to insert the Km^r cassette. Therefore I designed two primers which were used to PCR amplify, using Sf301 genomic DNA as the template, 973 bp of the *argW* UF with the *Nsi*I site positioned centrally (see Figure 6.7). The primers were designated *argW* UFF# and *argW* UFR# (see Table A2. 2) and were also designed to incorporate *Xba*I sites into the ends of the PCR

product, so that the region could then be cloned into pBluescript and pDS132. The region chosen had at least 97% nucleotide identity to the *E. coli* and *Shigella* genomes available on the NCBI, so was therefore very likely to be well conserved in S116.

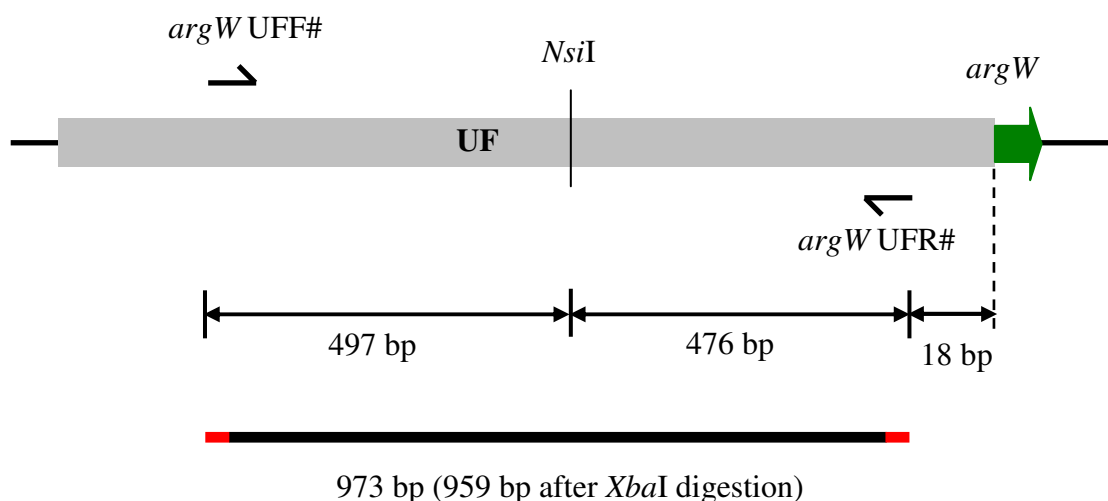


Figure 6.7. The Sf301 *argW* UF showing the positions of the primers used to amplify the region used for homologous recombination with the S116 *argW* UF.

The UF is shown in grey, and the location of the *Nsi*I site indicates where the Km^r cassette will be subsequently ligated. The thick black line indicates the expected PCR amplicon with flanking *Xba*I sites shown in red. Figure is not to scale.

6.3.4 Construction of the *argW* Km^r tagged UF

The PCR generated *argW* UF region was amplified using hot-start PCR with an extension time of 1 min. A relatively high concentration of template (~200 ng) was used and the reaction was cycled only 25 times, this was to minimise the number of PCR errors incorporated into the product (Horton *et al.*, 1989), to reduce the chance of the *Nsi*I site being disrupted. The product was electrophoresed, gel extracted, cleaned and 500 ng was digested with *Xba*I and ligated into pBluescript/*Xba*I to produce pJL5 (see Figure 6.8 (a)). The ligation was electroporated into *E. coli* DH5 α (see 2.14) and the cells plated onto LA containing 100 μ g/ml Ap and 40 μ g/ml X-gal, to screen for ampicillin resistant white colonies that contained the insert. Potential pJL5 candidates were checked by digestion with *Xba*I and *Nsi*I (see Figure 6.8 (b)).

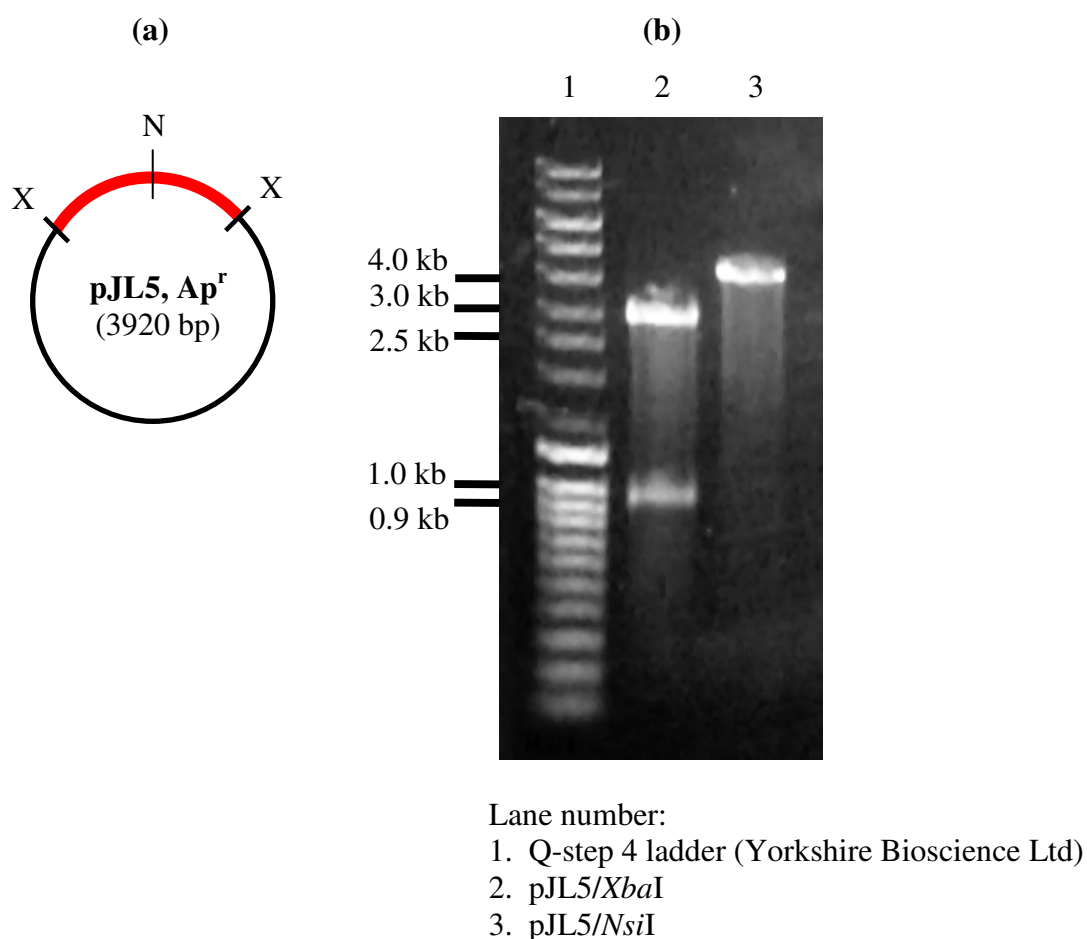


Figure 6.8. pJL5

(a). pJL5, the black segment represents pBluescript and the red segment represents the *argW* UF region. X and N stand for *Xba*I and *Nsi*I sites respectively, figure is not to scale.

(b). Agarose gel showing pJL5 digested with *Xba*I and *Nsi*I. Lane 2 shows the two *Xba*I digest fragments of 959 bp (*argW* UF region) and 2961 bp (pBluescript). Lane 3 shows pJL5 linearised by digestion with *Nsi*I.

6.3.5 Insertion and orientation of the Km^r cassette in the *argW* UF region

Next, the hot-start PCR amplified Km^r cassette (1361 bp) was electrophoresed, gel extracted, cleaned and 500 ng was digested with *Nsi*I and ligated into pJL5/*Nsi*I. The ligation was electroporated into *E. coli* DH5 α and the cells plated onto LA containing 50 μ g/ml Km and 100 μ g/ml Ap. Some of the Ap^r + Km^r transformants were screened to verify the orientation

of the Km^r cassette relative to the surrounding *argW* UF DNA. This was an important step, because previous studies have shown that the orientation of the cassette when in the host chromosome can alter its expression levels and therefore the number of recombinants that are recovered from the allelic exchange (Murphy and Campellone, 2003). The authors found that in some situations the number of recombinants was 10 fold higher if the cassette was positioned so that it was colinear with the surrounding genes, rather than in the opposite orientation. I therefore aimed to utilise the vector which contained this optimal conformation as shown in Figure 6.9.

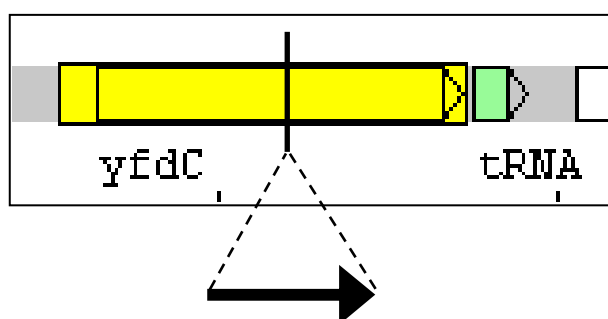


Figure 6.9. View of the Sf301 *argW* UF region taken from Artemis showing the optimal orientation for the Km^r cassette

The green box is the *argW* gene; its orientation is shown by the associated arrowhead. The yellow area indicates the region in the UF where homologous recombination would take place. Within this region is the non-essential gene *yfdC*, its orientation is shown by the associated arrowhead. The vertical black line indicates where the *yfdC* gene would be disrupted by the Km^r cassette, the black arrow (not to scale) indicates the preferred orientation of the Km^r cassette as it is colinear with *yfdC* and *argW*.

The candidate plasmids that I had constructed from pJL5 could have had the insert DNA in any of four different conformations (see Figure 6.10). I constructed all four plasmids *in silico* using EditSeq, and then using MapDraw (both programs are DNASTar™ software), searched

for restriction enzymes present in the multiple cloning site (MCS) of pBluescript and in the insert DNA that could be used to distinguish between the four orientations.

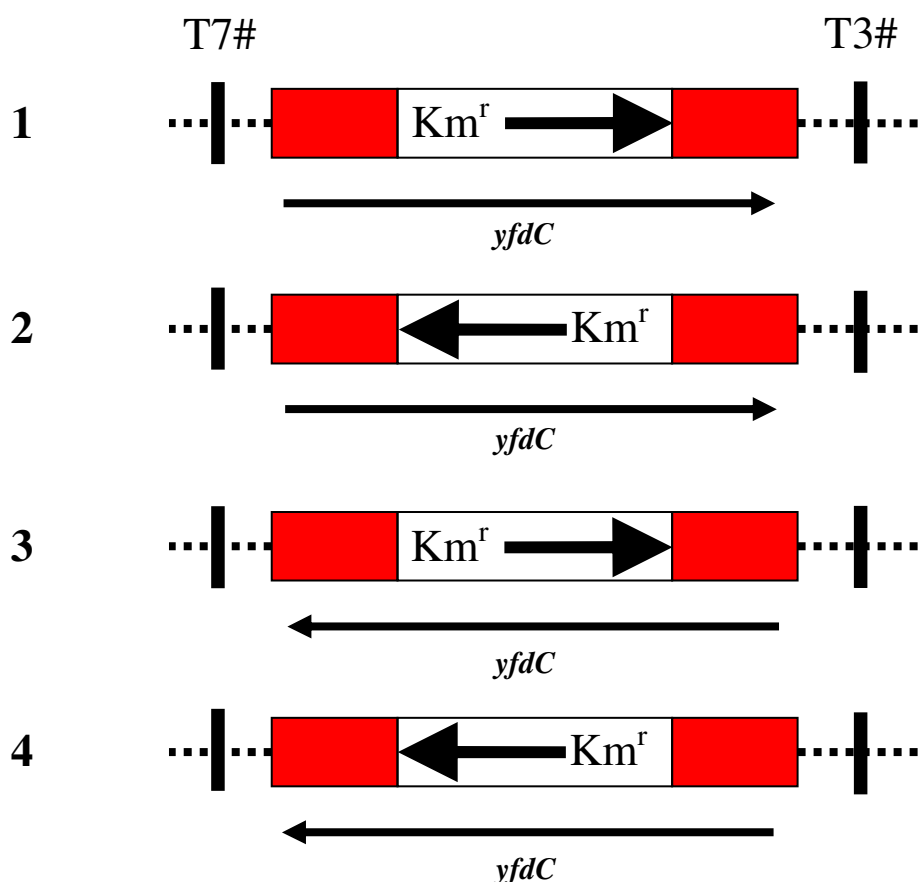


Figure 6.10. The four possible conformations of the insert DNA in the pJL5/*NsiI*:: Km^r cassette/*NsiI* constructs.

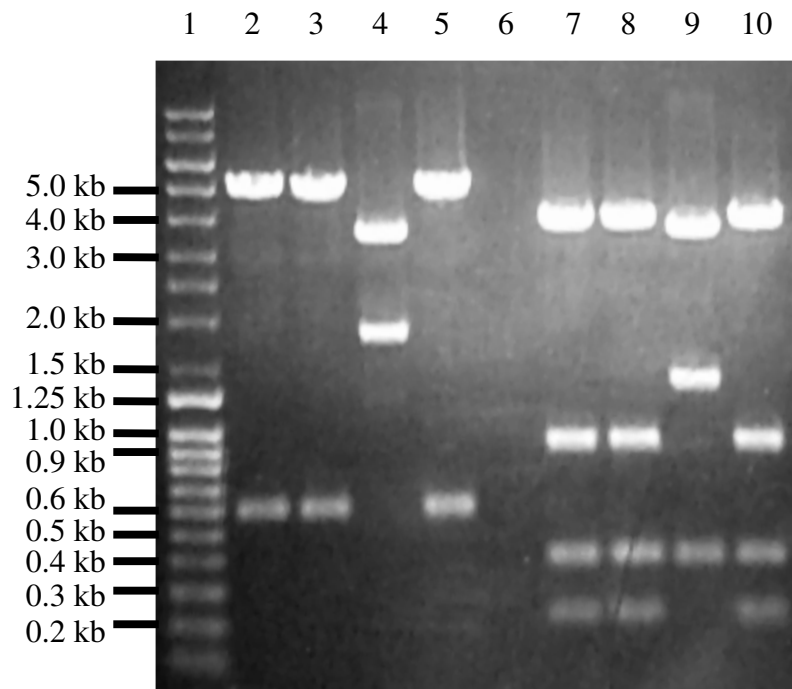
The red area is the *argW* UF region which contains the *yfdC* gene, disrupted by the Km^r cassette in white. The arrows indicate the orientation of the Km^r cassette and *yfdC* gene. The dashed lines represent the pBluescript vector backbone and the black vertical bars represent the locations of the T7 and T3 primers. Figure is not to scale.

I found that *SalI* cuts once in the MCS of pBluescript and once towards the 3' end of the Km^r cassette, so this enzyme could be used to verify the orientation of the cassette. Also *PstI* cuts once in the MCS, twice in the cassette and once in the *argW* UF region, so this enzyme could be used to determine the orientation of the UF sequence and the cassette (see Table 6.2).

Table 6.2. Restriction fragments produced by the four possible pJL5/*Nsi*I::Km^r/*Nsi*I constructs.

	Plasmid conformation and fragment sizes produced (bp)			
Restriction enzyme	1	2	3	4
<i>Sal</i>I	593, 4674	1810, 3457	610, 4657	1827, 3440
<i>Pst</i>I	206, 387, 923, 3751	387, 434, 923, 3523	434, 610, 923, 3300	206, 838, 923, 3300

Four transformants were screened using both of the enzymes separately (to be absolutely sure of the conformations, as partial digestion of some of the *Pst*I fragments could cause confusion). As the above plasmids are all derived from a single clone of pJL5, it was possible that I could have generated plasmids 1 and 2, or 3 and 4, but no other combination, Figure 6.11 shows the results of the digests.



Lane number:

1. Q-step 4 ladder (Yorkshire Biosciences Ltd)
2. Plasmid 1/*SalI*
3. Plasmid 2/*SalI*
4. Plasmid 3/*SalI*
5. Plasmid 4/*SalI*
6. Empty lane
7. Plasmid 1/*PstI*
8. Plasmid 2/*PstI*
9. Plasmid 3/*PstI*
10. Plasmid 4/*PstI*

Figure 6.11. Agarose gel showing the *SalI* and *PstI* digests of four potential pJL5/*NsiI*::Km^r cassette/*NsiI* constructs.

This clearly shows that plasmids 1, 2 and 4 have the insert in conformation 1 (see Figure 6.10) and plasmid 3 has the insert in conformation 2 (in the *PstI* digest the 434 bp and 923 bp fragments were not cut, this gave a partial digest band of 1357 bp).

I therefore had constructs with the Km^r cassette ligated in both orientations relative to the *argW* UF region; these were designated pJL6 and pJL7 (see Figure 6.12).

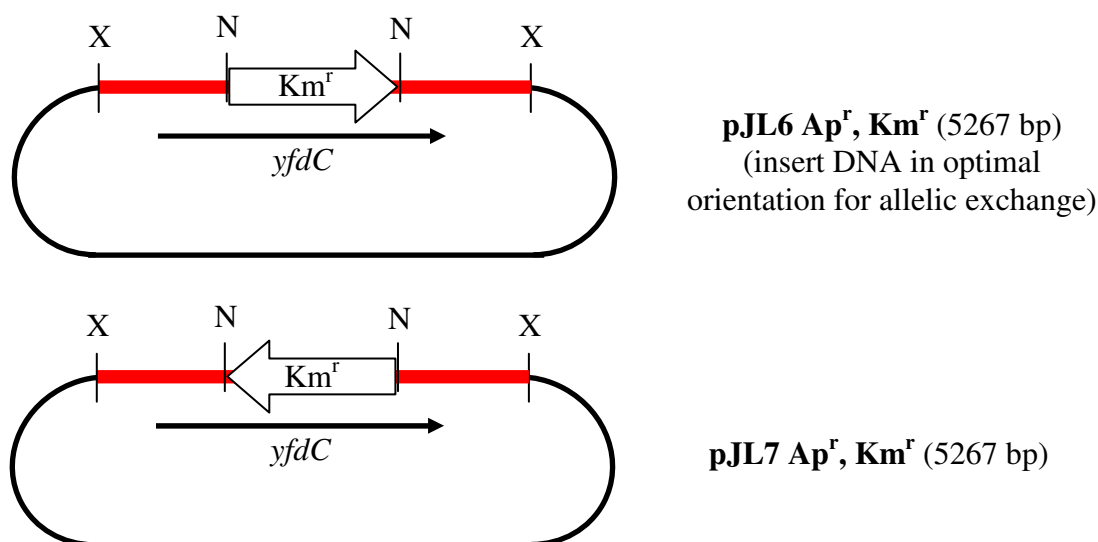


Figure 6.12. pJL6 and pJL7.

The black segment represents pBluescript and the red segment represents the *argW* UF region, the white arrow is the Km^r cassette and its orientation, the black arrow indicates the orientation of the *yfdC* gene. X and N stand for *Xba*I and *Nsi*I sites respectively. Drawings are not to scale.

6.3.6 Insertion and orientation of the mutant *argW* UF region in pDS132

1 µg of pJL6 was digested with *Xba*I to produce two fragments of 2961 bp (pBluescript) and 2306 bp (the *argW* UF mutant construct). This was electrophoresed, the 2306 bp fragment gel extracted, cleaned and ligated into the suicide vector pDS132 (Cm^r) which was also *Xba*I digested. The ligation was electroporated into *E. coli* CC118λ*pir* as this strain supports the replication of the suicide plasmid to low copy-number. The cells were plated onto LA plus 50 µg/ml Km and 30 µg/ml Cm. Transformants were screened by restriction endonuclease digestion to verify the orientation of the Km^r cassette relative to the *sacB* gene in pDS132. This is because the Km^r cassette has a strong promoter, if orientated in the opposite direction to the *sacB* gene, it may affect the expression of the *sacB* genes product when inserted into

the recipient *Shigella* chromosome (K. Rajakumar personal communication); therefore I aimed to obtain constructs with the insert DNA in both orientations to maximise the chances of generating recombinants from the conjugation experiments. The plasmids generated from the two orientations that the construct could be ligated into pDS132 were named pJL8 and pJL9 (see Figure 6.13).

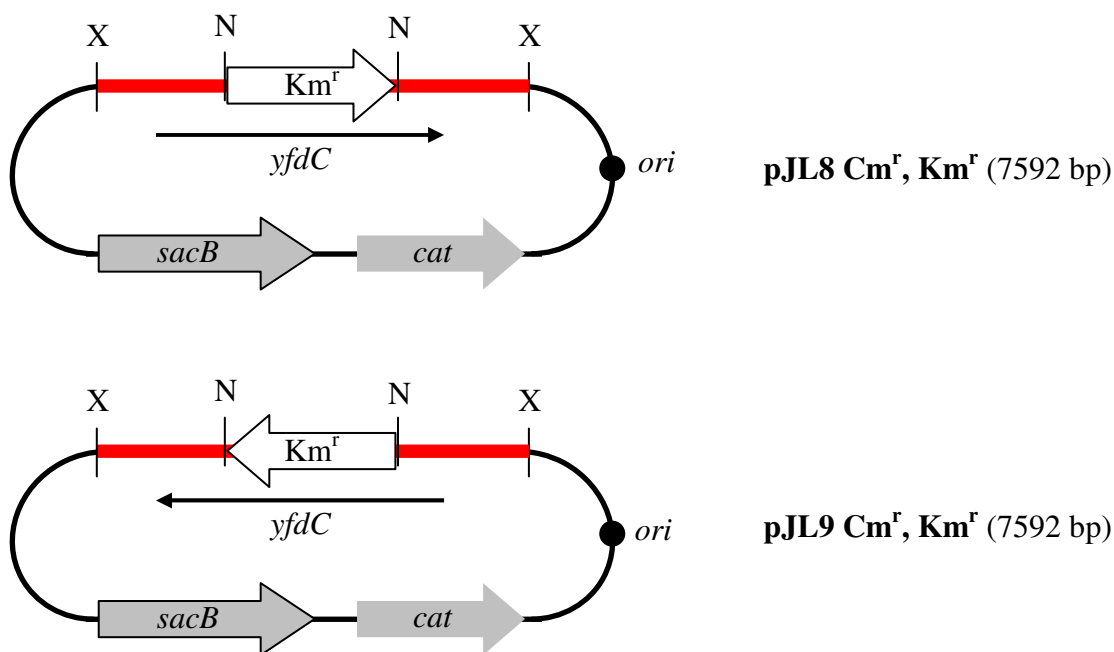


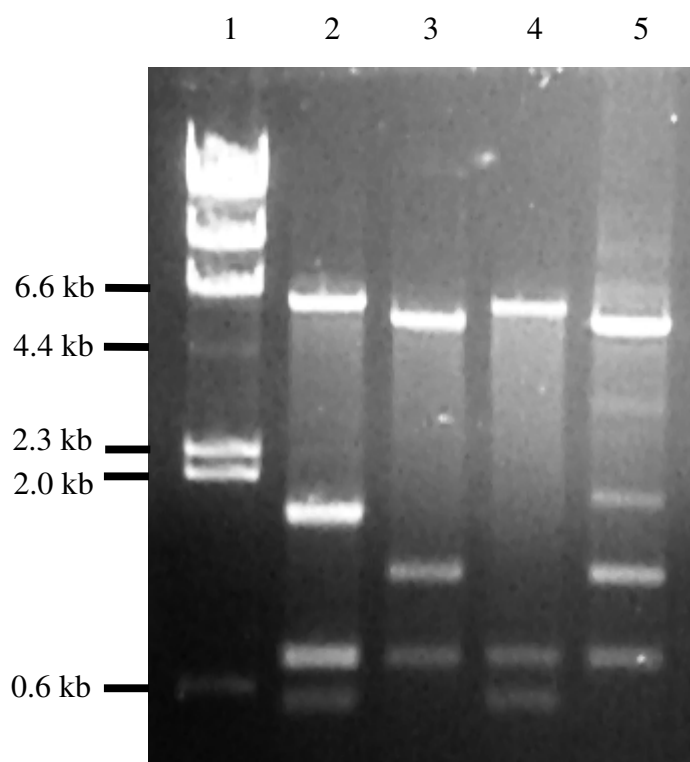
Figure 6.13. pJL8 and pJL9: pDS132 derived suicide constructs used to deliver the mutant *argW* UF region to S116.

pJL9 is the preferred choice as the *Km^r* cassette is colinear with the *sacB*, however, both constructs were used in the allelic exchange work to maximise the chances of generating recombinants.

Using EditSeq and MapDraw (DNASar™ software) I found that *SphI* cuts once in the MCS of pDS132 and twice in the insert DNA. I therefore used this enzyme to screen four recombinants to find both pJL8 and pJL9 (see Table 6.3 and Figure 6.14).

Table 6.3. *SphI* restriction fragments produced by the *argW* UF region suicide plasmid constructs.

	Plasmid and fragment sizes Produced (bp)	
Restriction enzyme	pJL8	pJL9
<i>SphI</i>	676, 1152, 5764	502, 676, 6414



Lane number:

1. λ /HindIII ladder
2. Plasmid 1/*SphI*
3. Plasmid 2/*SphI*
4. Plasmid 3/*SphI*
5. Plasmid 4/*SphI*

Figure 6.14. Agarose gel showing the *SphI* digests of four potential pJL8/pJL9 candidates.

This shows that plasmid 2 is pJL8 and plasmid 3 is pJL9. Plasmids 4 and 1 are partially digested pJL8 and pJL9 respectively.

I therefore had suicide constructs with the Km^r cassette in both orientations with respect to the *sacB* gene in pDS132; the CC118 λ *pir* strains that harboured these constructs were frozen and stored (see Table 2.3). The constructs were then electroporated separately into *E. coli* SM10 λ *pir* and the cells plated onto LA plus 50 μ g/ml Km and 30 μ g/ml Cm. Transformants harbouring pJL8 and pJL9 were frozen and stored (KR242 and KR243 respectively, see Table 2.3). Both strains were used in conjugations with the recipient *Shigella* strain.

6.3.7 Generation of X102 and X103: ampicillin resistant derivatives of S116

After a typical conjugation experiment, the mixture of donor and recipient cells must be plated on selective medium, so that only the transconjugant recipient cells survive and grow to single colonies. In this case, another selection was therefore required, in addition to kanamycin resistance which was used to select for transconjugant *Shigellas*, but could not be used to select between donor and recipient as *E. coli* SM10 λ *pir* is also Kanamycin resistant.

S116 was found to be sensitive to Ap, Cm, Tc and Str and also did not grow on *Shigella* specific minimal medium (Ahmed *et al.*, 1988) which is often used to select for *Shigella* after the conjugation step, as *E. coli* SM10 λ *pir* are unable to grow on this medium. So initially none of these could be used to select for S116. One possible solution was to also add X-gal to the medium and use blue/white selection to distinguish between donor and recipient. *E. coli* SM10 λ *pir* cells turn blue when on plates containing X-gal because they harbour the *lac* operon on their chromosome, this contains the *lacZ* gene, which encodes β -galactosidase. This enzyme hydrolyses X-gal, and one of the products from the reaction forms a blue precipitate, so the colonies on the plate appear blue. *Shigella* however, generally do not harbour the *lac* genes (Ito *et al.*, 1991) and would therefore not produce the blue colour in the presence of X-gal. This was the case with S116; its colonies were white when plated on LA plus X-gal. So LA plus Km and X-gal could have been used to screen for transconjugant

S116 clones, however due to the high density of cells plated on the medium and the low frequency of recombination events, this would not be a robust enough method to select for the transconjugant *Shigellas*. Therefore, S116 was made ampicillin resistant by electroporating it with pWSK29 (low copy-number Ap^r plasmid (Posfai *et al.*, 1997) and pBluescript (High copy-number Ap^r plasmid), separately. These plasmids both confer ampicillin resistance to 100 µg/ml in *E. coli*, but in *Shigella* this was previously unknown. After electroporation the cells were plated onto LA plus 100 µg/ml Ap and X-gal, this was included because both pBluescript and pWSK29 contain the 5' terminal part of the *lacZ* gene so the plasmids can be used to complement the mutant *lacZ* gene present in certain strains of *E. coli* such as DH5α so that blue/white screening can be performed. So there was a very slim chance that this would enable S116 to produce β-galactosidase if it has a partial *lacZ* in its chromosome, therefore producing blue colonies, this however was very unlikely (Ito *et al.*, 1991). This was important to determine as I planned to use the white phenotype as the final test to prove that my transconjugant S116 clones really were *Shigella* and not a contaminating *E. coli*. Both clones grew well to single colonies on this medium, and the colonies were white as expected. These clones were named X102 (S116 harbouring pWSK29) and X103 (S116 harbouring pBluescript).

6.3.8 X102 allelic exchange

X102 was used in conjugations with KR242 and KR243 (see section 2.16 for the methodology), as in X102 the pWSK29 is present in low copy number and was less likely to interfere with the suicide construct acquired during conjugation than the high copy number pBluescript. Transconjugant X102 clones were selected for on LA plus 50 µg/ml Km and 100 µg/ml Ap, then streaked onto LA plus 40 µg/ml X-gal to confirm they were *Shigellas*.

6.3.9 X102 sucrose selection

X102 transconjugants derived from conjugations with both the pJL8 and pJL9 bearing donor strains were used in a sucrose selection step (see 2.16.2 for methodology) to select for clones that had undergone a double crossover, lost the suicide vector and retained the Km^r cassette. Initially, in addition to this protocol, prior to plating on 6% (w/v) sucrose agar, the LB cultures were diluted 1 in 1000 into 5 ml of 6% (w/v) sucrose broth and incubated at 30°C, 200 rpm for 12 hr to enrich for sucrose resistant colonies. These cultures were then diluted to 10⁻⁵ to 10⁻⁷ – fold in LB and 100 µl aliquots spread onto 6% (w/v) sucrose agar plates and incubated overnight at 30°C. Sucrose resistant clones were plated onto LA plus 50 µg/ml Km and 100 µg/ml Ap to select for X102 derivatives that still harboured the Km^r cassette, as when the double crossover event occurs, there is a 50% chance that the wild type allele is reconstituted.

6.3.10 Screening of X102 potential transconjugants.

Sucrose resistant, Km resistant, X102 transconjugants were screened firstly, with primers that were positioned external to the allelic exchange region, to check for a possible double crossover event and the presence of the Km^r cassette in the *argW* UF. The primers were designated *argW* U#2 which is positioned upstream in the *argW* UF, and *argW* tRNArev, which is positioned on the complementary strand in the *argW* tRNA gene (see Figure 6.15).

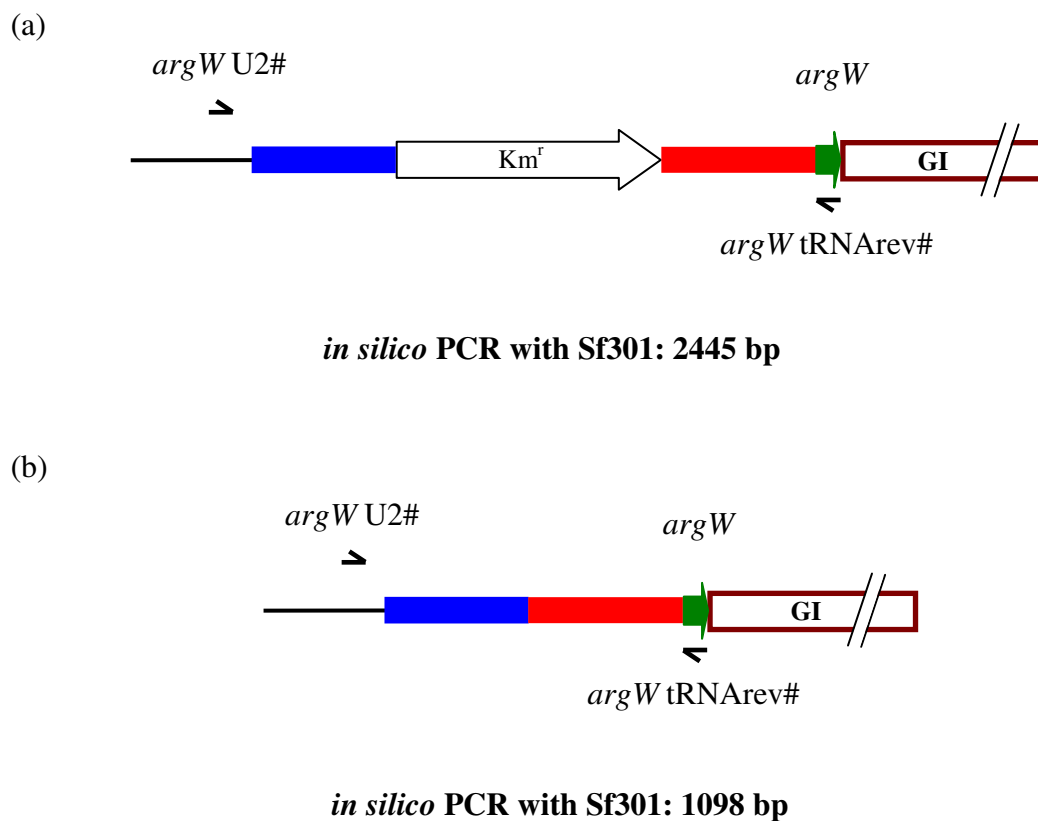
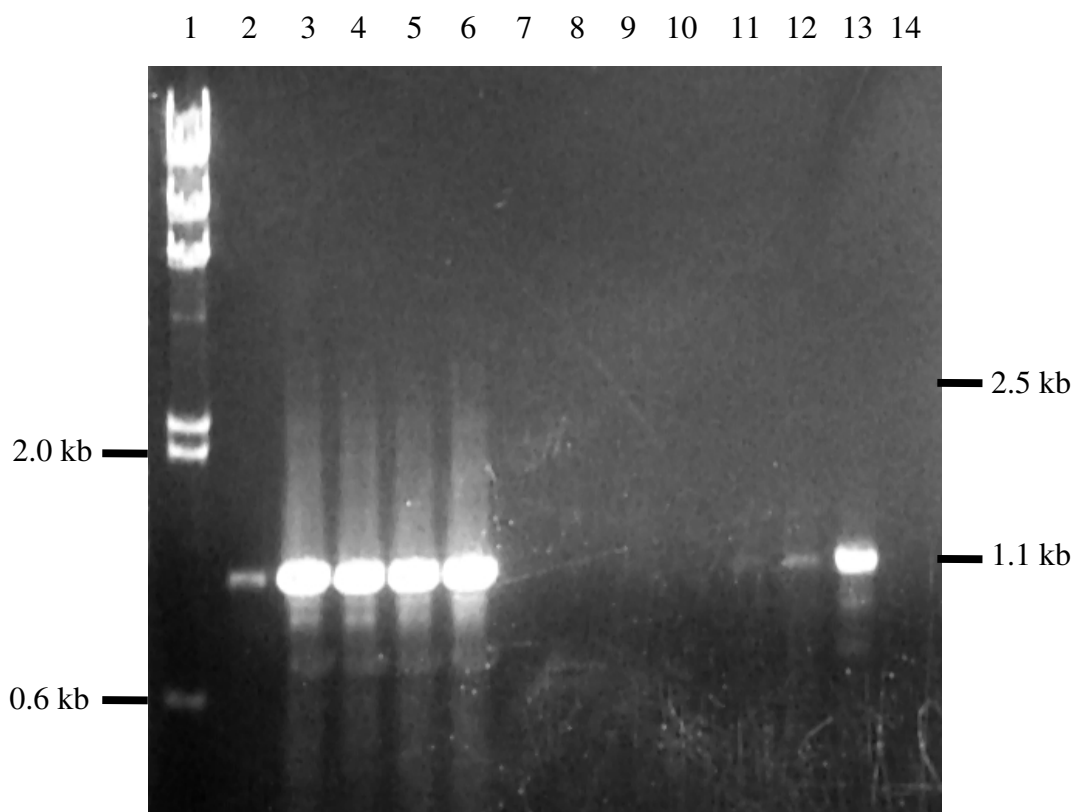


Figure 6.15. Schematics showing the *argW* UF region with the presence (a) or absence (b) of the Km^r cassette, and the respective *in silico* PCR amplicons using Sf301 as the template genome.

The half arrows represent the primers used to verify the event. The blue and red boxes indicate the parts of the UF region that are separated by the Km^r cassette. Drawings are not to scale.

Colony PCR on five potential transconjugants derived from conjugations with both the pJL8 and pJL9 bearing donor strains were performed (standard, touch-down PCR, extension time 2.5 min) (see Figure 6.16).



Lane number:

1. λ /*Hind*III ladder
2. pJL8 derived transconjugant 1
3. pJL8 derived transconjugant 2
4. pJL8 derived transconjugant 3
5. pJL8 derived transconjugant 4
6. pJL8 derived transconjugant 5
7. Empty lane
8. pJL9 derived transconjugant 1
9. pJL9 derived transconjugant 2
10. pJL9 derived transconjugant 3
11. pJL9 derived transconjugant 4
12. pJL9 derived transconjugant 5
13. X102 control (no Km^r cassette in *argW* UF)
14. Negative PCR control (no template)

Figure 6.16. Agarose gel showing the results of the colony PCRs on both pJL8 and pJL9 derived potential transconjugant X102 *Shigellas*.

The absence of amplicons in lanes 8, 9 and 10 indicate that these clones are merodiploids that harbour the entire suicide construct in the *argW* UF and had not undergone the double crossover event.

The results clearly show that all of the pJL8 derived transconjugants do not harbour the Km^r cassette in the *argW* UF region; neither do the pJL9 derived transconjugants 4 and 5 as they all yielded amplicons that were the same size as the control X102. Whereas in lanes 8, 9 and 10 you can see that the pJL9 derived transconjugants yielded no amplicon, indicating that these have an insertion larger than at least 2.5 kb in the *argW* UF. This indicated that the sucrose selection had not caused these clones to undergo the second crossover event and they were most likely to be merodiploids that had acquired resistance to sucrose through a mutation in the *sacB* gene, indicating also, that the parent pJL9 derived transconjugant was a merodiploid that had undergone the single crossover event and has the entire suicide construct in the *argW* UF. To verify this, two of these clones, along with the original pJL9 derived merodiploid used in the sucrose selection step were then screened with another panel of colony PCRs to confirm the presence of the suicide construct in the *argW* UF and the position of the Km^r cassette as there were two possible conformations (see Figure 6.17 and Figure 6.18).

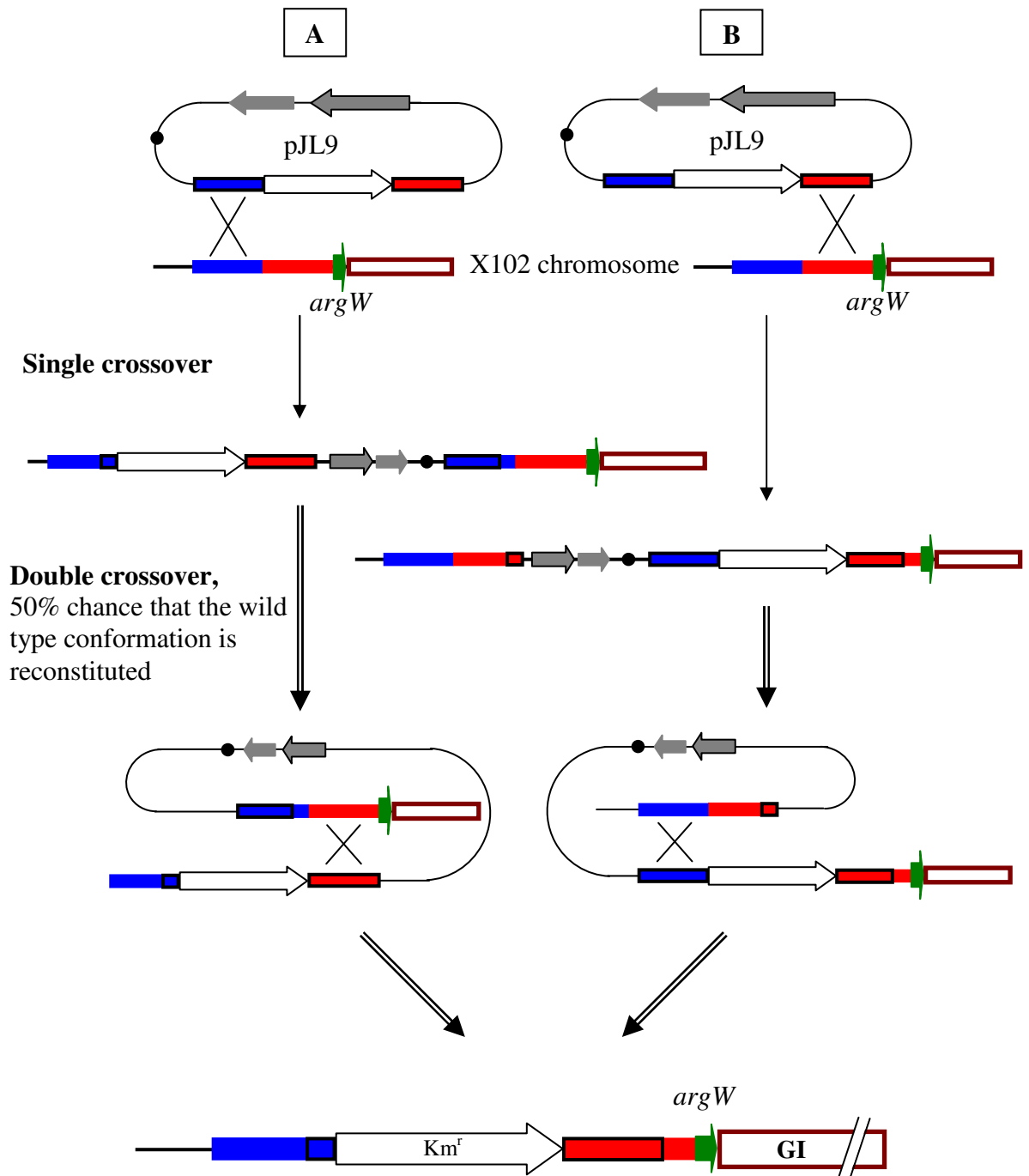
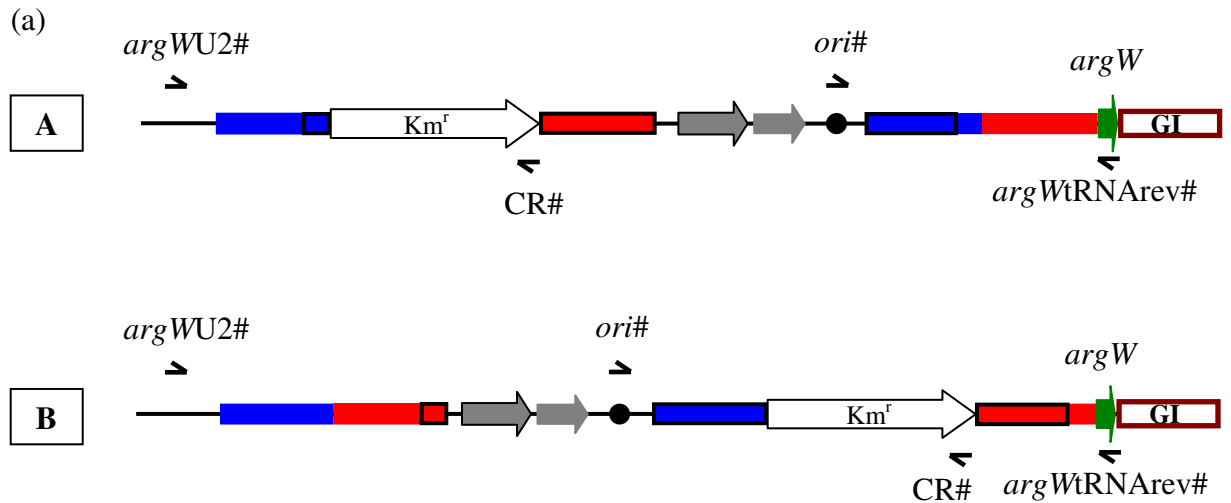
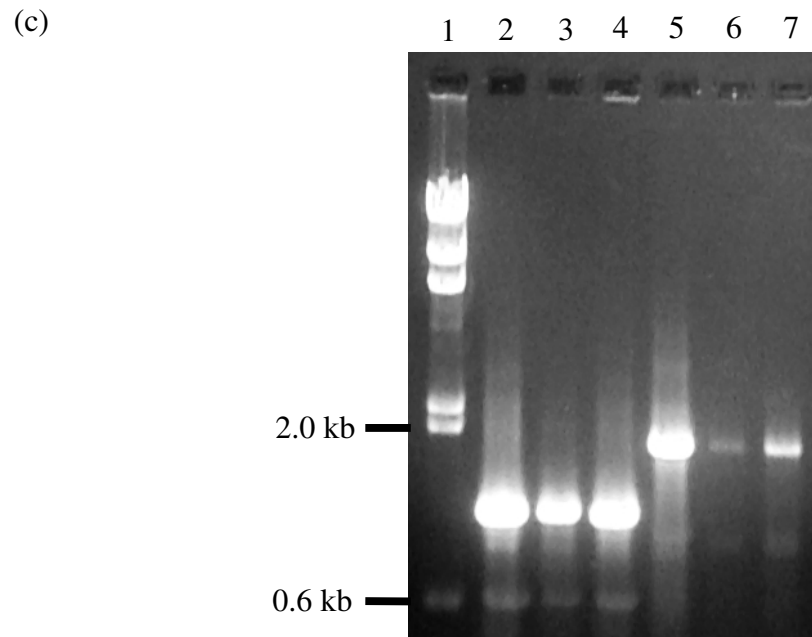


Figure 6.17. The single and double crossover events that could occur at the *argW* UF region in the chromosome of the pJL9 derived X102 transconjugant *Shigellas*.

The blue and red regions indicate the parts of the *argW* UF that are separated by the *Km^r* cassette. Drawings are not to scale.



- (b)
- A** $ori\# - argW\ tRNA^{Arev\#} = 1159\text{ bp}$, $argW\ U2\# - CR\# = 1894\text{ bp}$
- B** $ori\# - argW\ tRNA^{Arev\#} = 2506\text{ bp}$, $argW\ U2\# - CR\# = > 5\text{ kb}$



Lane number:

1. $\lambda/HindIII$ ladder
 2. pJL9 derived merodiploid
 3. pJL9 derived sucrose resistant transconjugant 1
 4. pJL9 derived sucrose resistant transconjugant 2
 5. pJL9 derived merodiploid
 6. pJL9 derived sucrose resistant transconjugant 1
 7. pJL9 derived sucrose resistant transconjugant 2
- $\left. \begin{array}{l} 3. \\ 4. \end{array} \right\} ori\# - argW\ tRNA^{Arev\#}\text{ PCRs}$
 $\left. \begin{array}{l} 6. \\ 7. \end{array} \right\} argW\ U2\# - CR\#\text{ PCRs}$

Figure 6.18. Potential merodiploids derived from X102 conjugations with pJL9.

(a) The two possible conformations of the pJL9 suicide construct present in the *argW* UF of X102 after a single crossover event, the half arrows indicate the primers used to distinguish between the two conformations (see Table A2. 2 for details of primers).

(b) The expected *in silico* PCR products of the primer pairs used to distinguish between conformations A and B, using Sf301 as the template.

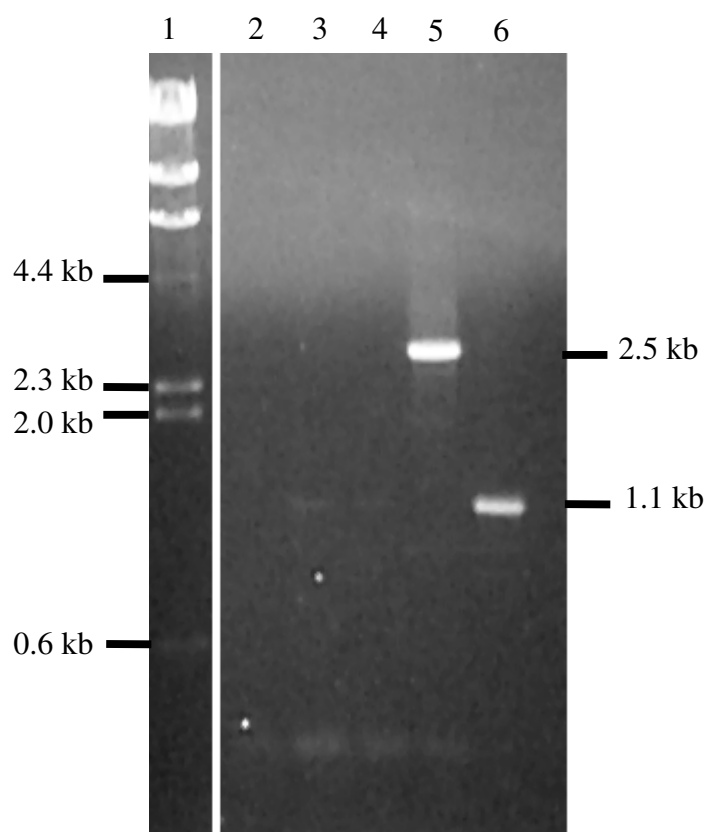
(c) Agarose gel showing the results of the PCRs on the original pJL9 derived merodiploid and two of the sucrose resistant derivatives, confirming that they were all merodiploids in conformation 'A'; this also indicated that the sucrose resistance of transconjugants 1 and 2 is very likely due to a point or frameshift mutation of the *sacB* gene which could in this case have rendered the protein product inactive. Drawings are not to scale.

The original pJL9 derived merodiploid was named X104, after screening more potential transconjugants from the pJL9 conjugation, a merodiploid in conformation B was found, this was named X105 (see Table 2.3). Both were stored as frozen glycerol stocks.

6.3.11 Refinement of the sucrose selection method

On further analysis I found that the growth of merodiploids in sucrose broth prior to plating on sucrose agar was likely to be counter-productive and it was more likely to produce many more sucrose resistant clones that had arisen through mutation of the *sacB* gene, rather than through loss of the suicide vector by a double crossover event. Therefore the sucrose selection was repeated with both X104 and X105, but the sucrose broth step was omitted, and dilutions of original 5 ml LB culture were plated directly onto 6% (w/v) sucrose agar (see 2.16.2). Sucrose resistant clones were plated onto LA plus 50 µg/ml Km and 100 µg/ml Ap to select for transconjugants that still harboured the Km^r cassette.

Colony PCR on four sucrose resistant, kanamycin resistant potential transconjugants derived from X104 were performed as before using the primers *argW* U2# and *argW* tRNArev# to check for clones that had undergone a double crossover event (see Figure 6.19).



Lane number:

1. λ /HindIII ladder
2. X104 derived transconjugant 1
3. X104 derived transconjugant 2
4. X104 derived transconjugant 3
5. X104 derived transconjugant 4
6. X102 control (no Km^r cassette in *argW* UF)

Figure 6.19. Agarose gel showing the results of the colony PCR on X104 derived potential double crossover transconjugant *Shigellae*.

Lane 5 shows that this clone produced an amplicon of around 2.5 kb and therefore was a true transconjugant that had undergone a double crossover event, with the Km^r cassette present in the *argW* UF (see Figure 6.15 (a) above). The absence of amplicons in lanes 2, 3 and 4

indicate that these clones are still merodiploids that harbour the entire suicide construct in the *argW* UF and had not undergone the double crossover event.

Transconjugant number 4 was named X106 (see Table 2.3) and stored as a frozen glycerol stock.

6.4 Marker rescue

To capture portions of the S116 *argW* associated novel GI, genomic DNA was extracted from the kanamycin resistant derivative X106 and used to generate restriction libraries in pWSK29. Each library was electroporated separately into *E. coli* DH5, the cells were then plated onto LA plus 50 µg/ml Km and 100 µg/ml Ap, to select for clones containing plasmids with inserts containing fragments of the X106 genomic DNA that harbour the Km^r cassette, the surrounding DNA and hopefully part of the GI (see Figure 6.20). To help select the enzymes used to create the restriction libraries, an *in silico* analysis of the 5 kb sequence upstream of *argW* in the known *Shigella* genomes was performed. It was found that in Sb227, *EcoRI*, *HindIII* and *SacI* all cut just upstream of the Km^r cassette in the conserved *argW* UF (4850 bp, 1385 bp and 2604 bp upstream of the *argW* tRNA respectively). These enzymes were therefore strong candidates to use for marker rescue, as the sequence data derived from S116 indicated that it was the most similar to Sb227, so there was a good chance that they cut in the same locations in the X106 *argW* UF, maximising the chances of capturing island DNA rather than mainly flanking sequence. Also, the S116 UF sequence obtained from SGSP-PCR was scrutinised and any enzymes that cut between the location of the Km^r cassette and the putative GI, were omitted, as were *PstI* and *SalI* as these cut in the Km^r cassette.



Figure 6.20. The Principle of marker rescue.

The use of enzymes that cut upstream of the Km^r cassette but as close to the tRNA locus as possible, increase the chances of capturing island DNA. *In silico* analysis of the conserved UF in the sequenced *Shigella* genomes helped select potential candidates. A and B represent hypothetical useful restriction enzyme cut sites. Figure is not to scale.

The following enzymes were selected to generate restriction libraries:

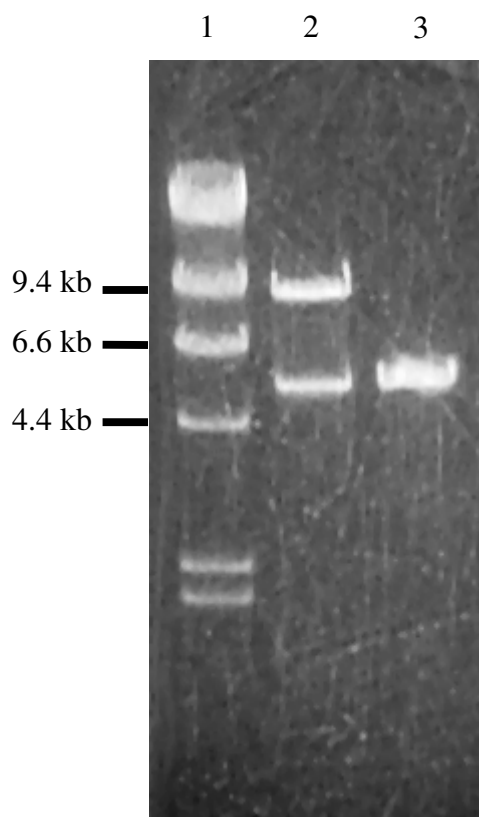
EcoRI, *HindIII*, *SacI*, *BamHI*, *KpnI*, *SacII* and *XbaI*.

6.4.1 Rescue of clones harbouring the Km^r cassette

After electroporation, and incubation in 1 ml of SOC, 900 μ l of cells from each library were pelleted by gentle centrifugation, resuspended to 200 μ l with LB and 100 μ l spread onto an LA plate containing 50 μ g/ml Km and 100 μ g/ml Ap, and 100 μ l spread onto an LA plate containing 50 μ g/ml Km. The plates were incubated at 37°C overnight, Km^r+Ap^r and Km^r clones were streaked to purity onto LA plus 50 μ g/ml Km and 100 μ g/ml Ap. In addition, the remaining 100 μ l from each electroporation was used to calculate the ligation efficiency of the libraries. 10^{-2} dilutions were plated onto LA plus 100 μ g/ml Ap and 40 μ g/ml X-gal to obtain the blue/white ratio, typical efficiencies of up to 20% white colonies were obtained.

Initially only the *EcoRI* and *HindIII* libraries produced $Km^r + Ap^r$ clones, representatives had their plasmid DNA extracted and were digested to size their inserts and end sequenced using

T3 and T7 to help map the positions of the fragments and restriction sites relative to the *argW* site.



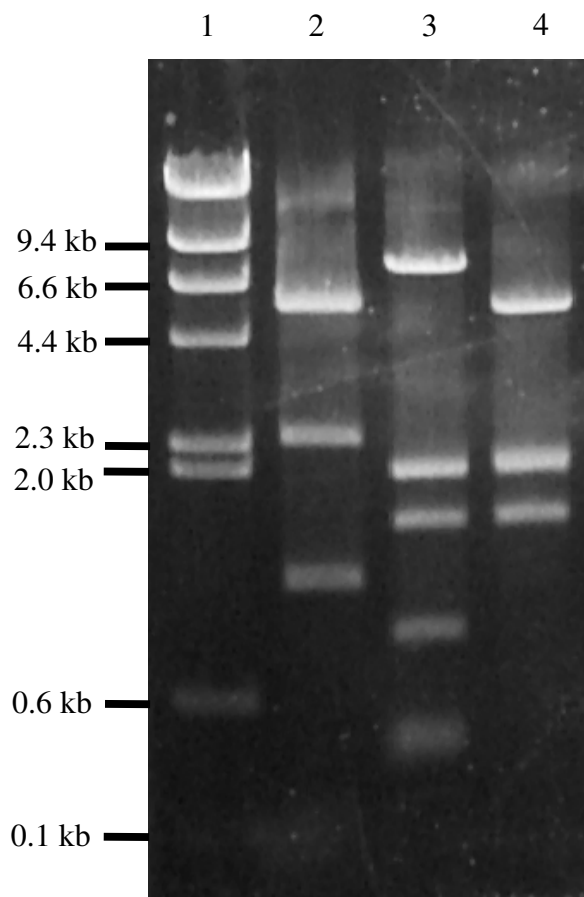
Lane number:

1. λ /*Hind*III ladder
2. *Eco*RI marker rescue clone/*Eco*RI (pJL17).....Insert of 9.3 kb
3. *Hind*III marker rescue clone/*Hind*III (pJL18).....Insert of 5.5 kb

Figure 6.21. Agarose gel showing the sizes of the insert fragments harboured by the X106 *Eco*RI and *Bam*HI marker rescue clones.

pWSK29 is 5.4 kb, the insert band in lane 3 ran to just above pWSK29 (5.4 kb), so the pJL18 insert is 5.5 kb.

The clones were also digested with other enzymes to size them more accurately (see Figure 6.22).



Lane number:

1. λ /HindIII ladder

2. pJL17/HindIII

Yielded bands of 5.4 kb (pWSK29)^a, 5.5 kb (the HindIII clone insert)^a, 2.4 kb, 1.1 kb, and 168 bp^b. Total of 14.6 kb, indicating that the insert is **9.2 kb**

3. pJL17/HincII

Yielded bands of 8.0 kb, 2.0 kb^a, 2.0 kb^a, 1.5 kb^c, 0.8 kb and 0.4 kb. Total of 14.7 kb, indicating that the insert is **9.3 kb**

4. pJL18/HincII

Yielded bands of 5.4 kb (pWSK29), 2.0 kb^a, 2.0 kb^a, and 1.5 kb^c. Total of 10.9 kb, indicating that the insert is **5.5 kb**

Figure 6.22. Agarose gel showing the restriction patterns of the X106 *Eco*RI and *Hind*III marker rescue clones.

^a Seen as a doublet on the gel

^b not seen on the gel, derived from the GI sequence data from this clone (see Figure 6.23)

^c Common to both clones, *Hinc*II fragment of 1.5 kb that cuts within the Km^r cassette and the novel P4-like integrase gene (see Figure 6.23).

6.4.2 Restriction map of the X106 *argW* GI U-arm

The restriction profile and end sequence data then enabled me to map the sequence surrounding the Km^r cassette.

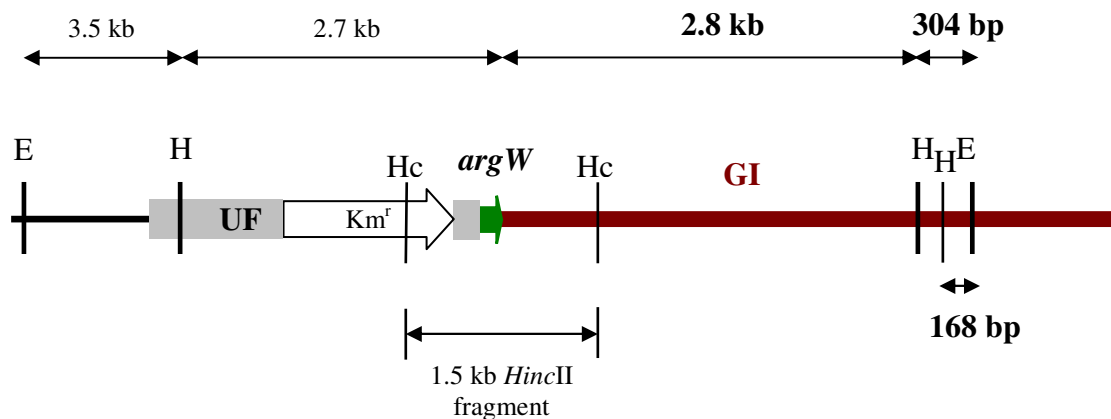


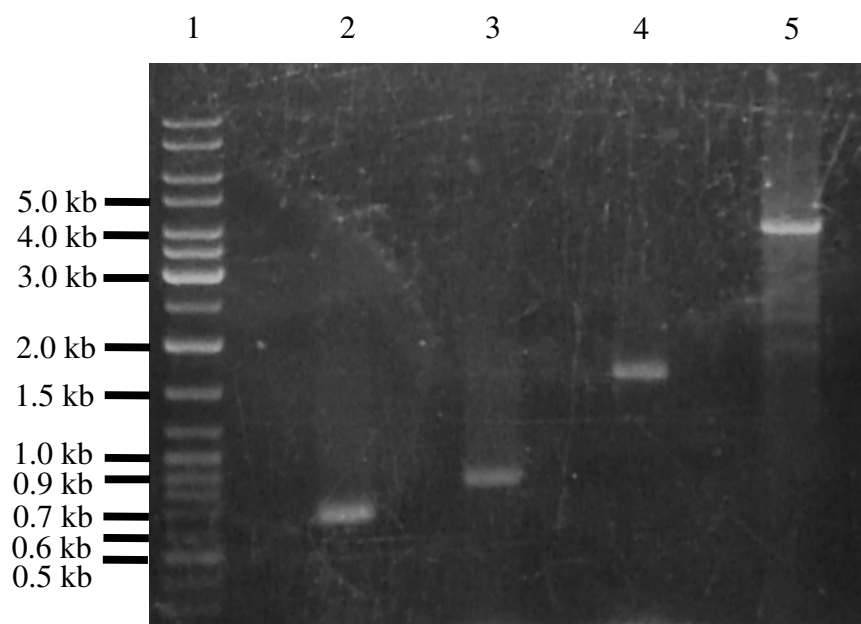
Figure 6.23. Relative positions of select restriction sites and fragment sizes around the X106 *argW* tRNA locus.

Inferred from calculations using the *in silico* Sb227 sequence, the *Eco*RI and *Hind*III clone end sequence data and the restriction fragment sizes. Accurate sizes shown in bp are derived from the marker rescue clone sequences. H, E and Hc stand for *Hind*III, *Eco*RI and *Hinc*II restriction sites respectively. Figure is not to scale.

These results indicated that I had walked 3.1 kb into the X106 *argW* GI. The 1187 bp of sequence derived from within the island (obtained from end sequencing of the marker rescue clones pJL17 and pJL18) was completely novel even at the translated level (see Figure 6.28). To gain more information on the novel GI, a primer was designed to anneal to the reverse complement of 36..60 of this novel sequence (see Figure 6.28), this primer was named X106GIinternal1#, and was used in PCRs with T7# to walk further into the GI, by using the S116 and X106 restriction libraries as template.

6.4.3 X106 (S116 Km^r strain) chromosome walking

The original X106/*SacI* library electroporation had a low efficiency, and no Km^r clones were recovered. However, I was unsure as to whether this was a truly negative result and as I was confident that there was a *SacI* site upstream of the Km^r cassette and *argW* tRNA gene in the UF (see 6.4 above) I hypothesised that there was an increased chance of recovering a *SacI* marker rescue clone. Therefore when performing primer walking experiments with X106Ginternal1#, to get more sequence data from within the novel GI, I also included the X106::pWSK29/*SacI* library. This library produced an amplicon of around 4.2 kb (see Figure 6.24).



Lane number:

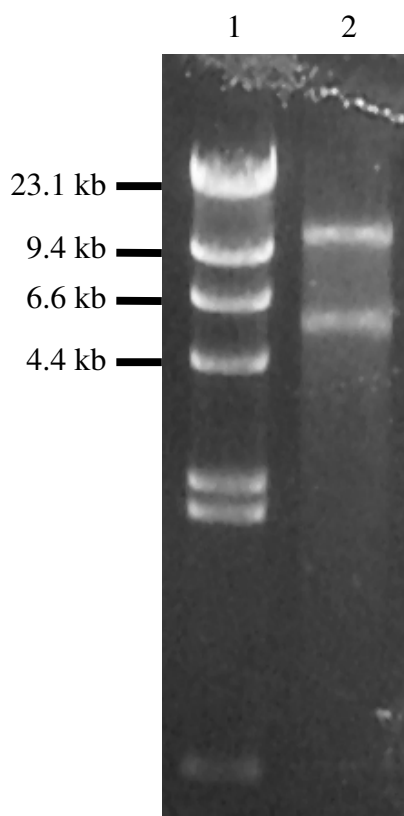
1. GeneRuler™ 1 kb ladder (Fermentas)
2. S116::pBluescript/*EcoRV* library, amplicon of 0.7 kb
3. S116::pBluescript/*HincII* library, amplicon of 0.9 kb
4. S116::pBluescript/*HindIII* library, amplicon of 1.7 kb
5. X106::pWSK29/*SacI* library, amplicon of 4.2 kb

Figure 6.24. Agarose gel showing the sizes of amplicons generated by primer walking across the X106 *argW* associated GI.

PCRs were performed with X106GIinternal1# and T7#; using different S116 or X106 genomic libraries as the template in each PCR (libraries that did not yield an amplicon are not shown).

This indicated that there was a *SacI* site 4.0 kb downstream of the *EcoRI* site which in turn was 3.1 kb into the novel GI. So I had walked 7.1 kb into putative island DNA.

The X106/*SacI* library was then electroporated into some fresh electrocompetent *E. coli* DH5 α cells, Km^r clones were recovered and subsequent plasmid extraction and *SacI* digestion showed that the clone harboured an 11.1 kb insert, as expected from the corresponding *in silico argW* UF data and primer walking results (see Figure 6.25).



Lane number:

1. λ /HindIII ladder

2. *SacI* marker rescue clone/*SacI* (pJL19).....Insert of 11.1 kb

Figure 6.25. Agarose gel showing the sizes of the restriction fragments harboured by the X106 *SacI* marker rescue clone.

pJL19 was end sequenced from T7# and T3# to map it relative to the *argW* tRNA and to obtain more sequence from within the putative GI

The majority of the island sequence obtained was found to be novel DNA; however two regions of 114 bp and 296 bp had 88% and 90% nucleotide identity respectively to the corresponding regions found 5.6 kb downstream of *argW* in the 14.1 kb prophage-like GI found only in the *E. coli* strains EDL933 and Sakai. This sequence has no other significant matches and its function is unknown (see Figure 6.28).

6.4.4 X106 *argW* GI integrase walking

As I had walked 0.7 kb into a novel P4-like integrase at the start of the GI, I aimed to obtain more sequence data downstream of this to see if the integrase was intact or truncated. To do this I designed a primer towards the end of the novel integrase sequence obtained, this was designated X106 *int* F. This was then used in a PCR with T7#, with pJL18 as the template, to walk further into the island (see Figure 6.26 and Figure 6.27).

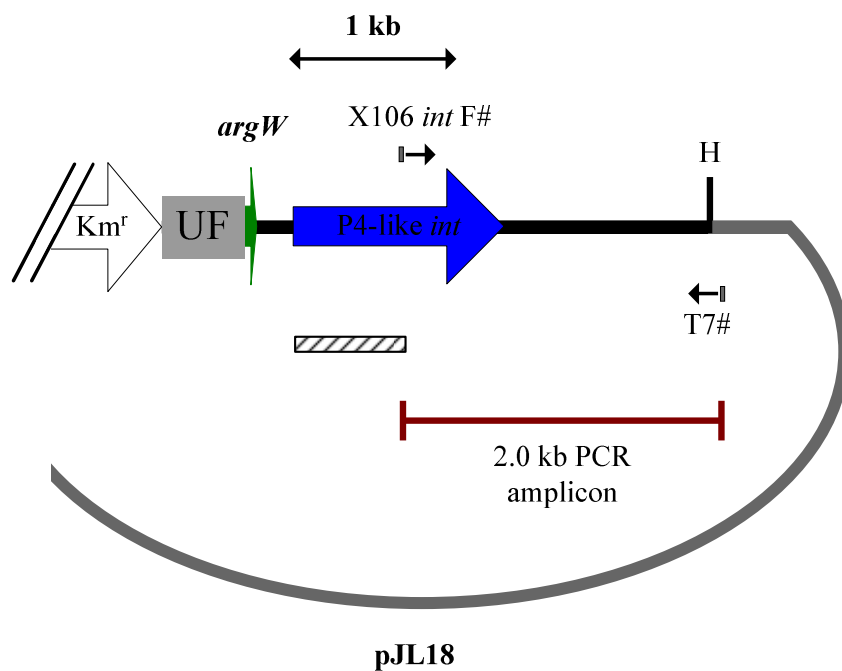
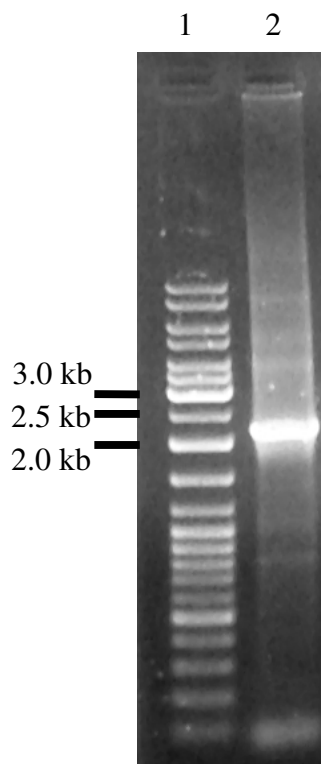


Figure 6.26. Schematic showing an *in silico* PCR using the X106 *int* F and T7 primers and pJL18 as the template to walk further into the novel X106 *argW* island.

The hatched bar represents the integrase gene sequence already obtained

Figure 6.26 shows that an amplicon of 2.0 kb was expected.



Lane number:

1. GeneRuler™ 1 kb ladder (Fermentas)
2. X106 *int* F# - T7# PCR using pJL18 as the template

Figure 6.27. Agarose gel showing the pJL18 marker rescue clone walking PCR using the X106 *int* F and T7 primers to walk further into the X106 *argW* associated novel integrase-like element.

Lane 2 shows that a 2.1 kb PCR product was obtained.

The PCR product was gel extracted, cleaned and sequenced from the X106 *int* F primer. The sequence data obtained is shown in Figure 6.28.

6.4.5 Schematic of the S116 *argW* associated novel prophage-like GI

See Figure 6.28

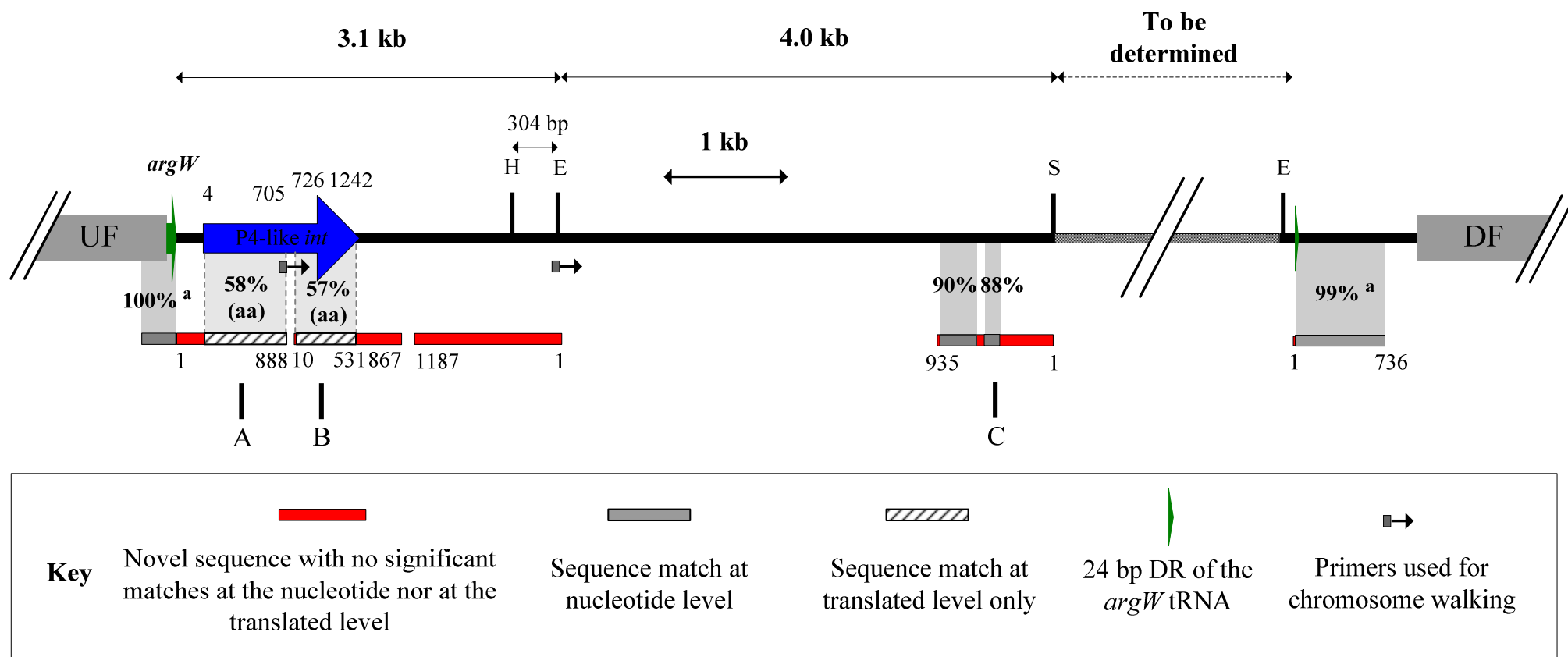


Figure 6.28. The S116 (*S. boydii* 1) *argW* associated novel GI.

H, E and S represent *Hind*III, *Eco*RI and *Sac*I restriction sites respectively.

^a These values represent the nucleotide identity to the Sb227 chromosome.

A: 888 bp with no significant hits at the nucleotide level. 185-886 has the highest amino acid similarity (58%) to 4-705 of a P4-like integrase in *Salmonella enterica* subsp. *enterica* serovar Paratyphi A str. ATCC 9150 (GenBank accession number YP 153333).

B: 867 bp with no significant hits at nucleotide level. 10-531 has 57% amino acid similarity to 726-1242 of a P4-like integrase in *Salmonella enterica* subsp. *enterica* serovar Paratyphi A str. ATCC 9150. However the same region has the highest similarity to P4-integrase-like elements in *Photobacterium luminescens* (60%) and various uropathogenic *E. coli* strains (59%). 532-867 has no significant hits at the nucleotide or at the translated level

C: 1-441 has no significant hits at nucleotide or at the translated level. 442-566 and 630-926 have 88% and 90% nucleotide ID respectively to phage-like DNA 5.6 kb into the 14.1 kb *argW* associated P4 prophage-like GI in the *E. coli* O157:H7 EDL933 and Sakai strains only. This sequence has no other significant matches and its function is unknown.

7.0 tRNA loci harbouring GIs that contribute to the virulence of *Shigella*

The results of the characterisation of tRNA loci that harbour GIs that I believe are playing a significant role in driving the virulence of *Shigella* are described in detail this section.

tRNA loci harbouring GIs that have been previously well characterised or less significant elements are described in section A2.6.

7.1 *leuX*

Table 7.1. SGSP-PCR results of *leuX* tRIP negative strain-tRNA loci

<i>leuX</i> U# - T7# SGSP-PCR results									
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)							Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	<i>Eco</i> RV	<i>Hinc</i> II	
K12 MG1655									
<i>S. dysenteriae</i> 3	S101	N ^a	N	N	N	N	~0.3	~1.0	710 [SK#]
<i>S. dysenteriae</i> 9	S102	N	N	N	N	N	~0.3	~1.0	431 [SK#]
<i>S. dysenteriae</i> 6	S103	N	N	N	~1.6 ^b	N			82 [SK#]
<i>S. flexneri</i> 1a	S104	N	N	~1.5	N	N			375 [SK#]
<i>S. flexneri</i> 2a	S106	N	N	~1.5	N	N			
<i>S. flexneri</i> 2b	S107			~1.5					
<i>S. sonnei</i>	S108	N	N	~2.0	N	N			753 [SK#]
<i>S. flexneri</i> X	S111			~1.5	N				
<i>S. sonnei</i>	S113			~2.2					546 [SK#]
<i>S. sonnei</i> bio a	S114			~2					
<i>S. sonnei</i> bio g	S115			~2					586 [SK#]
<i>S. boydii</i> 1	S116	N	~1.5	N	N	~0.7 F ^c			678 [SK#]
<i>S. boydii</i> 2	S117		~1.5						
<i>S. boydii</i> 3	S118	N	N	~0.7F, ~1.0 F	~1.6	N			
<i>S. boydii</i> 4	S119		~1.5				~0.3	N	520 [SK#]
<i>S. boydii</i> 7	S120	N	N	N	~0.3 F, ~0.6, ~1.6, ~3.0 F	N			

<i>leuX</i> D# - T7# SGSP-PCR results									
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)							Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	<i>Eco</i> RV	<i>Hinc</i> II	
K12 MG1655					~0.8				
<i>S. dysenteriae</i> 3	S101	N	N	N	N	N	~0.7 F	N	
<i>S. dysenteriae</i> 9	S102	N	N	N	~3.0 F	N	~1.4	~0.7	505 [T7#]

^a Indicates that no amplicon was generated

^b Text highlighted in bold indicates that the amplicon was sequenced

^c The addition of 'F' after the text indicates that the amplicon was faint

7.1.1 *S. flexneri*

The *leuX* U-arm results confirmed that the *S. flexneri* 1a, 2a, 2b and X strains all harbour the sequence that is present in the U-arm of the Sf301 *leuX* associated 7.5 kb islet (*leuX*-IF2, see Figure 7.1). The prevalence of IS and transposase-like elements within the Sf301 islet suggest that this was previously a larger GI that has been disrupted and lost most of its island-specific DNA, with the only remnants being the prophage-like integrase gene and an ORF that encodes a putative transcriptional regulator associated with the fermentation of deoxyribose. After blastn analysis, this ORF was found to only have significant DNA homologs in CFT073 (harboured on the *aspV* associated GI); another uropathogenic *E. coli*, strain 536 (GenBank accession number AJ617685, harboured on pathogenicity island V, which is associated with the *pheV* tRNA locus) and in intestinal and Extraintestinal pathogenic *E. coli* as part of an operon involved in the fermentation of deoxyribose. This operon enables strains to catabolise DNA to use it as a food source; strains that harboured the operon were able to out compete strains without it in coculture experiments and it is believed to provide a fitness advantage to pathogenic *E. coli* strains (Bernier-Febreau *et al.*, 2004).

These results suggest that this prophage was likely to have been acquired prior to the divergence of *Shigella* from an ancestral *E. coli* strain.

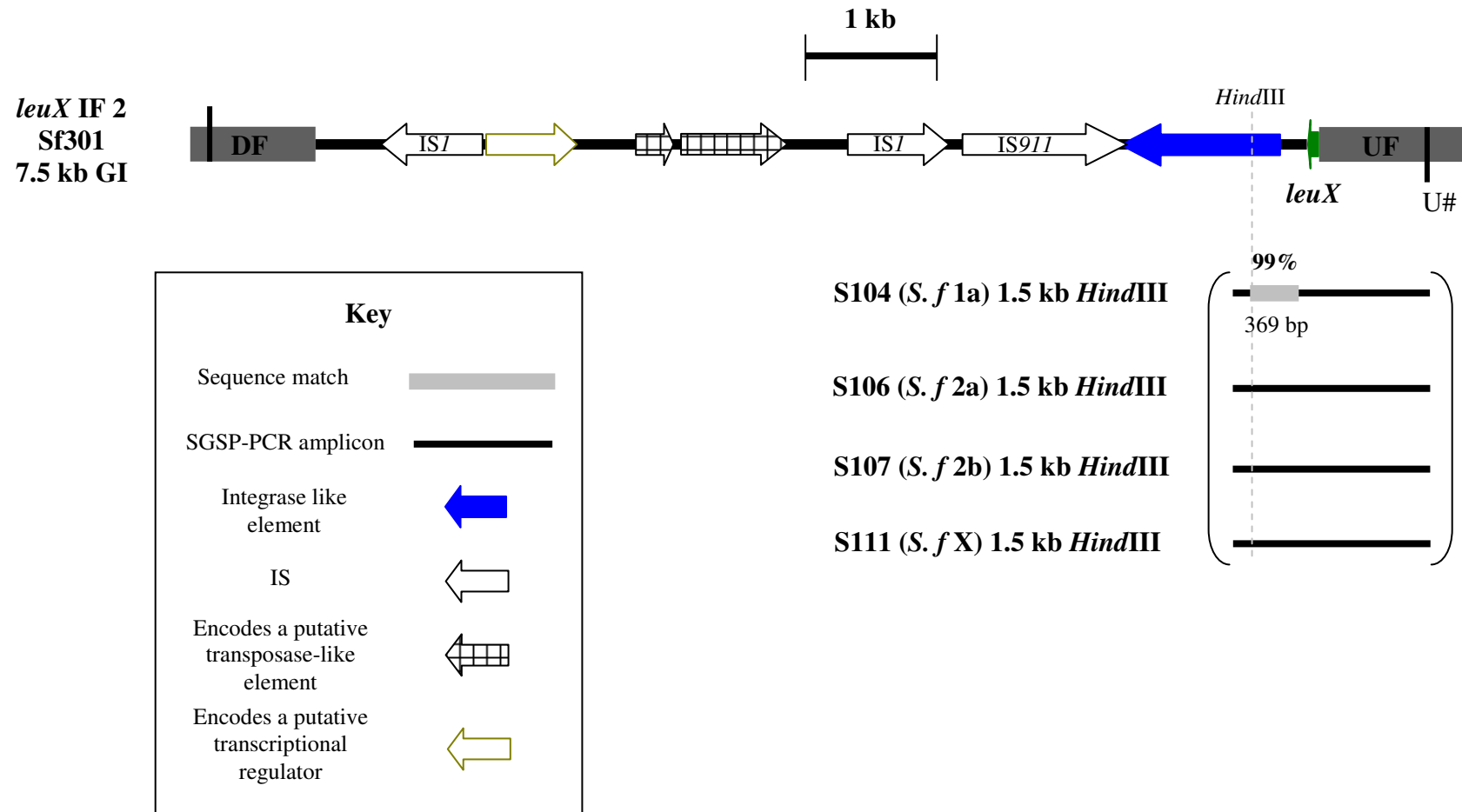


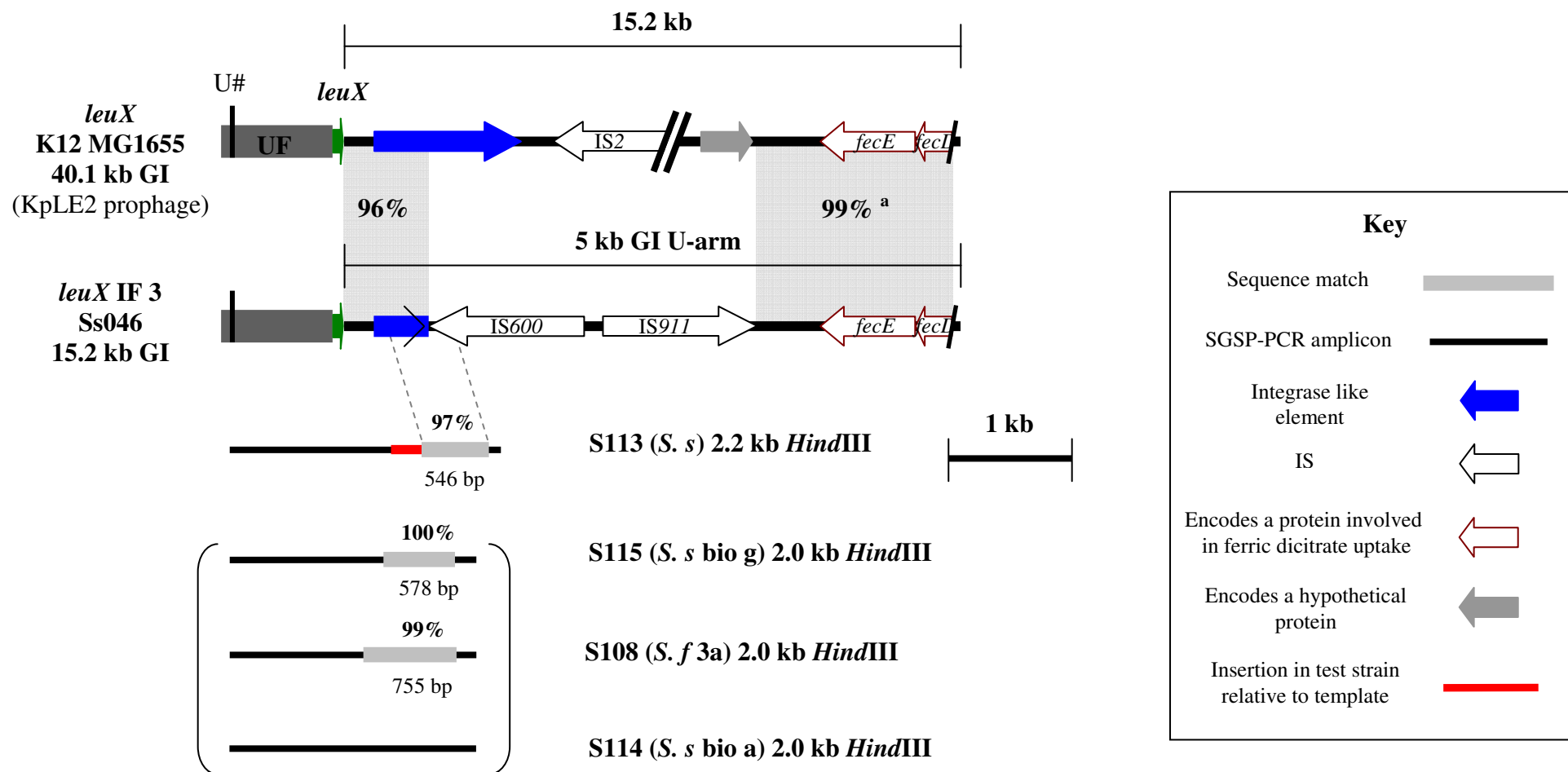
Figure 7.1. *leuX* U-arm SGSP-PCR results for the *S. flexneri* strains

7.1.2 *S. sonnei*

The U-arm results show that all of the *S. sonnei* strains harbour the same island DNA as is present in the U-arm of the *leuX* Ss046 15.2 kb GI (*leuX*-IF3). The presence of a truncated P4-like integrase gene, disrupted by an IS600 suggests that this island could be locked into the chromosome. In the centre of the Ss046 island is a cluster of genes known as the *fec* locus that encode a ferric dicitrate transport system (see Figure 7.2), this is a siderophore system that has been characterised previously in *E. coli* and *S. flexneri*. In *E. coli* K12 MG1655 it is present on the KpLE2 prophage, also associated with *leuX* (see Figure 7.2), whereas in *Shigella* it was discovered on the SRL PAI which is associated with the *serW* or *serX* loci (Luck *et al.*, 2001). However, a later study indicated that the *fec* locus is not always present on the SRL PAI, and as it is flanked by IS elements, it was suggested that the locus may be independently mobile (Turner *et al.*, 2003).

7.1.3 The *fec* locus is independently mobile

The above claim by Turner *et al* is therefore supported further in this study by the discovery of the *fec* locus on the *leuX* GI in Ss046, as it is also flanked by the same mobile genetic elements as are found in the SRL PAI (an IS911 at one end, as indicated in Figure 7.2 and an IS1 at the other). This provides more evidence that the *fec* locus is able to move within the chromosome of *Shigella* and that so far it has been found within *leuX* and *serW/serX* associated genomic islands.



^a 9.1 kb region comprising the *fecIRABCDE* genes that encode a ferric dicitrate uptake system.

Figure 7.2. *leuX* U-arm SGSP-PCR results for all of the *S. sonnei* strains

7.1.4 *leuX* island family 4 is a prophage restricted to some *S. boydii* strains

The U# results show that the *S. boydii* 1, 2 and 4 strains all harbour the same DNA as is present in the U-arm of the Sb227 *leuX* GI, designated in this study as *leuX*-IF4 (see Figure 7.3). The sequence walked into is the start of an intact P4-like integrase gene which in turn is found at the start of an 11.8 kb prophage-like element that is bound by a 19 bp imperfect DR of the *leuX* tRNA 3' end. Downstream of this is another prophage-like entity (see Figure 7.4), indicating that the Sb227 *leuX* GI is a mosaic made up of at least two tandem elements, however as the DF is inverted and displaced 182 kb upstream of the UF in this strain, the total size of the GI could not be determined by tRIP.

Interestingly before the Sb227 genome became available in November 2005, the *leuX* U# island sequences obtained from S116 and S119 had the highest nucleotide identity (94%) to P4-like integrase genes associated with *leuX* in the complete genomes of *Salmonella enterica* subsp. *enterica* serovar Typhi Ty2 (GenBank accession number AE014613), *Salmonella enterica* subsp. *enterica* serovar Typhi str. CT18 (GenBank accession number AL513382) and 93% to *Salmonella enterica* subsp. *enterica* serovar Paratyphi A str. ATCC 9150 (GenBank accession number CP000026). However they had no hits to the other two complete *Salmonella* genomes available (*Salmonella enterica* subsp. *enterica* serovar Choleraesuis str. SC-B67 and *Salmonella typhimurium* LT2) nor to any of the other fully or partially sequenced *E. coli* or *Shigella* genomes available on the NCBI database. These results indicate that this prophage-like element is found in a limited number of organisms, its presence in *S. boydii*, *S. typhi*, *S. paratyphi* and its absence in *S. typhimurium* suggests that it is restricted to strains that naturally cause disease in humans only. Its presence in so few strains and its context upstream of another prophage-like element also suggests that it is a relatively recently acquired entity.

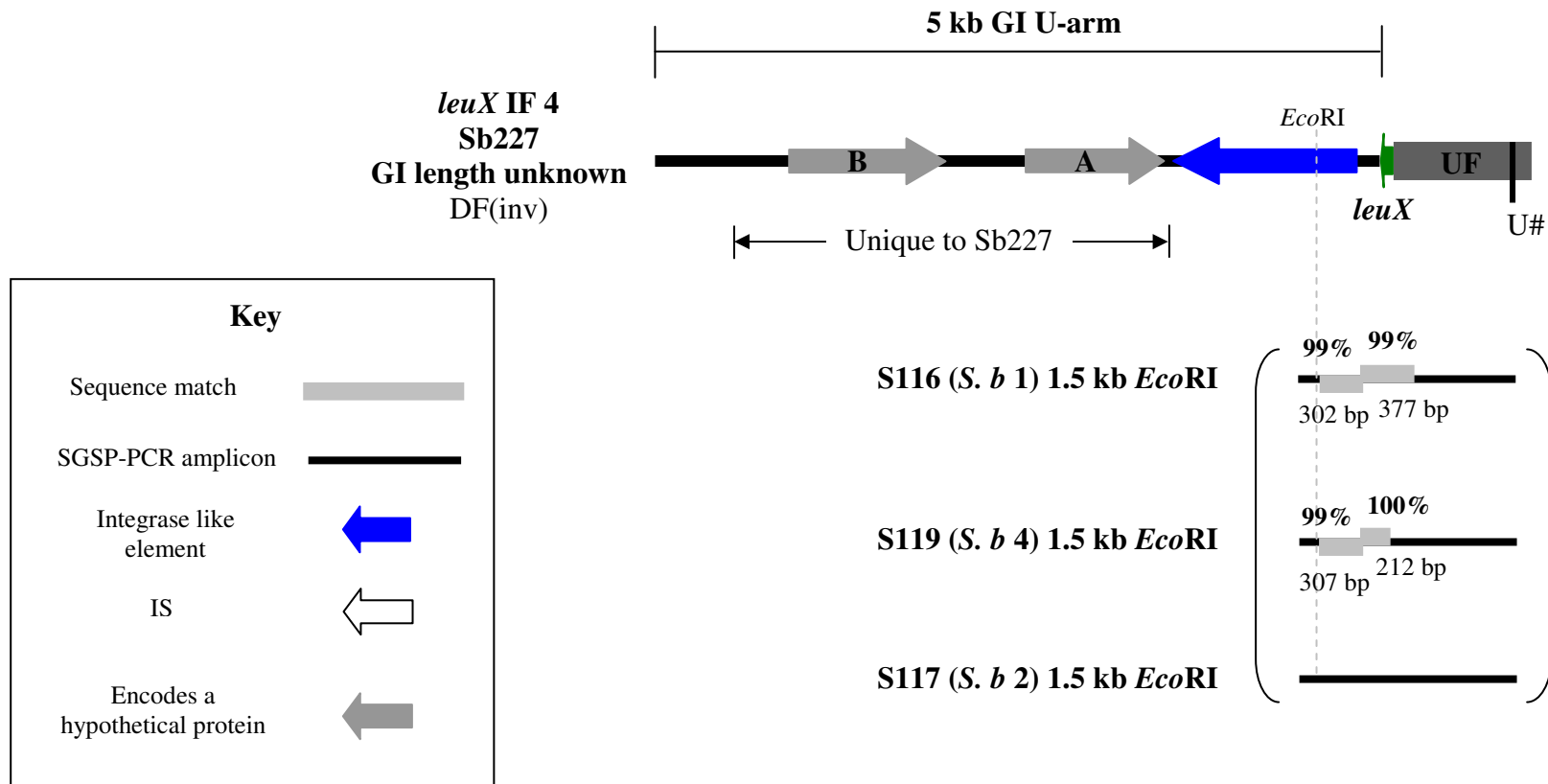


Figure 7.3. *leuX* U-arm SGSP-PCR results for *S. boydii* strains belonging to *leuX* island family 4

Further blastn analysis of the entire 11.8 kb Sb227 element showed that the 3.0 kb region downstream of the integrase gene, which contains two ORFs that encode hypothetical proteins is unique to Sb227 at the nucleotide level (see Figure 7.3). When translated, Blastx analysis indicated that the region had no hits in any of the other *E. coli* or *Shigella* genomes. The most significant match was across ORF 'A', which had only 37% amino acid similarity to a hypothetical protein in *Reinekea* sp. MED297 (GenBank accession number EAR10744), a novel gammaproteobacterium isolated from coastal sediments in the Sea of Japan in Russia. The novel nature of this DNA sequence, along with its relatively low GC content (35.5%), even compared with the surrounding prophage DNA and its small size, suggests that the two ORFs are 'ORFan'-like (Daubin and Ochman, 2004); which indicates that this DNA was acquired by HGT from an as yet uncharacterised organism, is most likely bacteriophage derived and/or could be a rapidly evolving region of the chromosome. The remainder of the element downstream of this, had 97% nucleotide identity to the early genes on the bacteriophage P4 complete sequence (GenBank accession number X51522) and to a P4-like prophage element found associated with the *ssrA* locus in *S. typhi* Ty2.

7.1.5 *S. boydii* and *S. dysenteriae* strains that harbour *leuX* island family

1

The *S. boydii* 3 and 7 strains and the *S. dysenteriae* 3, 9 and 6 strains were different to the above mentioned *S. boydii* strains, in that they did not harbour DNA similar to the 11.8 kb Sb227 P4-like island described above. Instead, directly downstream of *leuX* they all harboured DNA with the highest nucleotide identity to the sequence found at the start of the prophage-like element found distal to the 11.8 kb island in Sb227, designated in this study as *leuX*-IF1 (see Figure 7.4). As mentioned earlier, because the DF is inverted and displaced upstream of the UF in Sb227, the size of this GI could not be calculated by tRIP.

The Sb227 distal element also harbours a P4-like integrase gene; with nucleotide identity to other P4-like integrase genes found associated with *leuX* island DNA in *E. coli* and *Shigella* (see Table 7.2).

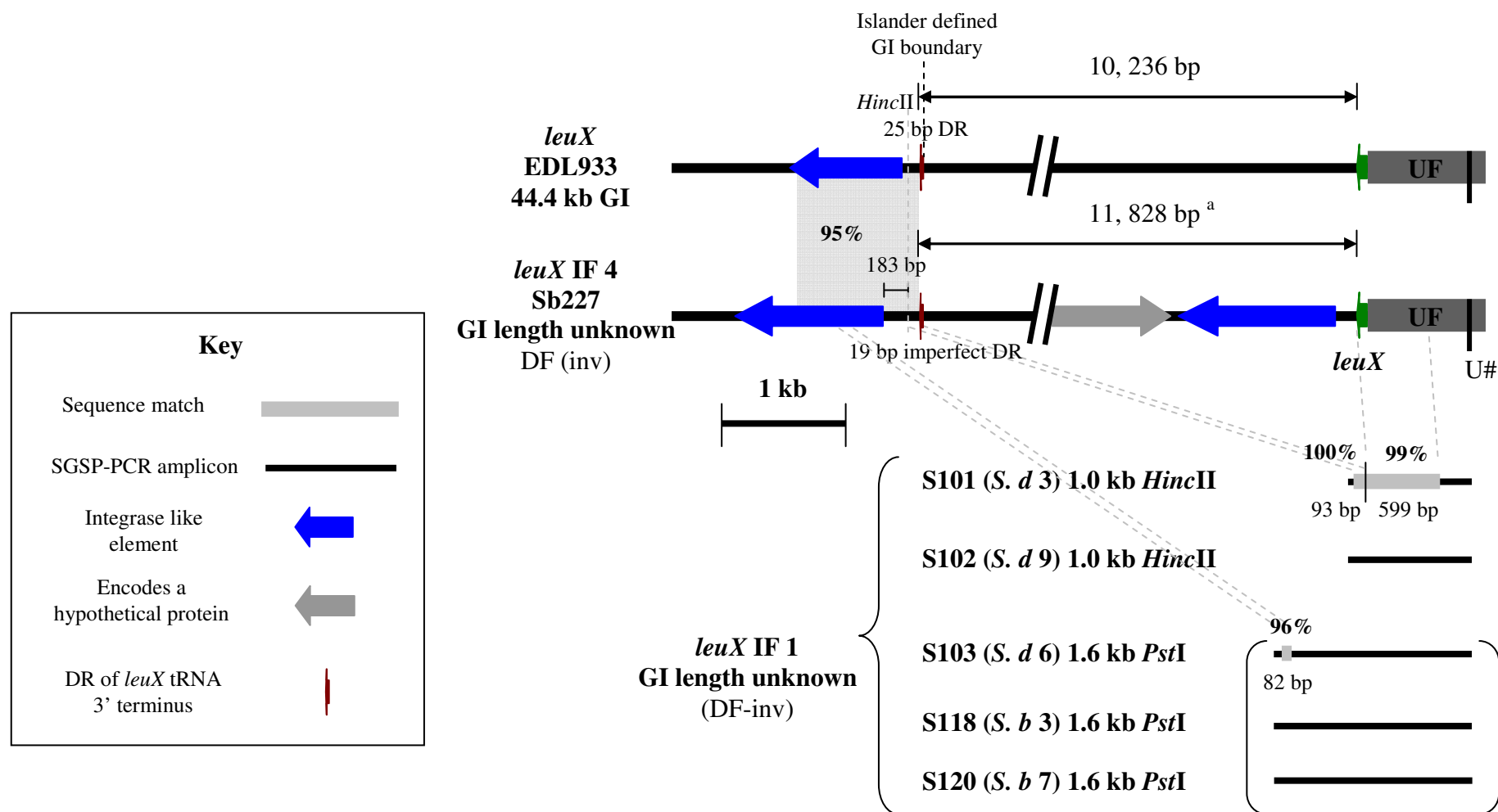


Figure 7.4. *leuX* U-arm SGSP-PCR results for the *S. boydii* and *S. dysenteriae* strains belonging to *leuX* island family 1

^a See Figure 7.3 for more details on the 11.8 kb element found immediately downstream of *leuX* in Sb227.

Table 7.2. Blastn comparison of the Sb227 P4-like integrase gene found 12.1 kb into the *leuX* associated GI, against the other complete *E. coli* and *Shigella* genomes.

Accession	Description	Max score	Tot. score	Query coverage	E-value	Max. identity
AE014075.1	<i>Escherichia coli</i> CFT073, complete genome	1858	2952	98%	0.0	95%
BA000007.2	<i>Escherichia coli</i> O157:H7 str. Sakai DNA, complete genome	1110	1367	98%	0.0	94%
AE005174.2	<i>Escherichia coli</i> O157:H7 EDL933, complete genome	1110	1367	98%	0.0	94%
U00096.2	<i>Escherichia coli</i> K12 MG1655, complete genome	753	753	80%	0.0	77%
AE005674.1	<i>Shigella flexneri</i> 2a str. 301, complete genome	549	1179	96%	9e-153	78%
AE014073.1	<i>Shigella flexneri</i> 2a str. 2457T, complete genome	549	1182	96%	9e-153	78%

Table 7.2 indicates that the Sb227 *leuX*-IF1 integrase gene has the most significant blastn score to CFT073, where the cognate *int* gene is also found directly downstream of *leuX* (see Figure 7.5).

These results indicate that the Sb227 distal integrase gene is more *E. coli*-like than *Shigella*-like, suggesting that Sb227 and the *S. boydii* and *S. dysenteriae* strains in Figure 7.4 may have acquired the element prior to *E. coli*'s divergence into the different pathotypes, and that they are of a distinct lineage to *S. flexneri* and *S. sonnei*.

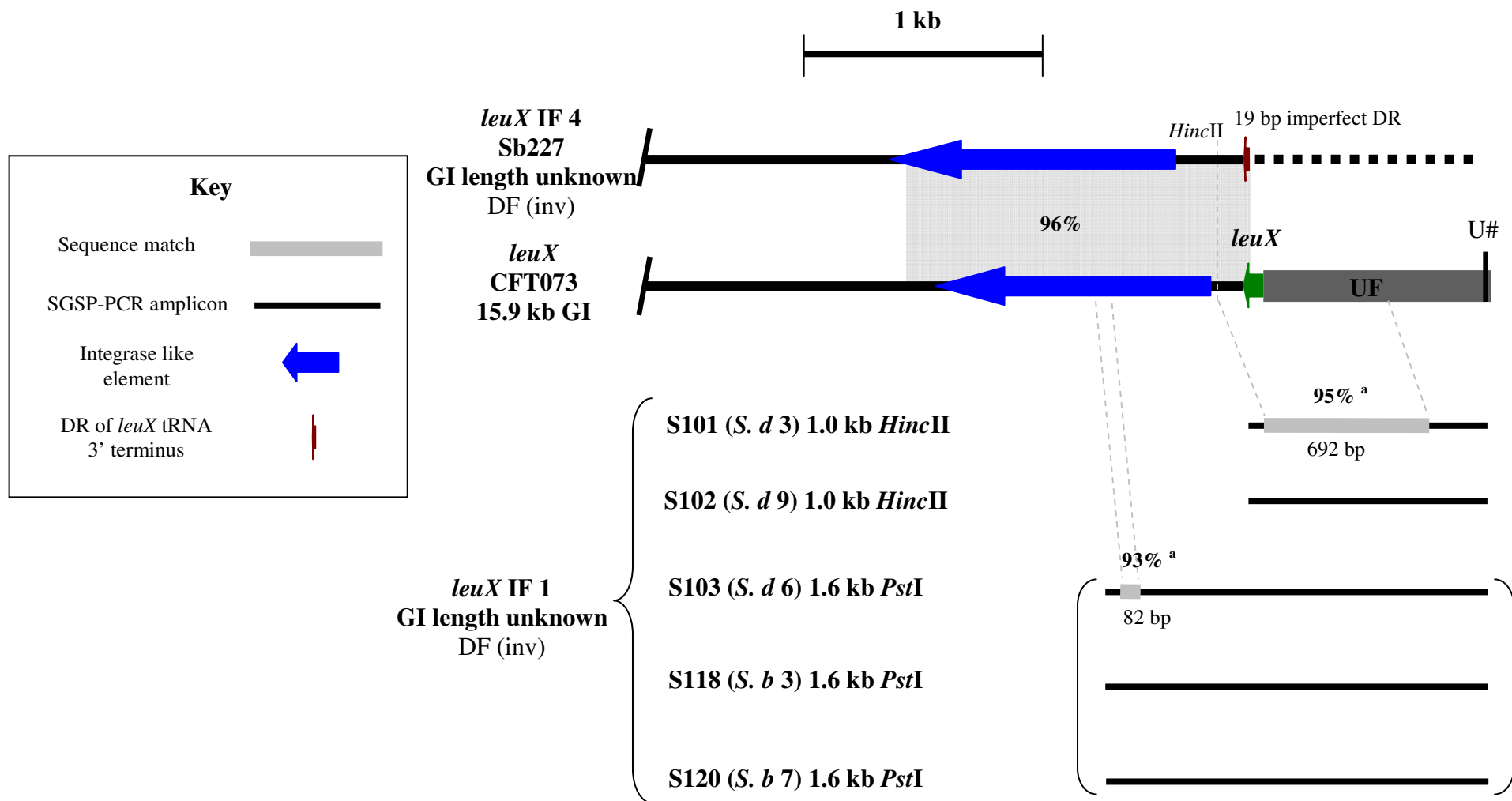


Figure 7.5. Sequence results of the *leuX* island family 1-like strains compared with the CFT073 *leuX* associated GI.

7.1.6 Diversity of other *leuX* associated GIs

In EDL933 (and Sakai), as in Sb227, the cognate island family 1 integrase gene is present at the start of a 34.2 kb prophage-like element, which in turn is found distal another 10.2 kb prophage-like element that is bound by a 25 bp DR of the *leuX* tRNA 3' end (see Figure 7.4). Blastn analysis of the 34.2 kb and 10.2 kb tandem GIs in EDL933 showed that apart from the integrase genes and one IS629, the rest of the sequence is only present in EDL933 and Sakai. Also the 15.9 kb CFT073 GI contains 50% strain-specific DNA, with only the integrase gene and one IS630 having significant nucleotide identity to any of the other *E. coli* or *Shigella* sequences available. This indicates that the *leuX* associated GIs are extremely diverse in their content across both *E. coli* and *Shigella*.

7.2 A *sigA* gene is harboured on the Sb227 distal *leuX* GI

The Sb227 *leuX* distal GI (*leuX*-IF1) is very interesting, as 14.7 kb downstream of the integrase gene is the *sigA* gene that encodes a temperature-regulated serine protease (SigA), which is a member of the serine protease autotransporters of Enterobacteriaceae (SPATE) family of autotransporters; it also has cytopathic and enterotoxic activity and is a major virulence factor of *Shigella*. This gene and its functions were characterised by (Al-Hasani *et al.*, 2000) where it was found present on the *she* PAI; that was first found associated with the *pheV* locus in *S. flexneri* (Rajakumar *et al.*, 1997), structural variants of the *she* PAI have since been found present in other members of *Shigella* (Al-Hasani *et al.*, 2001a) (see Table 1.3 for more details on the *she* PAI). Al-Hasani *et al.*, 2000 found that the SigA protein does not need any additional *Shigella* genes for its secretion from the cell and suggested that it is autonomously secreted. This shows that the *sigA* gene itself, if introduced into the chromosome of a *Shigella* strain could markedly enhance its virulence.

7.2.1 The *sigA* gene may be independently mobile

The *sigA* gene has not been previously reported to be present on a *leuX* associated GI, however it has been reported that there are two copies of the *sigA* gene on the Sb227 chromosome (Yang *et al.*, 2005); however this strain does not harbour the *she* PAI. These results indicate that the *sigA* gene is independently mobile. This claim is supported by the fact that blastn analysis indicated that the *sigA* gene and a small ORF directly upstream are flanked by IS elements, which are also found present in the same context in other *Shigella* strains (see Figure 7.6). In Sb227 the second *sigA* gene is found present 22.0 kb downstream of the *aspV* locus, which in turn using *in silico* tRIP was found to be occupied with a 1.0 kb islet. Therefore by definition, in this study, the second Sb227 *sigA* gene is not part of the *aspV* associated island DNA.

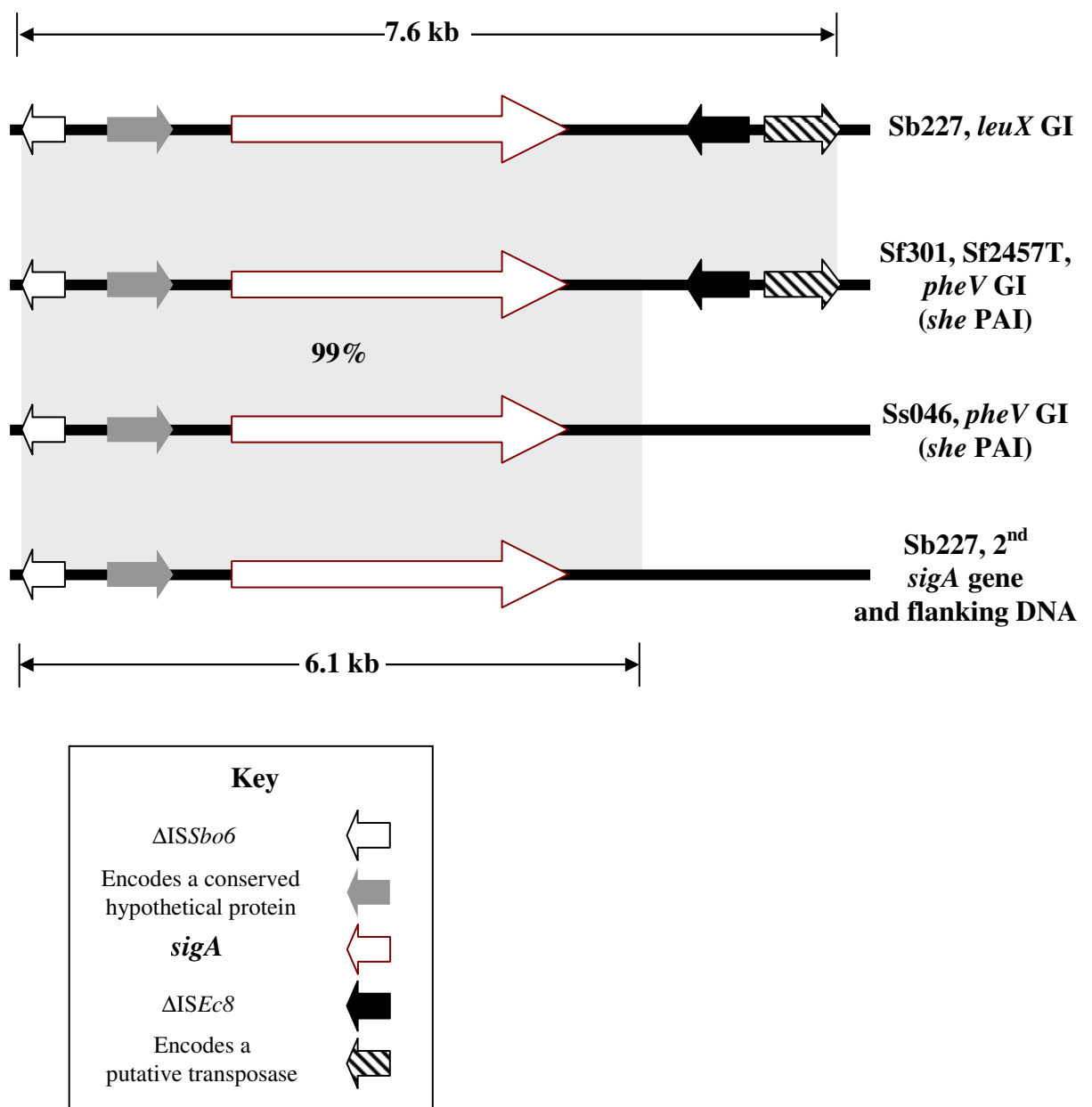


Figure 7.6. Schematic showing the location and context of the *sigA* region in the completed *Shigella* genomes available in the NCBI database.

The grey shaded region indicates sequences with 99% nucleotide identity. Figure is not to scale.

The results of this study show that the *leuX* tRNA locus is a hot spot for P4 prophage like GIs and other mobile elements across *E. coli* and *Shigella* and that there is significant variation between strains in the content of these elements and a high percentage of strain-specific DNA. This observation has also been reported recently in an excellent study by (Bishop *et al.*, 2005) on the *leuX* tRNA in *S. enterica*, which shows that P4 phages are drivers of diversity at this locus across *Salmonella* and *E. coli*.

7.3 S101 *leuX* island probing

As the *leuX* site is such a hot region for island activity, I decided to probe the S101 (*S. dysenteriae* 3) *leuX* island in more detail. As mentioned above, *sigA* was originally found associated with the *she* PAI which is found at the *pheV* locus. Previous studies have shown that S101 harbours the *sigA* gene, however, it does not harbour the *she* PAI at *pheV*, or any other *she* PAI markers apart from a P4-like integrase gene which was found to be associated with another *phe* tRNA (Al-Hasani *et al.*, 2001). The results of this study indicate that in S101 an island with a similar P4-like integrase gene to the *she*-PAI *int* is found associated with *pheU* (see section A2.6.4), however this may not be a *she*-PAI - like island, but could be a SHI-3 - like element, which has a similar P4-like *int* and has been found associated with *pheU* in *S. boydii* (Purdy and Payne, 2001). Given the similarity in tRIP profiles and characterised island content across the *S. boydii* and *S. dysenteriae* strains screened in this study at *leuX* and at other tRNA loci (see Table 5.1) to each other and to Sb227, and the absence of other *she*-PAI markers across these strains (Al-Hasani *et al.*, 2001), I hypothesised that the *sigA* gene may be harboured on the *leuX* GI in these strains. I chose S101 as the first strain to probe because it was also multidrug resistant due to it harbouring an SRL PAI-like element and therefore has a greater pathogenic potential and is an interesting strain to probe.

The method used for the S116 *argW* allelic exchange was also used in this case, however as S101 was already Ap^r, Cm^r, Sm^r and Tc^r, any of these resistance phenotypes could be used

along with the Km^r conferred by the inserted cassette to select for transconjugant S101 derivatives after the conjugation with the *E. coli* SM10 λ *pir* donor strain, so there was no need to engineer a resistant derivative of S101 as with the S116 island probing work.

7.3.1 *leuX* UF splicing overlap extension (SOE) PCR

A 1.0 kb region of the Sf301 *leuX* UF was selected for allelic exchange, it had at least 95% nucleotide identity to the *E. coli* and *Shigella* genomes available on the NCBI, so was very likely to be well conserved in S101. As with the *argW* experiments, I aimed to amplify this region using PCR with primers that incorporated *Xba*I restriction sites into the ends of the amplicon, so that the product could be cloned and manipulated. However as the *leuX* UF does not contain an *Nsi*I site, this had to be incorporated into the centre of the UF region, so that the *Nsi*I flanked Km^r cassette could be subsequently ligated into it. To do this, a technique known as splicing overlap extension (SOE) PCR was used, which was originally published by Horton *et al.* in 1989, (see section 2.8.5 for the modified method used).

Firstly, two separate standard PCR reactions were performed with primers designed to incorporate flanking *Xba*I and *Nsi*I sites into each of the amplicons; *leuX* UFF# and *leuX* UFIF# ('I' stands for internal) were used to generate a 507 bp product and *leuX* UFR and *leuX* UFIR were used to generate a 571 bp product (see Figure 7.8). The two internal primers were designed to be the reverse complement of each other with the *Nsi*I site placed in the centre (see Figure 7.7) and Table A2. 2).

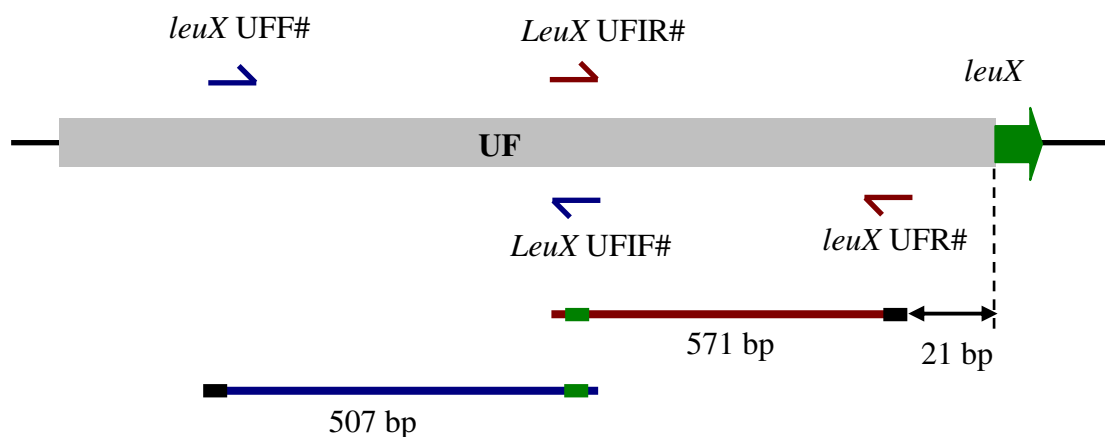
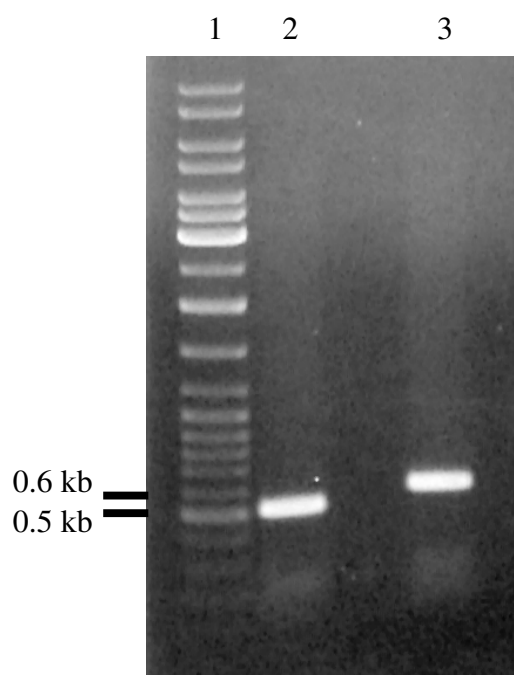


Figure 7.7. The Sf301 *leuX* UF, showing the positions of the primers used to amplify the region used for homologous recombination with the S101 *leuX* UF.

The thick blue and red lines indicate the expected PCR amplicons with the flanking *Xba*I and *Nsi*I sites shown as black and green boxes respectively. Figure is not to scale.



Lane number:

1. GeneRuler™ 1 kb ladder (Fermentas)
2. *leuX* UFF- *leuX* UFIF generated PCR amplicon
3. *leuX* UFR - *leuX* UFIR generated PCR amplicon

Figure 7.8. Agarose gel showing the two first round PCR amplicons used to generate the *leuX* UF SOE PCR product.

The remainder of each PCR was electrophoresed, gel extracted, cleaned and quantified. 100 ng of each of the products was then added as the template in a second round SOE PCR using the *leuX* UFF and *leuX* UFR primers. The ends of each amplicon where the *Nsi*I sites are incorporated are the reverse complement of each other, so they can anneal after the DNA is denatured to form a fusion product with a single *Nsi*I site in the centre. This product is then amplified by the external F and R primers (see Figure 7.9 and Figure 7.10)

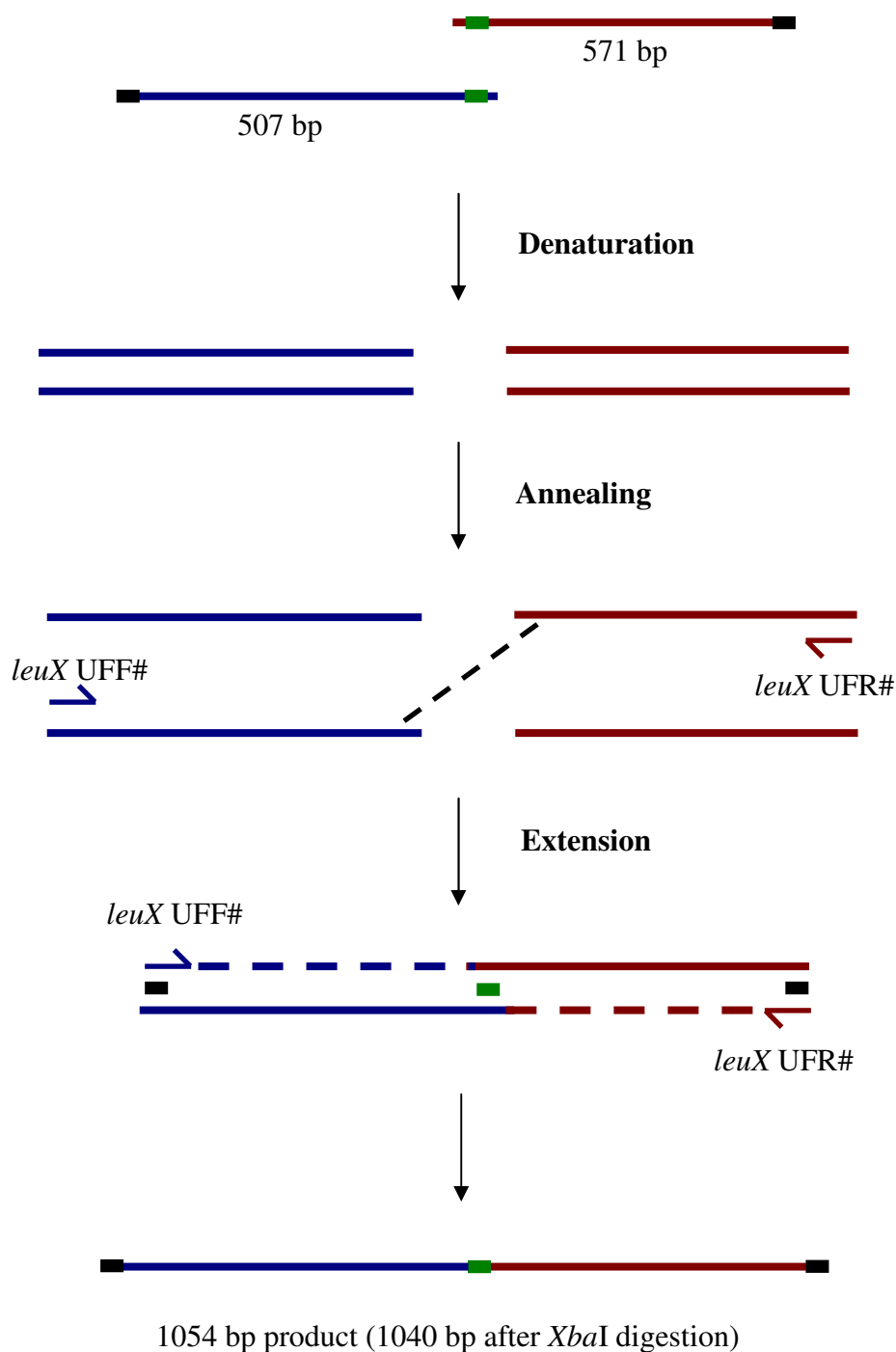
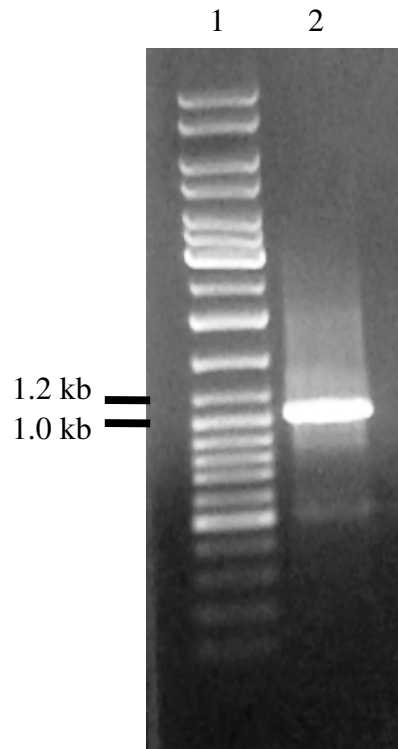


Figure 7.9. Schematic showing the principle of SOE PCR and the generation of the modified *leuX* UF region containing a central *NsiI* site and flanked by *XbaI* sites.

The *XbaI* and *NsiI* sites are shown as black and green boxes respectively. Figure is not to scale.



Lane number:

1. GeneRuler™ 1 kb ladder (Fermentas)
2. *leuX* UFF- *leuX* UFR generated SOE PCR amplicon

Figure 7.10. Agarose gel showing the 1054 bp SOE PCR product.

The fusion product was gel extracted, cleaned and 500 ng was digested with *Xba*I and ligated into pBluescript/*Xba*I to produce pJL1 (see Figure 7.11 (a)). The ligation was electroporated into *E. coli* DH5 α (see 2.14) and the cells plated onto LA containing 100 μ g/ml Ap and 40 μ g/ml X-gal, to screen for ampicillin resistant white colonies that contained the insert. Potential pJL1 candidates were checked by digestion with *Xba*I and *Nsi*I (see Figure 7.11 (b)).

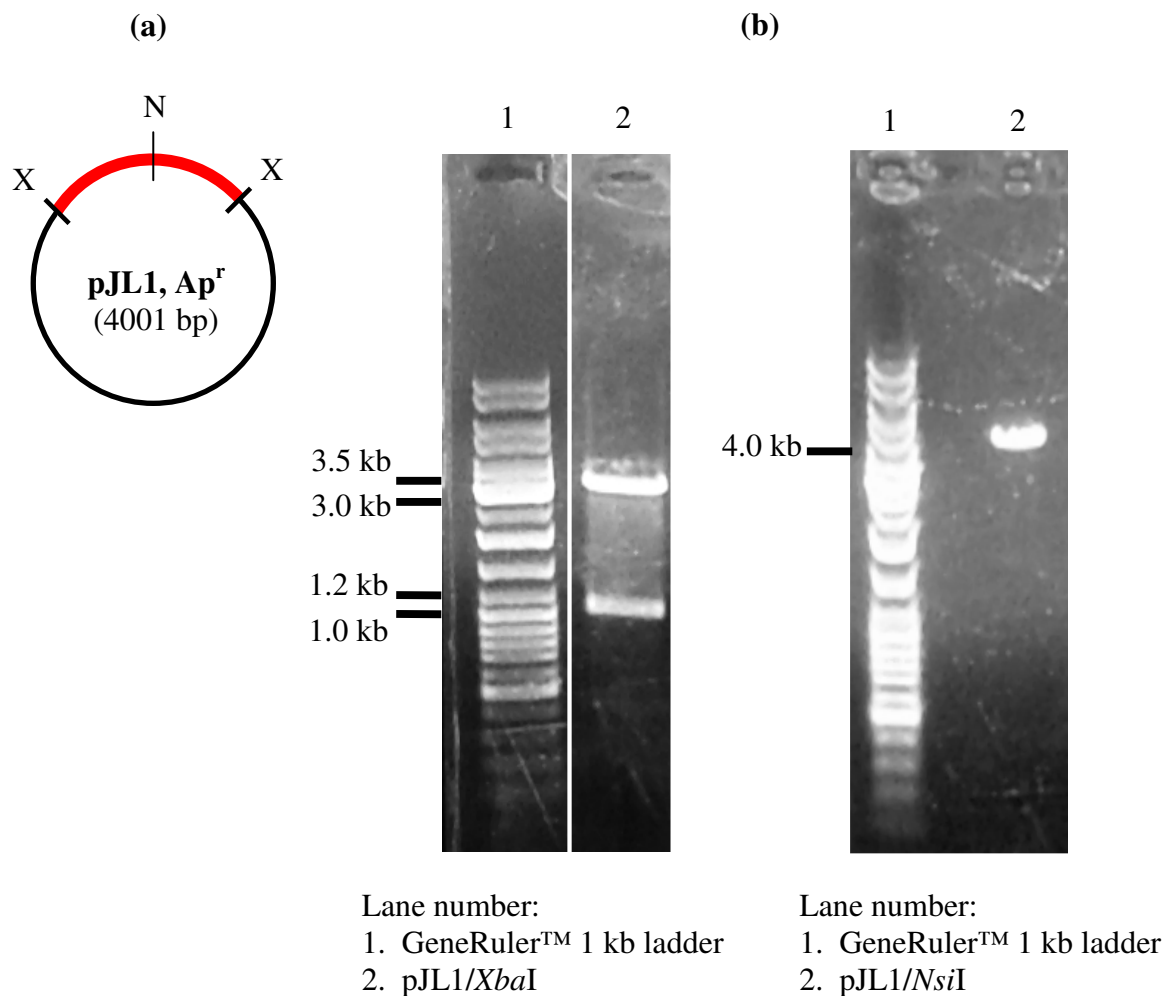


Figure 7.11. pJL1

(a). pJL1, the black segment represents pBluescript and the red segment represents the *leuX* UF region. X and N stand for *Xba*I and *Nsi*I sites respectively, figure is not to scale.

(b). Agarose gels showing pJL1 digested with *Xba*I and *Nsi*I respectively. This confirmed that the insert was the correct size and that the restriction sites still existed.

The subsequent generation of the *leuX* UF suicide constructs was performed in the same way as the corresponding *argW* UF suicide constructs (see section 6.3.5 to section 6.3.6). The *leuX* UF suicide constructs generated were named pJL3 and pJL4 and were electroporated into *E.*

coli SM10 λ *pir* to generate strains KR238 and KR239 respectively, the details of their construction are in section A2.5.

7.3.2 S101 allelic exchange

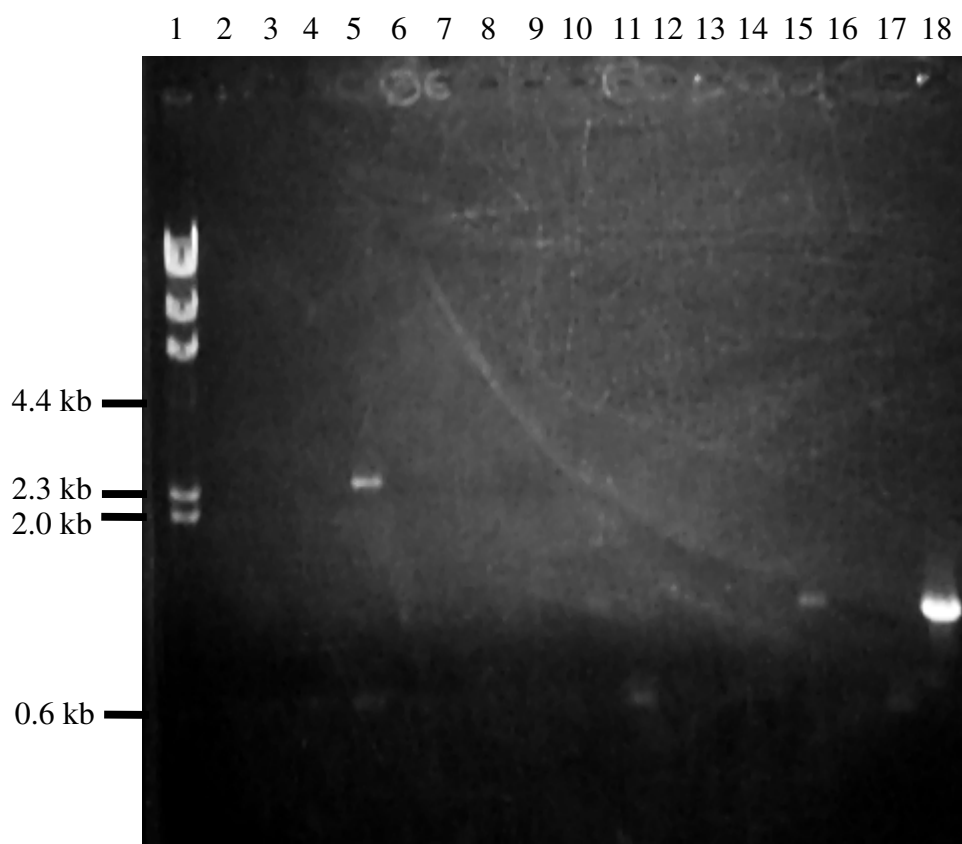
S101 was used in conjugations with KR238 and KR239 (see section 2.16 for methodology). Transconjugant S101 *Shigella* clones were selected for on LA plus 50 μ g/ml Km and 10 μ g/ml Tc (S101 is resistant to Tc, but SM10 λ *pir* is susceptible so was killed), then streaked onto LA plus 40 μ g/ml X-gal to confirm they were *Shigella* bacteria (see section 6.3.7).

S101 transconjugants were obtained from conjugations with both pJL3 and pJL4, representatives of each were stored as frozen stocks and these were named KR219 and KR220 respectively. Each of these was used in a sucrose selection step to select for transconjugants that had undergone a double crossover event (see 2.16.2 for methodology). Sucrose resistant clones were obtained from both experiments, but when streaked onto LA plus 50 μ g/ml Km and 10 μ g/ml Tc to confirm they had retained the Km^r cassette, 30% of the pJL4 transconjugants grew on this medium, but none of the pJL3 transconjugants grew.

As in the *argW* experiments, the sucrose resistant, Km resistant, S101 transconjugants were screened with primers that were positioned external to the allelic exchange region, to check for the double crossover event and the presence of the Km^r cassette in the *leuX* UF. The primers were designated *leuXU*#2 which is positioned upstream in the *leuX* UF, and *leuX* tRNA^{rev} (see Table A2. 2), which is positioned on the complementary strand in the *leuX* tRNA gene.

In silico PCR with Sf301 yields a product of 1153 bp when the Km^r cassette is absent. If the Km^r cassette was present, an amplicon of 2500 bp would be produced.

Colony PCR on 15 potential transconjugants derived from the conjugations with pJL4 were performed (standard, touch-down PCR, extension time 2.5 min) (see Figure 7.12).



Lane number:

1. λ /HindIII ladder
5. S101 transconjugant with Km^r cassette present in the *leuX* UF
9. Lane empty
18. S101 parent strain positive control

Figure 7.12. Agarose gel showing the results of the colony PCRs on the S101 potential transconjugants.

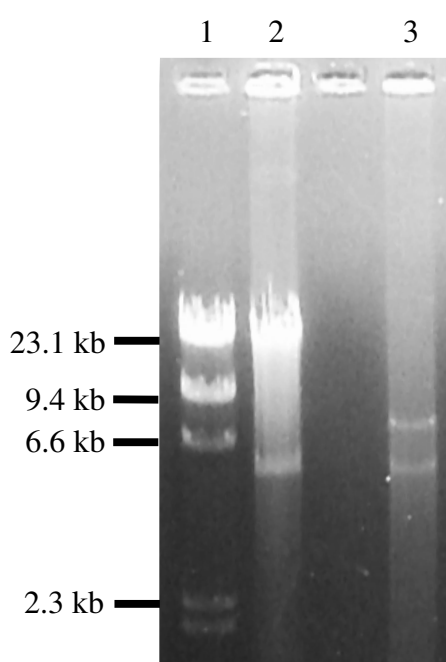
Lane 5 shows that this clone has undergone a double crossover and retained the Km^r cassette in the *leuX* UF region. Lane 18 shows the S101 positive control with no Km^r cassette present. All of the other lanes were likely to have only undergone a single crossover event. The transconjugant shown in lane 5 was named X101 and stored as a frozen glycerol stock (see Table 2.3).

7.4 X101 marker rescue

To capture portions of the S101 GI, genomic DNA was extracted from the kanamycin resistant derivative X101 and used to generate restriction libraries in pWSK29, as with the X106 *argW* GI (see 6.4). The enzymes selected to generate restriction libraries were:

EcoRI, *EcoRV*, *HindIII* and *BamHI*.

The *HindIII* and *BamHI* libraries produced Km^r and Ap^r clones, representatives had their plasmid DNA extracted and were digested to size their inserts (see Figure 7.13)



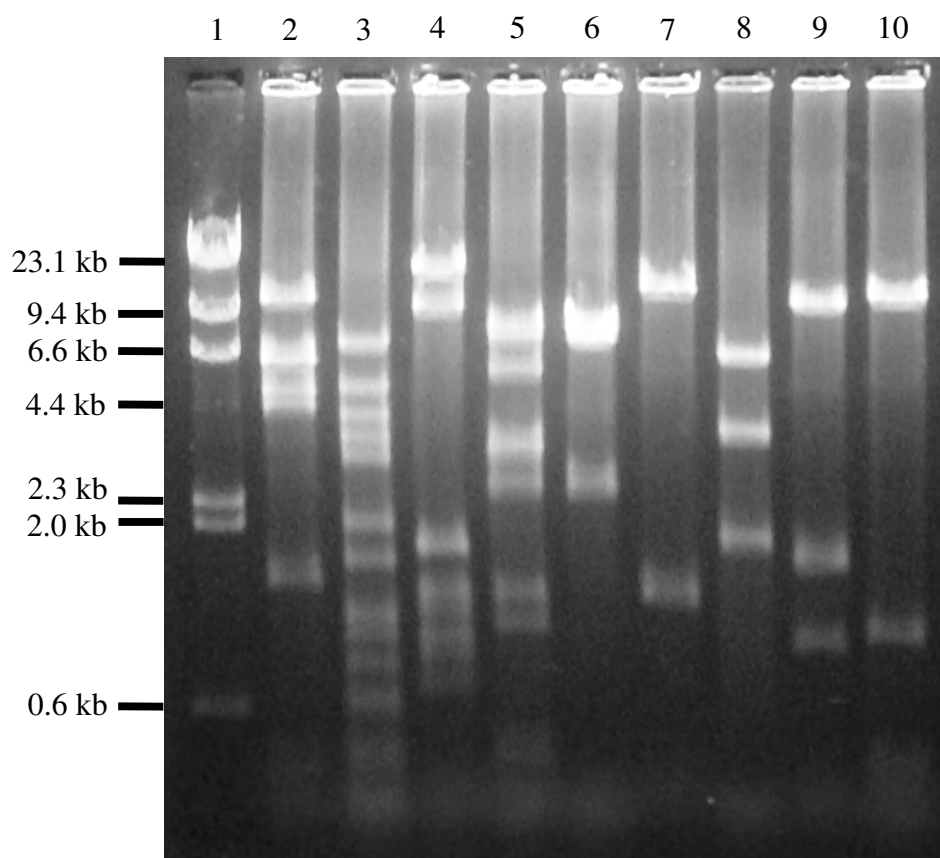
Lane number:

1. λ /HindIII ladder
2. *BamHI* marker rescue clone/*BamHI* (pJL10).....Insert of around 23.0 kb
3. *HindIII* marker rescue clone/*HindIII* (pJL11).....Insert of around 7.0 kb

Figure 7.13. Agarose gel showing the sizes of the insert fragments harboured by each of the X101 *leuX* marker rescue clones pJL10 and pJL11.

The common band is pWSK29 at 5.4 kb.

The clones were also digested with other enzymes to map and size them more accurately (see Figure 7.14).



Lane number:

1. λ /*Hind*III ladder
2. pJL10/*Nsi*I
3. pJL10/*Hinc*II
4. pJL10/*Pst*I
5. pJL10/*Eco*RV
6. pJL10/*Hind*III
7. pJL11/*Nsi*I
8. pJL11/*Hinc*II
9. pJL11/*Pst*I
10. pJL11/*Eco*RV

Figure 7.14. Agarose gel showing the restriction patterns of the X101 *Bam*HI and *Hind*III marker rescue clones pJL10 and pJL11.

The lengths of the restriction fragments were determined accurately using semi-logarithmic graph paper and the ID version 3.5.0 gel documentation package (Kodak). From these the sizes of the inserts were calculated to be:

pJL10 insert – 24.6 kb.

pJL11 insert – 7.0 kb.

The plasmids were end sequenced from T7 and T3 to map the sequence surrounding the Km^r cassette.

Figure 7.15 shows the sequence data obtained.

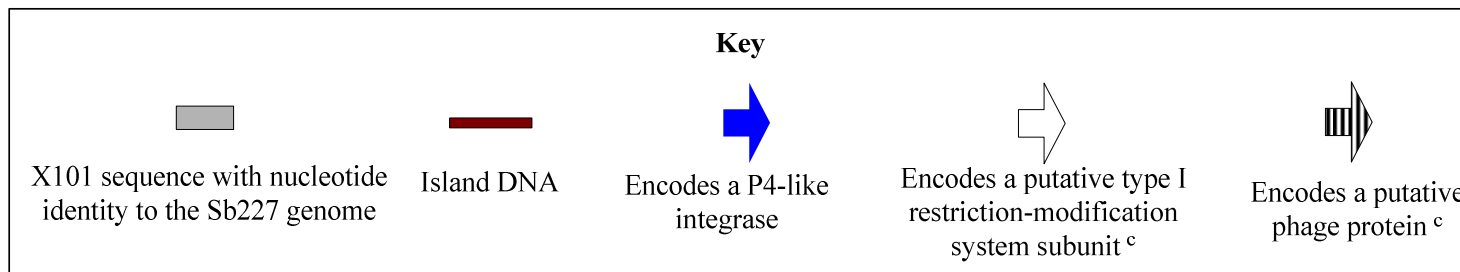
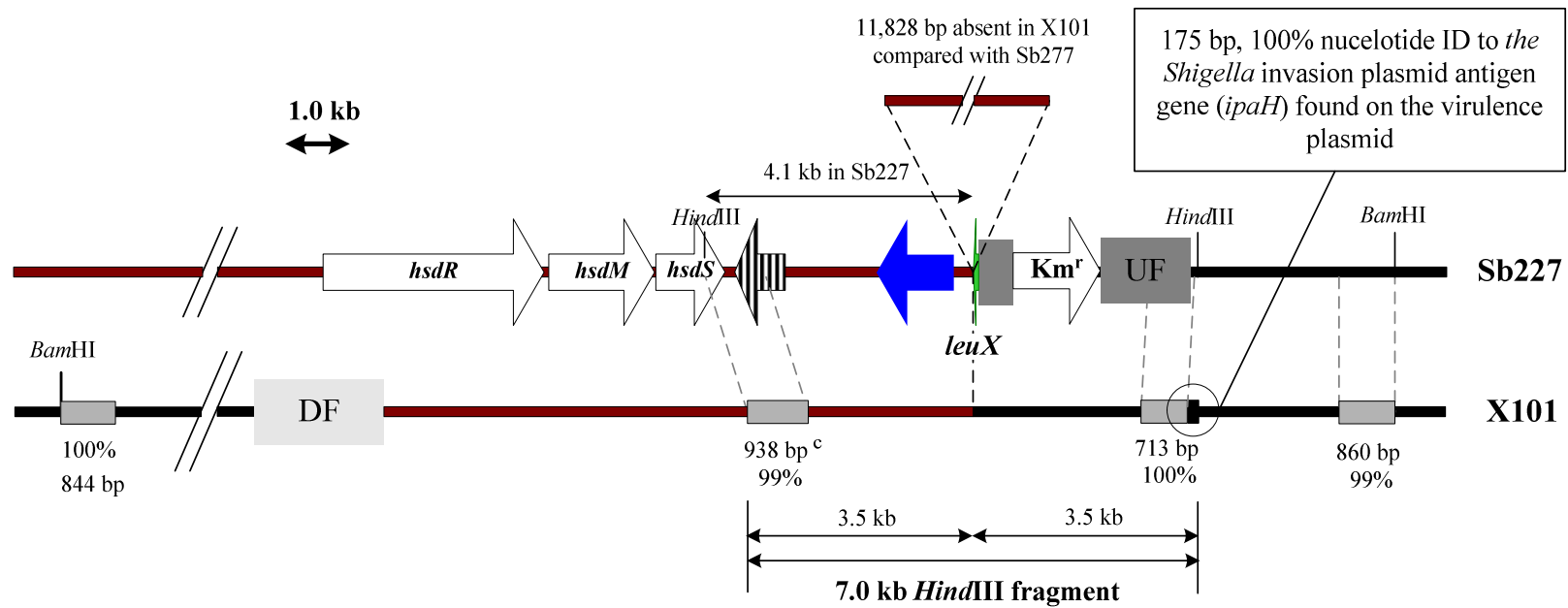


Figure 7.15. The S101 (*S. dysenteriae* 3 strain) Km^r derivative X101, marker rescue clone end sequence analysis.

Compared with the Sb227 chromosome to determine the size of the X101 *leuX* associated GI. Only selected restriction sites are shown

^a The sequence acquired was found to be 9.0 kb downstream of the start of the conserved DF. Subsequent Blastn analysis indicated that this sequence lies within a 9.1 kb region that comprises the DF, which in turn is well conserved and present in all of the other complete *E. coli* and *Shigella* genomes. In Sb227 the entire DF and distal DNA is inverted and found translocated 190 kb upstream of *leuX*, however, the sequence data from the X101 *Bam*HI marker rescue clone proves that in X101 the region is present downstream of the island and in the orientation matching most other strains. Therefore, in X101 the DF is likely to be downstream of the GI.

^b The GI lies on a 24.6 kb *Bam*HI fragment, in the sequenced *E. coli* and *Shigella* genomes the core regions flanking *leuX* are similar in length, therefore as the DF is likely to be present as shown in the diagram, the size of the island between the UF and DF in X101 was calculated to be: $24.6 - (9.0 + 6.5) = 9.1$ kb

^c This sequence was found present only in the sequenced *S. boydii* strains Sb227 and the *S. boydii* 18 strain BS512 (unfinished, GenBank accession number AAKA000000000).

The results of the marker rescue and end sequencing indicated that the entire X101 island lies on the *Bam*HI marker rescue clone and as the DF is likely to be downstream of the island, unlike in Sb227 where it is inverted. This enabled me to calculate the length of the island to be 9.1 kb. The sequence data from within the island indicated the presence of some of the elements that are present in the U-arm of the Sb227 *leuX* flanking GI (*leuX*-IF1), which is found 12.1 kb downstream of *leuX*, distal to another prophage-like element (*leuX*-IF4) inserted at *leuX* (see Figure 7.4 also). The region sequenced includes part of an ORF which encodes a putative phage protein and part of an ORF which encodes a putative specificity subunit for a type I restriction-modification (R-M) system. When checked against the entire microbial database on the NCBI, it was found that these were previously only found present in Sb227 and the partially sequenced *S. boydii* 18 strain BS512. In Sb227, the putative ORF encoding the R-M specificity subunit is found downstream of two ORFs which encode putative R (restriction) and M (modification) proteins, in turn making up the three putative subunits of a type I R-M system. These sequences also, are found only in Sb227 and BS512.

These results suggest that this could be a novel type I R-M system that is only present in strains of the same lineage and could control the acquisition of island DNA from other species and maybe strains from other lineages, this phenomenon has been reported before in *Staphylococcus aureus* (Waldron and Lindsay, 2006). This possibility is further backed up by the fact that the strains that harbour the R-M system are in cluster 1, a *Shigella* grouping system based on sequence comparisons of four core genes (Lan *et al.*, 2004). Sb227 and BS512 are both cluster 1 strains, being serotypes 4 and 18 respectively, as is S101, (*S. dysenteriae* 3). This cluster also includes (amongst others) *S. boydii* serotypes 1, 2 and 3 and *S. dysenteriae* serotypes 6 and 9; which are all represented in this study and harbour Sb227-like elements at the *leuX* locus. S120, however (*S. boydii* 7 strain) also harbours the Sb227-

like sequence, but is a member of cluster 2. Even so, it would be interesting to see if these strains also harboured the novel R-M system.

Given the likely size of the S101 *leuX* associated island and the context of the genes, it is possible that the putative *hsdR* and *hsdM* genes found in Sb227 are also present, however the location of the *sigA* gene in S101 is yet to be determined.

7.5 *serU*

Table 7.3. SGSP-PCR results of *serU* tRIP negative strain-tRNA loci

<i>serU</i> U# - T7# SGSP-PCR results							
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)					Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sa</i> I	
K12 MG1655			~2.0	~0.5			
<i>S. dysenteriae</i> 3	S101	N ^a	N	N	~2.5 ^b	N	683 [U#]
<i>S. dysenteriae</i> 9	S102	N	N	N	~2.8 ^c , ~0.6 ^d	N	729, 460 [SK#], 1216 UC ^f [U#], 487 [U#]
<i>S. dysenteriae</i> 6	S103	N	N	N	N	N	
<i>S. flexneri</i> 1a	S104	N	N	N	N	N	
<i>S. flexneri</i> 1b	S105	N	N	N	N	N	
<i>S. flexneri</i> 2a	S106	N	N	N	N	N	
<i>S. flexneri</i> 2b	S107	N	N	N	N	N	
<i>S. flexneri</i> 6	S110	N	N	N	N	N	
<i>S. flexneri</i> X	S111						
<i>S. flexneri</i> Y	S112						
<i>S. boydii</i> 1	S116	~1.8		N	N	N	866 [U#], 682 [T7#]
<i>S. boydii</i> 2	S117	~3.3		N	~0.7 F ^e	~2.3 F	615 [U#], 1099 UC [SK#]
<i>S. boydii</i> 3	S118	N		N	N	N	
<i>S. boydii</i> 4	S119	~1.8					
<i>S. boydii</i> 7	S120	~1.8					

<i>serU</i> D# - T7# SGSP-PCR results							
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)					Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	
K12 MG1655				~1.5			
<i>S. dysenteriae</i> 3	S101	N	N	~1.5	N	N	
<i>S. dysenteriae</i> 9	S102	N	N	~1.5	N	N	
<i>S. dysenteriae</i> 6	S103	N	N	~ 1.7	N	N	227 [SK#]
<i>S. flexneri</i> 1a	S104			~1.7	~1.2		
<i>S. flexneri</i> 1b	S105	N		~1.7	~1.2	N	
<i>S. flexneri</i> 2a	S106			~ 1.7	~1.2		769 [D#], 773 [T7#]
<i>S. flexneri</i> 2b	S107			~ 1.7	~1.2		856 [D#], 869 [T7#]
<i>S. flexneri</i> 6	S110			~ 1.5	N		553 [SK#]
<i>S. flexneri</i> X	S111			~1.7	~1.2		
<i>S. flexneri</i> Y	S112			~1.7	~1.2		
<i>S. boydii</i> 1	S116	N		~ 1.5	N	N	574 [D#], 580 [T7#]
<i>S. boydii</i> 2	S117	N	N	N	N	N	
<i>S. boydii</i> 3	S118	N	N	~1.7	N	N	
<i>S. boydii</i> 4	S119			~1.5			
<i>S. boydii</i> 7	S120			~2.5 F			

^a Indicates that no amplicon was generated

^b Text highlighted in bold indicates that the amplicon was sequenced

^c Underlined text indicates the data for the specific SGSP-PCR amplicon (see 7.6 for details)

^d A non-specific SGSP-PCR amplicon that comprised novel sequence (see 7.6 for details)

^e The addition of 'F' after the text indicates that the amplicon was faint

^f The addition of 'UC' after the text indicates that the unclipped sequence was used in the analysis. The sequence run failed by the Phred20 (Ewing and Green, 1998) standards used by the sequencing company; however the low quality sequence still provided some meaningful information.

7.5.1 The *S. flexneri serU* associated prophage-like GI

In the representative strains investigated of *S. flexneri* serotypes 1a, 1b, 2a, 2b, X and Y, the *serU* D-arm results confirm that they harboured the same sequence as the DNA associated with the corresponding D-arm region in the 22.3 kb Sf301 *serU* prophage-like island (*serU*-IF1). This GI has classic prophage-like signatures, such as the presence of an intact prophage integrase gene, and is flanked by DRs; it is also present in the Islander database. This provides strong evidence that the entire island was acquired by HGT, most likely bacteriophage mediated. This GI is described as *Shigella* island number 34, or ‘*ipaH* island IV’ by Jin *et al.*, 2002 as it harbours an *ipaH* gene (see Figure 7.16). The authors speculate that the ‘*ipaH* islands’ present in Sf301 (five over 1 kb in total) were originally associated with *Salmonella* bacteriophage P27 because all five *ipaH* genes are found next to genes that may encode proteins sharing about 75% identity with a hypothetical protein of unknown function from phage P27. Also as most of the other genes on the ‘*ipaH* islands’ have homologies to genes of different phages, they believe that *S. flexneri* then acquired the ‘*ipaH* islands’ from different phages. They also mention that the chromosomal ‘*ipaH* islands’ may have been derived from different sources or via different routes to the *Shigella* virulence plasmid associated *ipaH* genes, because the plasmid borne *ipaH* genes are not associated with any genes that are paralogs of phage P27. The exact function of these chromosomal *ipaH* genes is unknown, but recent studies have shown that they are involved in affecting host gene expression during infection (see section 7.7 below). The speculation that this island is a mosaic, bacteriophage derived entity is further backed by the presence of the integrase gene, which is described as a ‘putative integrase for prophage CP-933U’, which in turn is associated with *serU* in EDL933, again highlighting the specificity of certain phage integrases for particular tRNA loci (Williams, 2002). Also other ORFs present in the GI have homology to genes on the prophages CP-933P, O, N, M and K, all found present in EDL933.

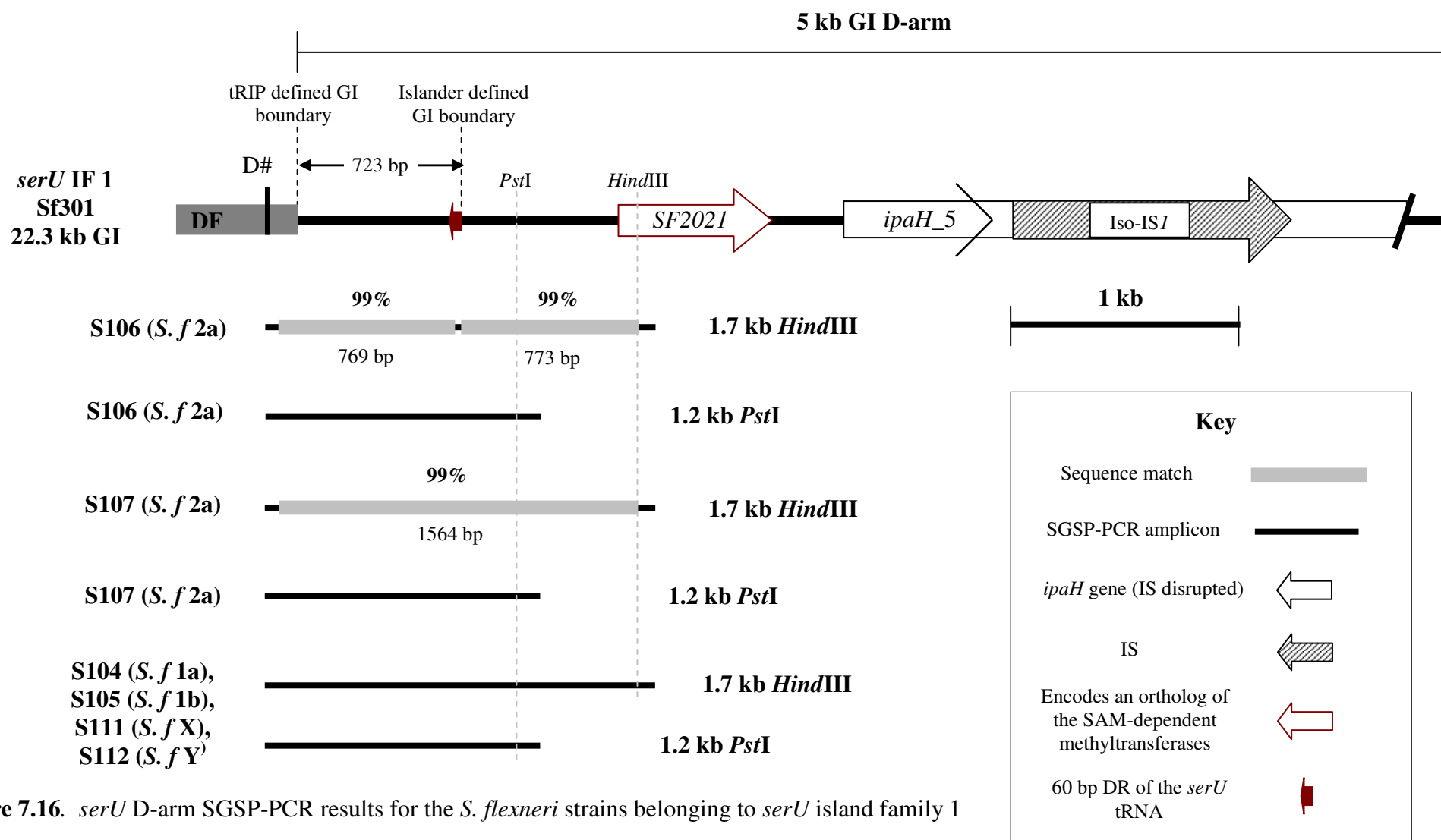


Figure 7.16. *serU* D-arm SGSP-PCR results for the *S. flexneri* strains belonging to *serU* island family 1

The *serU* associated GI in S110 (*S. flexneri* 6 strain) has been classified as ‘uncharacterised’ because only a D# amplicon was obtained and this walks into the *yodB* region described in section 7.5.3 below. However, this does tell us that the putative island is not Sf301-like at the D-arm, but is more similar to Sb227 and K12 MG1655, in contrast to the other tRIP negative *S. flexneri* strains. This is also the case at other tRNA loci (see Table 5.1), providing strong evidence that S110 is of a distinct lineage to the other *S. flexneri* strains.

7.5.2 The *S. boydii* and *S. dysenteriae* strains harbour the same prophage-like GI at the *serU* locus

In the representative strains investigated of *S. boydii* serotypes 1, 2, 4 and 7, the *serU* U-arm results show the presence of a putative prophage integrase gene and a putative exodeoxyribonuclease VIII gene with the sequence results indicating the highest nucleotide identity to the Sb227 *serU* associated prophage-like island integrase gene (*serU*-IF2, see Figure 7.17). These two elements are also found present in the U-arm of the 46.6 kb *serU* associated prophages in EDL933 and Sakai (which are included in the Islander database); however, the rest of the GI differs in content to the EDL933 element. These results show, as in Sf301 that the Sb227 *serU* GI is a mosaic, phage derived GI, harbouring elements found at *serU* in other pathogenic *E. coli* strains.

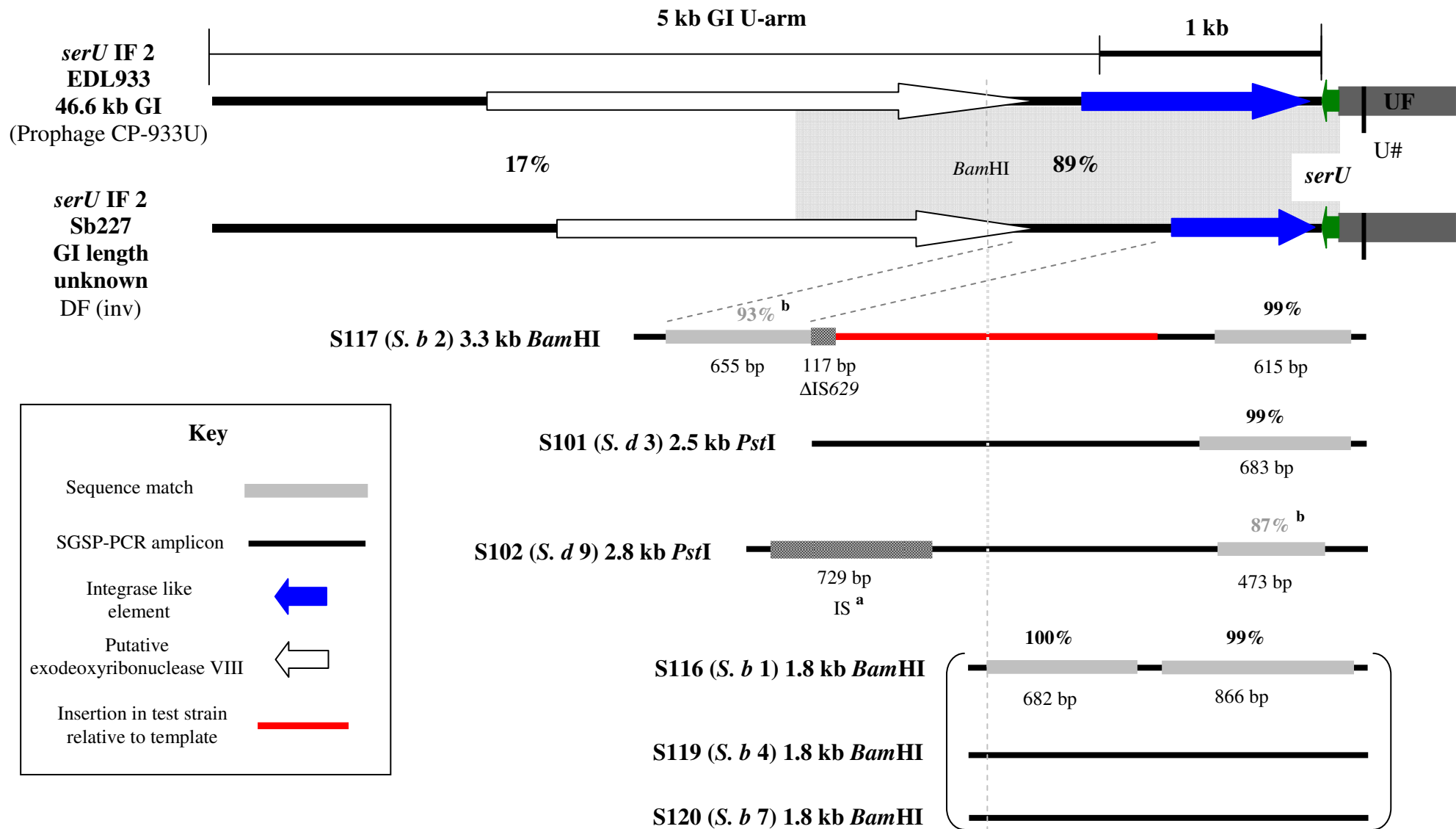


Figure 7.17. *serU* U-arm SGSP-PCR results for the *S. boydii* and *S. dysenteriae* strains belonging to *serU* island family 2

^a Mosaic IS elements

^b Sequence runs failed, however the unclipped (low quality) sequences hit to the regions indicated

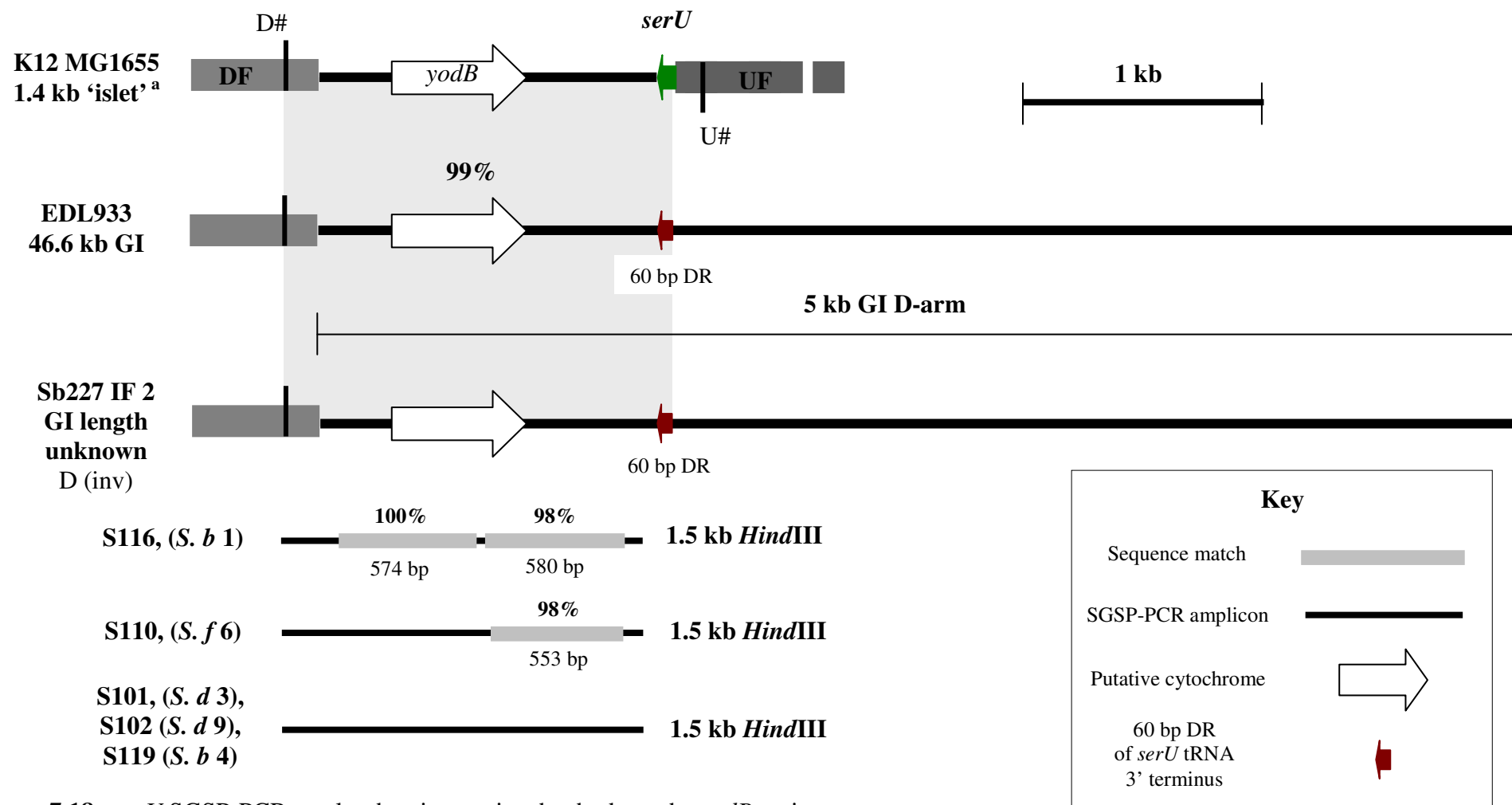
The length of the Sb227 GI could not be determined by *in silico* tRIP as the DF is inverted and translocated 48.0 kb upstream of the *serU* locus. However, further analysis using blastn and Artemis visualisation indicated that directly downstream of *serU* is a 20.6 kb element that is comprised of the above mentioned genes, then ORFs that encode hypothetical proteins, phage-like genes and IS elements; also like in the D-arm of the Sf301 *serU* associated GI, an *ipaH* gene is present. The remainder of the element associated with the DF and displaced as mentioned above is 8.6 kb long and bounded by a 31 bp imperfect repeat of the *serU* 3' terminus. The displaced D-arm harbours phage-like genes and also contains an *ipaH* gene, further analysis indicated that a 4.2 kb region of the displaced sequence has 100% nucleotide identity to the region that contains the *ipaH* gene found at the end of the 20.6 kb *serU* associated element, indicating that at some time the *ipaH* gene and some of the surrounding phage-like DNA has been duplicated. The presence of a number of inverted IS elements flanking both the duplicated sequence and the entire displaced region, suggests that these mobile elements may have firstly duplicated this region of the *serU* associated prophage, and then displaced the sequence and DF region. This in turn suggests that this *ipaH* locus may also be independently mobile.

The U-arm results from the *S. dysenteriae* strains show that in S101 and S102 (serotypes 3 and 9) there is a putative prophage integrase gene, with the sequence results indicating the highest nucleotide identity to the Sb227 *serU* associated prophage integrase gene, as with the characterised *S. boydii* strains. However the restriction patterns are different, and S102 has mosaic IS elements downstream of the integrase gene.

7.5.3 The *S. boydii* and *S. dysenteriae serU* GI D-arms

The D-arm results show that the strains representative of *S. boydii* serotypes 1 and 4, *S. dysenteriae* serotypes 3 and 9 and S110 (*S. flexneri* 6 strain) have walked into a region that contains the *yodB* gene (see

Figure 7.18). This is the same region that is present in the D-arms of the *serU* associated island DNA in K12 MG1655, Sb227, EDL933 and nine other *E. coli* genomes present in the NCBI database, but is not present in Sf301, Sf2457T and CFT073. Its presence in so many genomes made it apparent that this region may comprise core chromosomal DNA and not necessarily island DNA.



^a Re-defined as flanking DNA, see below.

After more detailed analysis using Blastn it was found that a 1013 bp region surrounding the *yodB* gene was probably deleted in Sf301, Sf2457T and CFT073, also there is another 432 bp between the region deleted and the inwardly directed *serU* in K12 MG1655 that is at least 96% conserved in all corresponding *E. coli* and *Shigella* genome sequences available at the time of analysis, that should have been used to place the *serU* D#; the *serU* locus in K12 MG1655 should have been classified as ‘empty’ (see Figure 7.19). If this had been discovered earlier in the study, the *yodB* region would have been designated as part of the DF, however, as the 432 bp conserved DF region was not detected in the original screen of the six genomes studied, the *serU* DF was defined as starting at a point 1.4 kb distal to its likely true location and the D# was placed in this region.

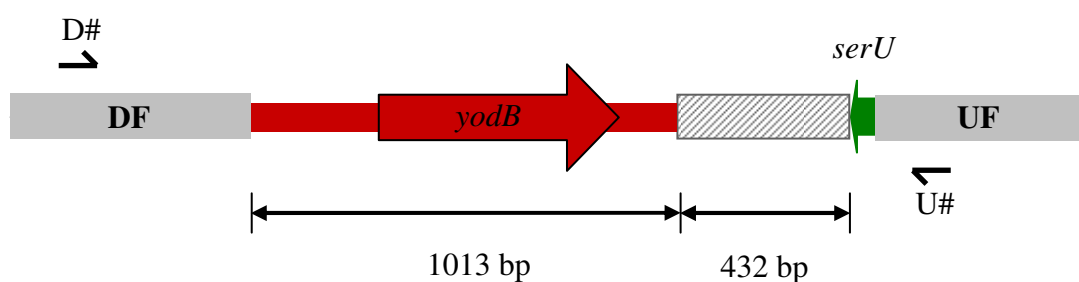


Figure 7.19. View of the *E. coli* K12 MG1655 chromosome, showing the 1013 bp region in the *serU* DF that is deleted in Sf301, Sf2457T and CFT073.

The deleted region is highlighted in red. The grey hatched region indicates the sequence conserved in all of the sequenced *E. coli* and *Shigella* strains, where the D# should have been placed. The position of the originally defined U and D flanking regions are shown as grey boxes, with the respective U and D primers used for tRIP and SGSP-PCR also indicated. Figure is not to scale.

Interestingly the 1013 bp region deleted in Sf301 has a GC content of 41.4%, and the GC content of the 5 kb region directly downstream of *serU* in K12 MG1655 is 43.8%, lower than

the average for the *E. coli* genome which is 50.8%. *yodB* is described as a 'putative cytochrome b561 homolog' (COG 3038) and (Gerdes *et al.*, 2003) defined it as non-essential to K12 MG1655 after a transposon footprinting study. They also give it an ERI (evolutionary retention index) of 0.313. "The ERI for a gene is defined as the fraction of organisms in a diverse set of 33 bacterial species which contain an ortholog of the gene in their genomes. The higher the number the more conserved this gene is in these different genomes, with a score of 1 meaning it is present in all 33 genomes (33/33=1)". So in this case an ortholog is present in only about a third of the genomes, and the authors proposed that for a gene to be essential it must be present in at least 80% of the genomes tested. Also the authors speculate that the 30 kb region downstream of *serU* is all non-essential. Prior to this, (Lawrence and Ochman, 1998) performed a study of the K12 MG1655 genome and defined *yodB* as a horizontally transferred gene. Given this evidence it is likely that the region downstream of *serU* in K12 MG1655 is ancient island DNA that was acquired by *E. coli* before it diverged, hence the finding that it is well conserved in many strains.

Overall, after more detailed analysis, I have re-defined the *yodB* region deleted in some strains as part of the downstream flanking DNA, because further analysis found that there is still 432 bp of well conserved DNA upstream of this region, which is adjacent to *serU* in an 'empty' strain such as K12 MG1655. The D-arms of the *serU* tRIP-negative strains indicated in

Figure 7.18 were therefore re-classified as 'to be confirmed' (see Table 5.1) as the suboptimal location of the D# alone was unlikely to account for negative tRIP PCR results. Therefore, in future experiments a new *serU* D# should be placed in this location in order to walk into more recently acquired putative island DNA.

7.6 The S102 (*S. dysenteriae* 9 strain) novel sequence

serU U# SGSP-PCR with the S102 *Pst*I library produced two amplicons of similar intensity (see Figure 7.20). Generally, when multiple bands were obtained in SGSP-PCR, as the method had been optimised, the specific amplicon was the brightest and any non-specific bands were much fainter. However, in this case it was difficult to tell which amplicon was the specific SGSP-PCR product, because PCR amplification of the larger product (2.8 kb) was likely to have been significantly less efficient than that of the smaller band (0.6 kb), it should be the lesser species in the amplification reaction (Don *et al.*, 1991). As this is not the case and the bands were the same intensity, it was postulated that the 2.8 kb amplicon was more likely to be the specific product.

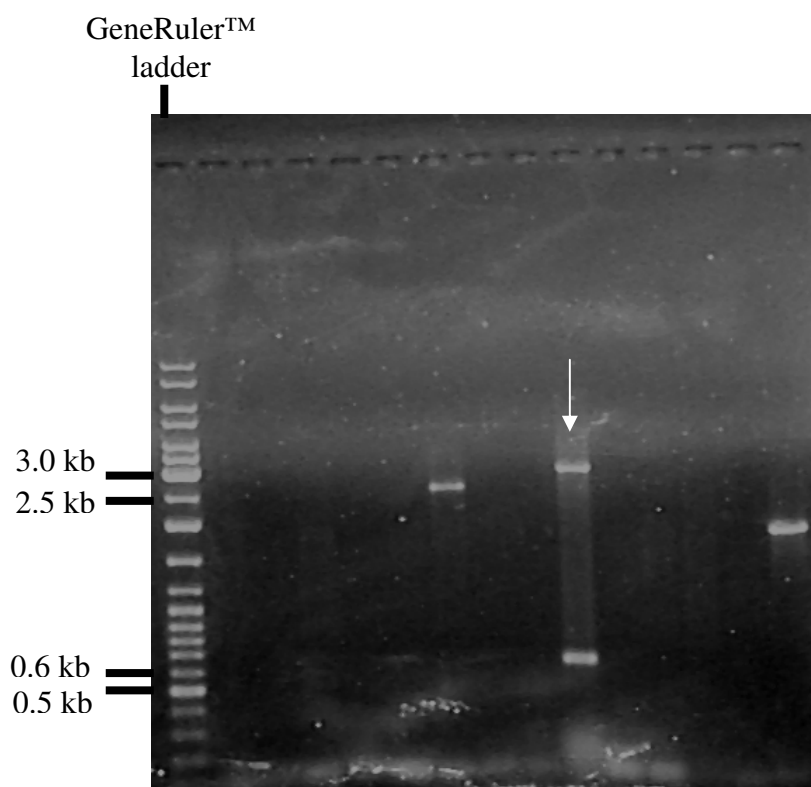


Figure 7.20. Agarose gel of a *serU* U# SGSP-PCR showing the 2.8 kb and 0.6 kb amplicons produced by the S102 (*S. dysenteriae* 9 strain)/*Pst*I library as indicated by the white arrow.

Both amplicons were sequenced, and the 2.8 kb product was found to be the specific amplicon as hypothesised and the *serU* GI sequence was associated with this (see Figure 7.17). The 0.6 kb amplicon was non-specific, but turned out to be an interesting discovery nevertheless. When the 0.6 kb amplicon was sequenced from both ends and the vector sequence trimmed off, a contig of 483 bp of sequence was obtained. After analysis with Blastn, it was found that the end of the sequence (464-483) walks into the *serU* U#, to just 3 bp short of the 5' end of the primer, however the rest of the sequence (463 bp) had only thirteen partial hits to the entire NCBI database, with the only prokaryotic hit being to the *Rhodopirellula baltica* SH 1 genome (GenBank accession number BX294144) across 20 bp of the sequence, with 100% ID, a score of 40.1 and an E-value of 7.2. An E-value of any greater than 0.05 means that the hit is statistically insignificant and is more likely to have occurred due to chance, rather than it being the correct match. This proves firstly that the amplicon is non-specific as it is not associated with the *serU* UF, and secondly that the entire sequence is completely novel with the only known region being the PCR incorporated U# sequence. (see Figure 7.21)

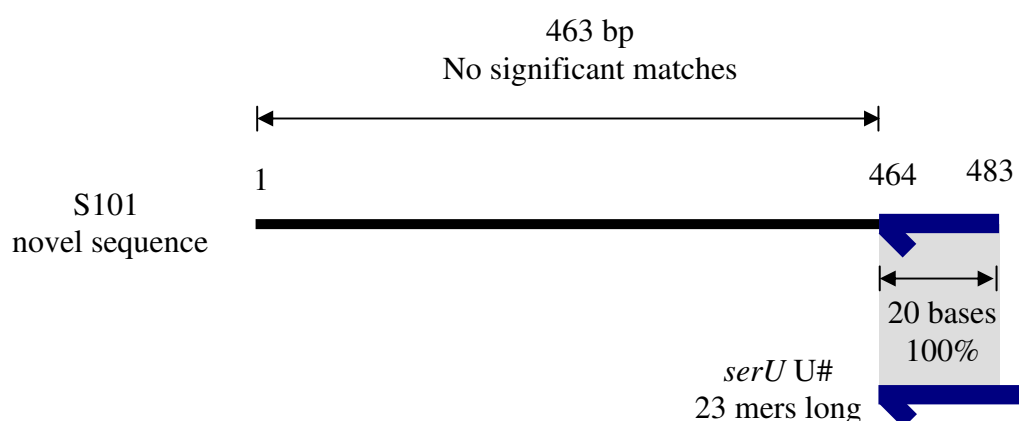


Figure 7.21. Blastn analysis of the S101 (*S. dysenteriae* 3 strain) novel sequence discovered by SGSP-PCR with the *serU* U#.

20 bases of the sequence hits to the *serU* U# with 100% nucleotide identity (incorporated through non-specific binding) and the rest of the sequence has no significant matches in the NCBI database. Figure is not to scale.

The novel sequence has a GC content of 57.8%, exactly 7% higher than the average for the K12 MG1655 genome, and as it is completely novel, it is very likely to be horizontally acquired DNA.

tBlastx analysis of the novel sequence indicated that the most significant hit was from bases 1 to 207, with 59% amino acid identity to an ‘unnamed protein product’ in *Photorhabdus luminescens* subsp. *laumondii* TTO1 genome (GenBank accession number BX571866).

7.7 Linkage of the *ipaH* genes with the *serU* associated prophages in *Shigella*

After performing a Blastn study, it was found that in Sf301, there are a total of seven chromosomal *ipaH* genes, and two of these are disrupted by an iso-ISI. One of the genes disrupted is the one found on the *serU* associated prophage. In Sb227 there are a total of six chromosomal *ipaH* genes, two of them are identical, possibly due to a duplication event and both are present as part the *serU* associated prophage DNA (see section 7.5.2). Both of the *serU* associated Sb227 *ipaH* genes have 99% nucleotide ID to the cognate *ipaH* gene in Sf301 after removal of the iso-ISI in the latter strain. All of the chromosomal *ipaH* genes in these two sequenced strains are associated with prophage DNA, again highlighting the role phages have played in the dissemination of these elements. The *ipaH* gene on the Sf301 *serU* prophage-like GI is annotated as having 77% identity to the virulence plasmid borne *ipaH* 9.8 gene, the product of which is known to be transported via the virulence plasmid-encoded type III secretion system (TTSS) into the nuclei of host cells, where it may affect expression of host genes involved in the inflammatory response (Toyotome *et al.*, 2001). Also, recently a

study on the *S. flexneri* 2a strain YSH6000 has proved that each of the chromosomally encoded IpaH proteins are also secreted by the TTSS, and that they modulate the inflammatory response of the host (Ashida *et al.*, 2007). This provides direct evidence that the chromosomal *ipaH* genes have an important role in the virulence of *Shigella*. However, as the Sf301 *serU* prophage *ipaH* gene is disrupted it is unlikely to be functional in this strain. Even so, the *serU* associated islands found in the panel of *Shigella* strains tested in this study, could be playing a key role in virulence if they harbour intact *ipaH* genes, plus they could be involved in the continual dissemination of *ipaH* genes in *Shigella*.

In EDL933, the *serU* associated element is the 46.6 kb CP-933U prophage, that harbours the recently characterised *espJ* and *tccP* genes, which form an operon and encode effector proteins that are translocated into the host cells by the LEE PAI-encoded type III secretion system (TTSS) (see (Garmendia *et al.*, 2005) for an excellent review). This gives us insights into the overall role that the *serU* associated GIs play; as in both *Shigella* and *E. coli*, the islands harbour genes that encode effector proteins that are translocated by type III secretion systems into the host cell, which are in turn encoded by separate genetic entities (the virulence plasmid-encoded TTSS in *Shigella* and the *pheV* associated LEE PAI-encoded TTSS in *E. coli* O157:H7).

7.8 The *Shigella serU* prophage is a selfish element

Another interesting observation regarding the serine tRNA sites across the *Shigella* strains in this study that harbour *ipaH* bearing prophage-like GIs at *serU* is that generally the other serine tRNA loci in these genomes are not occupied by other significant GIs. The exceptions are S107 and S110 which harbour SRL-related PAIs at *serX* and S101 which harbours an SRL-related PAI at *serW*. Besides these three cases, Table 5.1 shows that in the strains in which the *serU* site is occupied by an *ipaH* island, the *serW* locus is empty and the *serX* locus

is only occupied by a 1.5 kb islet (see section A2.6.8, *serX* results). This pattern indicates that the presence of this *serU* prophage may prevent the other serine tRNA genes from being occupied by similar prophage-like entities. This hypothesis is further supported by the presence of the ORF designated *SF2021* in the D-arm of the Sf301 island (see Figure 7.16); this ORF is also found present in the same context in Sb227. It encodes an ortholog of the SAM-dependent methyltransferases (COG0500), which are part of prokaryotic restriction-modification systems that in turn are often associated with prophage DNA; indicating that the island could at some time have been acting in a selfish manner (Kobayashi, 2001), protecting the host organism from invasion by other similar prophage DNA, and also promoting its own survival in the chromosome of *Shigella*.

Overall, at the *serU* locus, 2 different GI ‘families’ were discovered across the 19 *Shigella* strains. The island content in *S. flexneri* appears to be well conserved, with all but one of the strains (S110, *S. flexneri* 6), possessing sequences identical to the Sf301 associated *serU*-linked prophage. Whereas, in the *S. boydii* and *S. dysenteriae* strains, the islands are more similar to the Sb227 prophage-like island. In all of the characterised strain-tRNA loci, the elements walked into have ‘classic’ island signatures and are represented in Islander. All are prophage-like, indicating that this site is a hot-spot for phage activity across all *Shigella* strains but the *S. sonnei* strains investigated; in the case of the latter, all three strains tested possessed empty *serU* loci.

7.9 *aspV*

Table 7.4. SGSP-PCR results of the *aspV* tRIP negative strain-tRNA loci

<i>aspV</i> U# - T7# SGSP-PCR results									
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)							Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	<i>Eco</i> RV	<i>Hinc</i> II	
K12 MG1655				~1.0					
<i>S. flexneri</i> 1a	S104	N ^a	N	N	N	N			Southern = Sf301 U-arm ^h
<i>S. flexneri</i> 1b	S105	N	N	N	N	N			Southern = Sf301 U-arm ^h
<i>S. flexneri</i> 2a	S106	N	N	N	N	N			Southern = Sf301 U-arm ^h
<i>S. flexneri</i> 2b	S107	N	N	N	N	N			Southern = Sf301 U-arm ^h
<i>S. sonnei</i>	S108	N	N	~1.2 F	N	N	N ^{*f}	N [*]	Southern = Ss046 U-arm ^h
<i>S. flexneri</i> X	S111								Southern = Sf301 U-arm ^h
<i>S. flexneri</i> Y	S112								Southern = Sf301 U-arm ^h
<i>S. sonnei</i>	S113	N	N	N	N	N			
<i>S. sonnei</i> bio a	S114	N	N	N	N	~1.5 F			
<i>S. sonnei</i> bio g	S115	N [*]	N [*]	N [*]	N [*]	N [*]	N [*]	N [*]	Southern = Ss046 U-arm ^h
<i>S. boydii</i> 7	S120	N	N [*]	N [*]	N [*]	N			

aspV D# - T7# SGSP-PCR results									
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)							Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	<i>Eco</i> RV	<i>Hinc</i> II	
K12 MG1655			~2.3	~2.3					
<i>S. flexneri</i> 1a	S104		~ 1.5 *						2562 UC [SK#]
<i>S. flexneri</i> 1b	S105	~0.8 F, ~1.8 F	~0.6 F, ~ 1.4 ^{bc}	N	~1 F ^d	~1.7 F			533 [SK#], 224 [D#]
<i>S. flexneri</i> 2a	S106	~1.8 F	~0.6 F, <u>~1.4</u>	N	~1 F	~1.7 F, ~2 F			
<i>S. flexneri</i> 2b	S107		~ 1.5 *						2203 UC [D2#]
<i>S. sonnei</i>	S108	~0.8 F, ~1.8 F	N	N	N	~1.7 F	N *	~ 1.5 *	1085 UC ^e [SK#], 132 [D2#]
<i>S. flexneri</i> X	S111		~ 1.5 *						1651 UC [D#]
<i>S. flexneri</i> Y	S112		~1.5 *						
<i>S. sonnei</i>	S113	~0.8 F, ~1.8 F, ~2 F *	N *	N *	N *	N *			
<i>S. sonnei</i> bio a	S114	~ 0.8 F , ~1.8 F	N *	N	N	N			652 [SK#], 676 [D#] NS ^g
<i>S. sonnei</i> bio g	S115	N *	N *	N *	N *	N *	N *	~ 1.5 *	422 [SK#]
<i>S. boydii</i> 7	S116	N	N *	N *	N *	N			

^a Indicates that no amplicon was generated

^b Text highlighted in bold indicates that the amplicon was sequenced

^c Underlined text indicates the data for the specific SGSP-PCR amplicon

^d The addition of 'F' after the text indicates that the amplicon was faint

^e The addition of 'UC' indicates that the unclipped sequence was used in the analysis. The sequence run failed by the Phred20 (Ewing and Green, 1998) standard used by the sequencing company; however the unclipped (low-quality) sequence still provided some meaningful information.

^f The presence of a * after the text indicates that the secondary U or D primer was used in the SGSP-PCR reaction. As SGSP-PCR with the original D# produced multiple faint amplicons and after blastn analysis with the smallest word (sliding window) size of 7 bases, it was found that both the original U# and D# have non-specific hits to various known *E. coli* sequences at their 3' termini, which could increase the likelihood of non-specific amplicons. Therefore a new pair of primers were designed that were more stringent, these were designated *aspV* U2# and D2# (see Table A2. 1).

^g The addition of 'NS' indicates that the amplicon was found to be non-specific after sequence analysis.

^h Indicates that a Southern hybridisation was performed with an *aspV* U# probe to confirm the *aspV* U-arm GI status of this strain, see section 7.9.1 for more details.

7.9.1 *aspV* U-arm Southern hybridisation

The *aspV* U# SGSP-PCRs yielded no results with all 5 of the original enzyme libraries in the initially screened tRIP-negative test strains, and as this was early in the study I was unsure that the SGSP-PCR technique was robust. Therefore I decided to check if the U# region even existed across the *Shigella* strains, as the strains being tested were previously uncharacterised and it was possible that the negative SGSP-PCRs were due to the absence of the U# sequence. To check for the presence of the U#, a 'screening' Southern hybridisation was performed on *Hind*III digested genomic DNA from the isolates, *Hind*III was chosen because initially four of the tRIP-negative *S. flexneri* isolates were U# SGSP-PCR negative and I hypothesised that they were likely to have the same *aspV* GI as is present in Sf301. An *in silico* *Hind*III digest of Sf301 showed that the U# lies on an 11.7 kb fragment which walks 6.3 kb into the U-arm of the *aspV* associated GI (see Figure 7.22). Southern hybridisation could therefore determine the U-arm GI status of the test isolates, as the results could be compared against the

sequenced *E. coli* and *Shigella* genomes available. Also K12 MG1655 which is tRIP-positive at *aspV* was used as a control as the U# lies on a 2465 bp *Hind*III fragment.

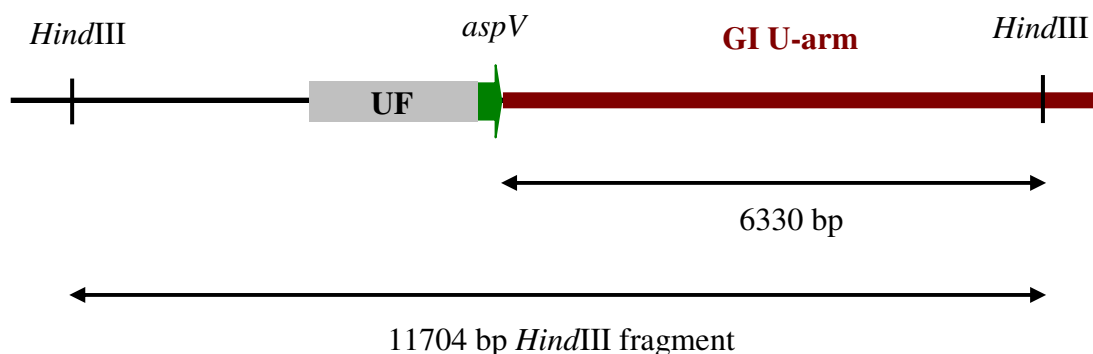


Figure 7.22. *Hind*III restriction map of the region surrounding the Sf301 *aspV* tRNA locus.

Figure is not to scale.

Two primers were designed (*aspV* UF probe F# and *aspV* UF probe R# (see Table A2. 2) to amplify by PCR from the K12 MG1655 genome a small region of the UF that surrounds the *aspV* U# (see Figure 7.23). This region had at least 98% nucleotide identity to the cognate sequences in the U flanking regions of all of the sequenced *E. coli* and *Shigella* strains available on the NCBI database and was therefore very likely to have high nucleotide identity to the same region in uncharacterised strains. The PCR product was then purified, Digoxigenin (DIG)-labelled and used as the probe in a Southern hybridisation (see 2.15 for the protocol).

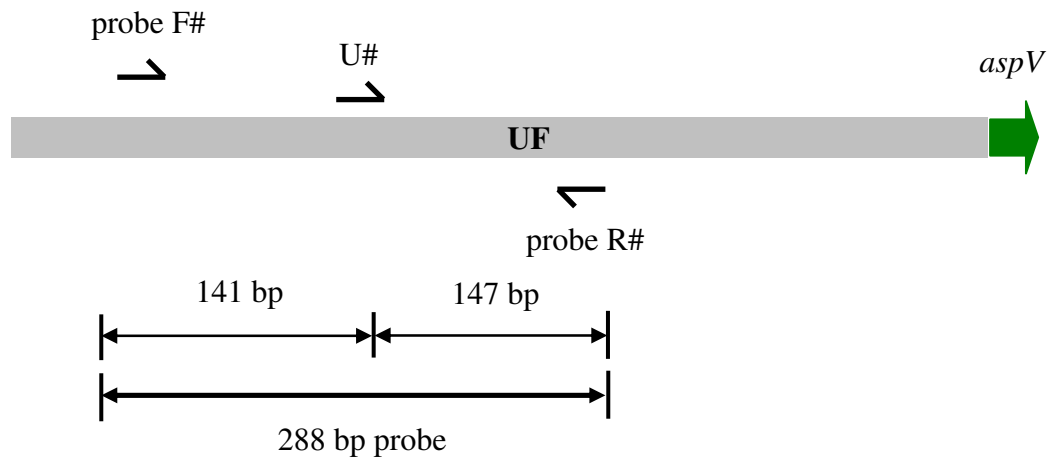
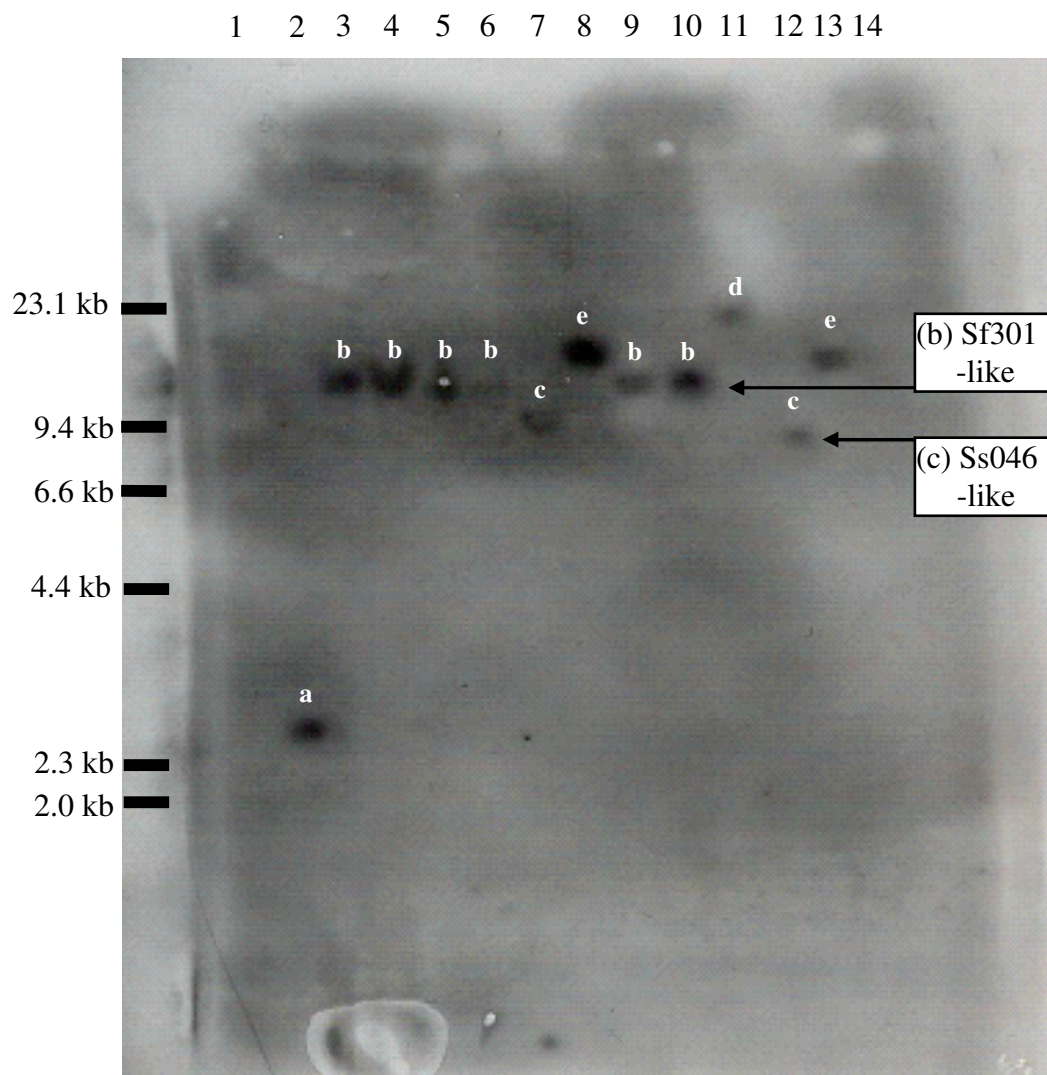


Figure 7.23. Relative positions of the *aspV* UF probe primers on the K12 MG1655 chromosome.

Figure is not to scale.

Some of the strains included in the Southern hybridisation had not been screened by tRIP at the time of this experiment, but they were included out of interest to confirm the presence of the U#. Figure 7.24 shows the results of the Southern hybridisation

7.9.2 Southern hybridisation results



Lane Number:

1. λ /HindIII ladder
2. *E. coli* K12 MG1655(control, tRIP +ve)
3. *S. flexneri* 1a (S104)(tRIP -ve)
4. *S. flexneri* 1b (S105)(tRIP -ve)
5. *S. flexneri* 2a (S106)(tRIP -ve)
6. *S. flexneri* 2b (S107)(tRIP -ve)
7. *S. sonnei* (S108)(tRIP -ve)
8. *S. flexneri* 6 (S110)(tRIP not performed at time of study, tRIP +ve after screen)
9. *S. flexneri* X (S111)(tRIP not performed at time of study, tRIP -ve after screen)
10. *S. flexneri* Y (S112)(tRIP not performed at time of study, tRIP -ve after screen)
11. *S. sonnei* bio a (S114)(tRIP not performed at time of study, tRIP -ve after screen)
12. *S. sonnei* bio g (S115)(tRIP not performed at time of study, tRIP -ve after screen)
13. *S. boydii* 4 (S116).....(tRIP not performed at time of study, tRIP +ve after screen)
14. *S. boydii* 7 (S120)(tRIP not performed at time of study, tRIP -ve after screen)

Figure 7.24. Photograph of the *aspV* Southern hybridisation membrane using the DIG system

The *aspV* tRIP status of each strain is described below the gel picture.

The results show that the U# region is present as expected on a fragment of around 2.5 kb in the control strain K12 MG1655 (a). In all of the tRIP negative *S. flexneri* isolates (apart from S108 that was later confirmed as a *S. sonnei* by an API-20E test see section 5.7), the U# region lies on the same size *Hind*III restriction fragment as is found in Sf301 (see (b), indicating that in lanes 3, 4, 5, 6, 9 and 10 the probe has bound to a fragment of about 11.7 kb), confirming my hypothesis that they harbour the Sf301 GI U-arm at the *aspV* locus. In S115 (*S. sonnei* bio g) and S108 (*S. sonnei*), the probe lies on a fragment of around 8.8 kb (c), whereas in S114, it is on a fragment of around 22.0 kb (d). This initially showed that in these isolates the putative GI is different to that in Sf301, also that there is variation in this region between the *S. sonnei* isolates. Further analysis after the *S. sonnei* 046 (Ss046) (GenBank accession no. NC 007384) chromosome was completely sequenced (Yang *et al.*, 2005), showed that in this strain the U# region is present on an 8841 bp *Hind*III fragment that walks 7364 bp into the GI, therefore showing that S108 and S115 (c) are very likely to have the same U-arm as Ss046, whereas S114 is different.

S110 and S119 had the same profile (e) which was distinct from all of the other strains. These two strains turned out to be tRIP positive, with U#-D# amplicons of 1.7 kb, indicating the presence of an islet of 1.0 kb between the UF and DF. Sequencing analysis confirmed this islet to be an *IS1* (see Table A2. 5), this was later found to correspond to the *aspV* locus in the fully sequenced *S. boydii* 4 227 (Sb227) strain (GenBank accession no. NC 007613) (Yang *et al.*, 2005) which has an *in silico* tRIP product of 1719 bp. This result, as with other loci indicates how the S110 *aspV* GI content is more similar to Sb227 than Sf301, suggesting that it comes from a different lineage to the other *S. flexneri* strains. The only strain that remained uncharacterised was the *S. boydii* 7 isolate as no signal was detected in lane 14.

7.9.3 The Sf301 *aspV* associated GI

The *aspV* U-arm results confirmed by the Southern hybridization and the D-arm SGSP-PCR results show that all of the *S. flexneri* strains (except for S110) harbour the same DNA as is present in the 57.7 kb tRIP-defined Sf301 *aspV* associated GI (*aspV*-IF1), see Figure 7.26 also). This GI does not harbour an integrase gene homolog or any DRs so is therefore not defined by Islander. However, part of this Sf301 island has been described before as the ‘*sci*’ island by Jin *et al.*, 2002 as it harbours paralogs of the *sciCDEFF* operon, which is found on the *Salmonella enterica* centrisome 7 genomic island (SCI), which is associated with the *aspV* locus in *S. enterica* subspecies I strains, (Folkesson *et al.*, 2002) and bacteriophage P22 HK620. The authors describe the GI as being 22,789 bp long; whereas when defined by tRIP there is another 34.9 kb of island DNA downstream of the *sci* element (see Figure 7.25).

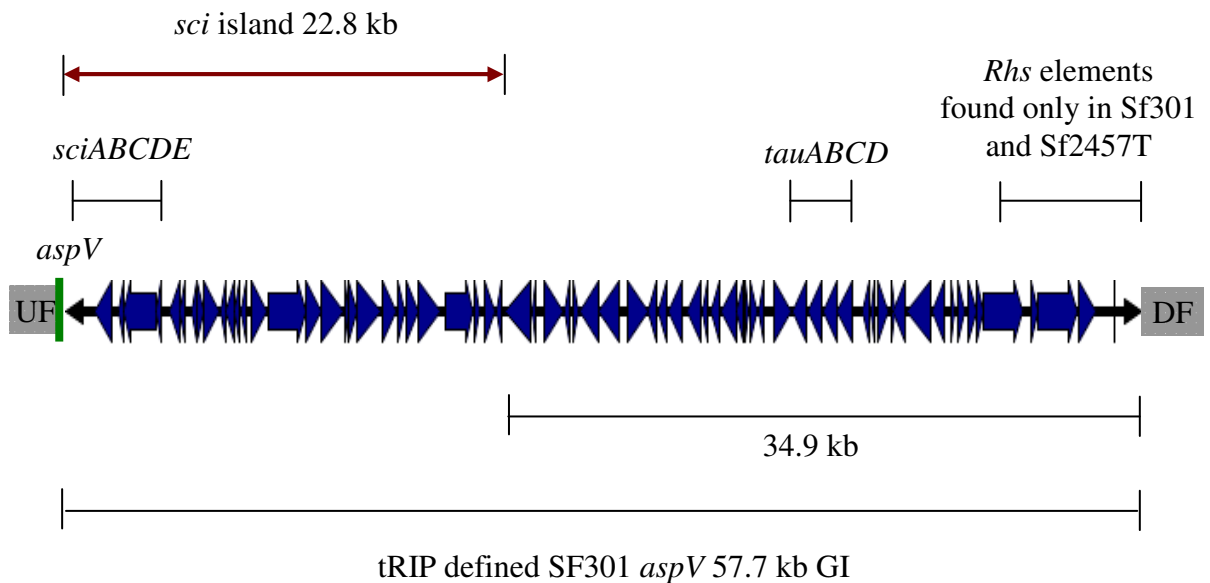


Figure 7.25. The Sf301 *sci* island as defined by Jin *et al.*, 2002, and the corresponding tRIP defined GI found associated with the *aspV* locus in Sf301.

The grey boxes indicate the *aspV* UF and DF, the green line shows the position of the *aspV* tRNA and the blue regions indicate ORFs present in the corresponding GI (taken from coliBASE [(Chaudhuri *et al.*, 2004)] www.colibase.bham.ac.uk). Figure is not to scale.

7.9.4 Putative functions of the *Shigella aspV* associated *sciABCDE* gene cluster

Genes that encode proteins similar to the proteins encoded by the *S. enterica* subspecies I *sci* island are also found in other Gram-negative bacteria that live in close contact to and have the capacity to manipulate eukaryotic cells, such as *Vibrio cholerae*, *Pseudomonas aeruginosa*, *Yersinia Pestis* and EDL933. These Sci-like proteins have either known or putative roles in virulence and may encode a secretion system and maybe species-specific effector molecules. Also deletion of the *sci* island in *S. enterica* attenuates their virulence, resulting in 50% less ability to enter cultured eukaryotic cells (Folkesson *et al.* 2002). This suggests that in *Shigella*, the *aspV* associated *sci* elements also encode novel virulence associated proteins.

As the results of this study indicate that these elements are likely to be found in the majority of *S. flexneri* strains, it would be interesting to determine the function of the Sci-like proteins.

7.9.5 The *Shigella tauABCD* gene cluster

The Sf301 34.9 kb tRIP defined distal GI ‘extension’ found downstream of the *sci* island contains various IS elements, ORFs that encode hypothetical proteins and the *tauABCD* gene cluster (see Figure 7.25) which is found in *E. coli*, *Shigella* and other members of the Enterobacteriaceae and confers the ability to utilise taurine as a sulphur source. The expression of these genes is regulated by sulphur or cysteine starvation and they are part of the ‘toolbox’ for survival in nutrient limiting conditions (van der Ploeg *et al.*, 1996). Blastn analysis and Artemis visualisation of the complete *E. coli* and *Shigella* genomes indicated that in most strains these genes are not associated with island DNA. The only exceptions were Sf301 (and Sf2457T) and Sb227, where the genes are harboured on the *aspV* and *thrW* GIs respectively. In these strains the genes are bound by inverted *IS1* elements, this suggests that these mobile elements have played a role in the transposition of the genes between core and island DNA and that the *tauABCD* gene cluster could be independently mobile.

7.9.6 The *S. flexneri* novel *Rhs* elements

The D-arm results show that the tRIP-negative *S. flexneri* strains all harbour the same DNA as is found in the D-arm of the Sf301 *aspV* associated 57.7 kb GI (see Figure 7.26) and subsequent sequence analysis indicated that the region walked into was previously only found present in Sf301 and Sf2457T. The sequence run walks into *SF0268*, an ORF annotated as encoding a ‘putative *Rhs*-family protein’. *Rhs* (recombination hot spot) elements are regarded as accessory genetic elements, eight have been found so far in *E. coli* and there are five in K12 MG1655. Their function is unknown, however they are known to recombine at a frequency of 10^{-5} in *E. coli* K12 strains, causing duplication of the DNA found between them

(Lin *et al.*, 1984). They are likely to be horizontally acquired as they are not found in all *E. coli* strains, they have very different GC contents to the core *E. coli* genome and are composites assembled from components with different evolutionary histories which has led to the hypothesis that their acquisition was relatively recent (see (Zhao *et al.*, 1993) and (Wang *et al.*, 1998) for excellent studies on *Rhs* elements in *E. coli*). In *E. coli* the *Rhs* core ORF is typically 3.7 kb long and has a GC content of around 60%; in Sf301 *SF0268* has a GC content of 56.4% and is therefore likely to be the terminus of a novel putative *Rhs* core-ORF which actually comprises the putatively annotated ORFs *SF0266*, *SF0267* and *SF0268*; because on further analysis, this region was found to be 3.7 kb in length and has a GC content of 58.2%. Also the DNA between *SF0268* and the conserved DF has a GC content of 39.3%; this is likely to be the ext/ds region of the putative *Rhs* element, these are known to be variable, often unique regions of DNA that are AT-rich, having GC contents around 40% and can be up to 1.0 kb in length. Therefore, this evidence suggests that the content and context of the entire region is characteristic of an *Rhs* element and the DNA walked into in the tRIP-negative *S. flexneri* strains is part of a unique *Rhs* element found previously only in Sf301 and Sf2457T. Whether these elements are currently playing a role in the rearrangement of DNA in *S. flexneri* is yet to be determined.

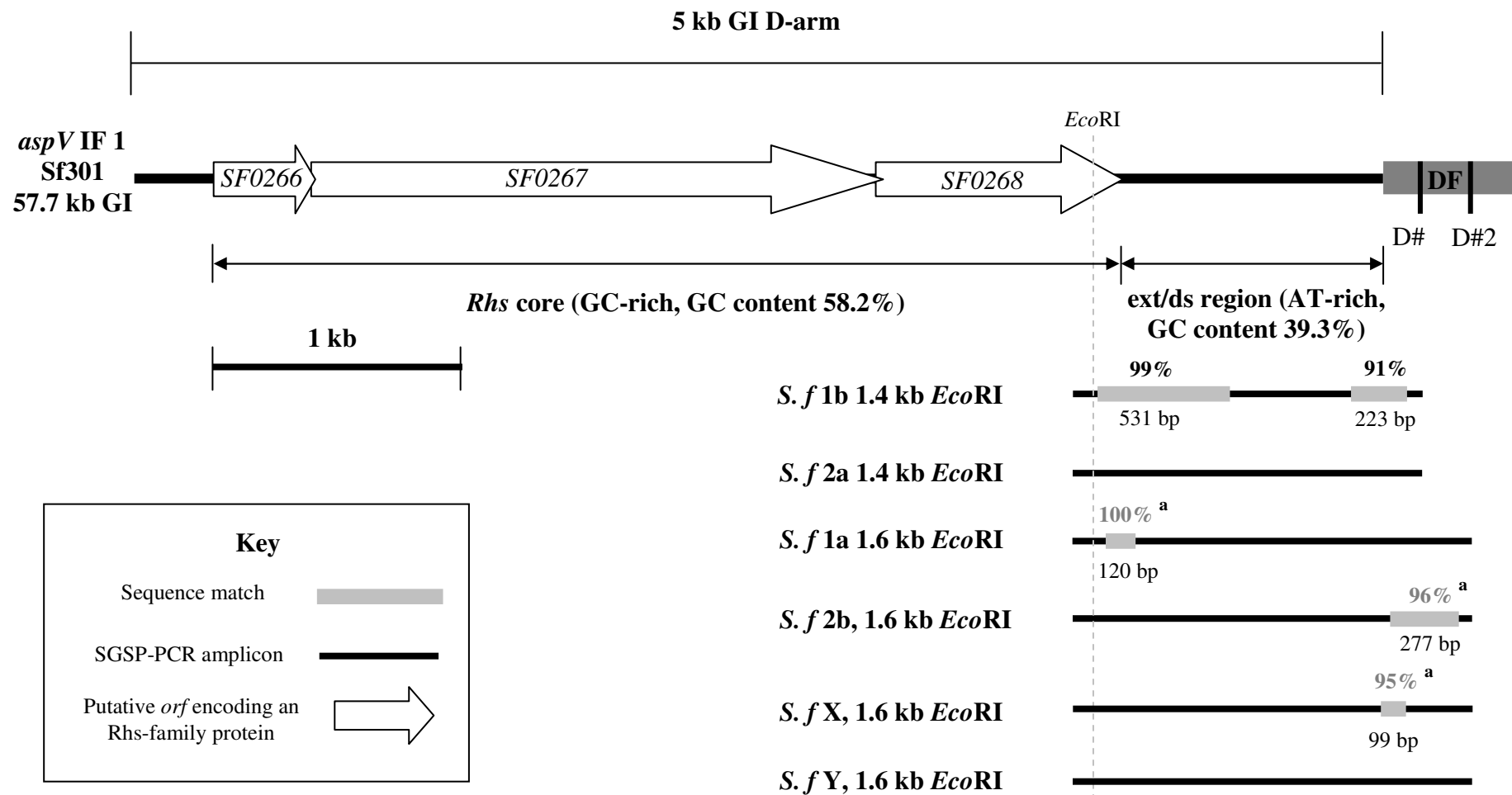


Figure 7.26. *aspV* D-arm SGSP-PCR results for the *S. flexneri* strains belonging to *aspV* island family 1

^a Sequence runs failed, however the unclipped (low quality) sequences hit to the regions indicated

7.9.7 *S. sonnei*

The D-arm results for the characterised *S. sonnei* strains show that they harbour the same sequence as is found in the corresponding D-arm in the Ss046 *aspV* associated 36.1 kb GI (*aspV*-IF2, see Figure 7.27). The sequences acquired walk into the last ORF in the GI, a gene named *yhhI*, annotated as encoding a ‘putative receptor’, however its COG assignment indicates that it encodes a member of the transposases (COG5433) and the Blastn analysis of the S115 SGSP-PCR amplicons SK# derived sequence showed that it also hits to predicted transposases and H repeat-associated protein genes in other *E. coli* genomes with 95% nucleotide identity. H repeats (‘H’ standing for Hinc as they contain *HincII* sites) are interesting in that they are found associated with some *Rhs* elements just downstream of the *ext/ds* region (see above in the *S. flexneri* analysis for some details on *Rhs* elements). They have similar features to IS, however Zhao *et al.*, 1993 found no transposition activity associated with the H repeats, but the authors believe they still have a role in site-specific recombination. Also the *yhhI* gene in *E. coli* K12 is described as encoding an ‘H repeat-associated protein, *RhsE*-linked, function unknown’. The gene is in a different location in the chromosome to that in Ss046, however the context is similar in both genomes, with the H repeat being downstream of an *Rhs* core ORF. Therefore the evidence shows that in the *S. sonnei* strains characterised in this study, as in *S. flexneri*, the D-arm of the *aspV* associated GIs are comprised of *Rhs* like elements, however the sequences are completely different; the *S. sonnei* *Rhs* sequences have identity to *E. coli* *Rhs* elements, whereas the *S. flexneri* *Rhs*-like sequences are unique.

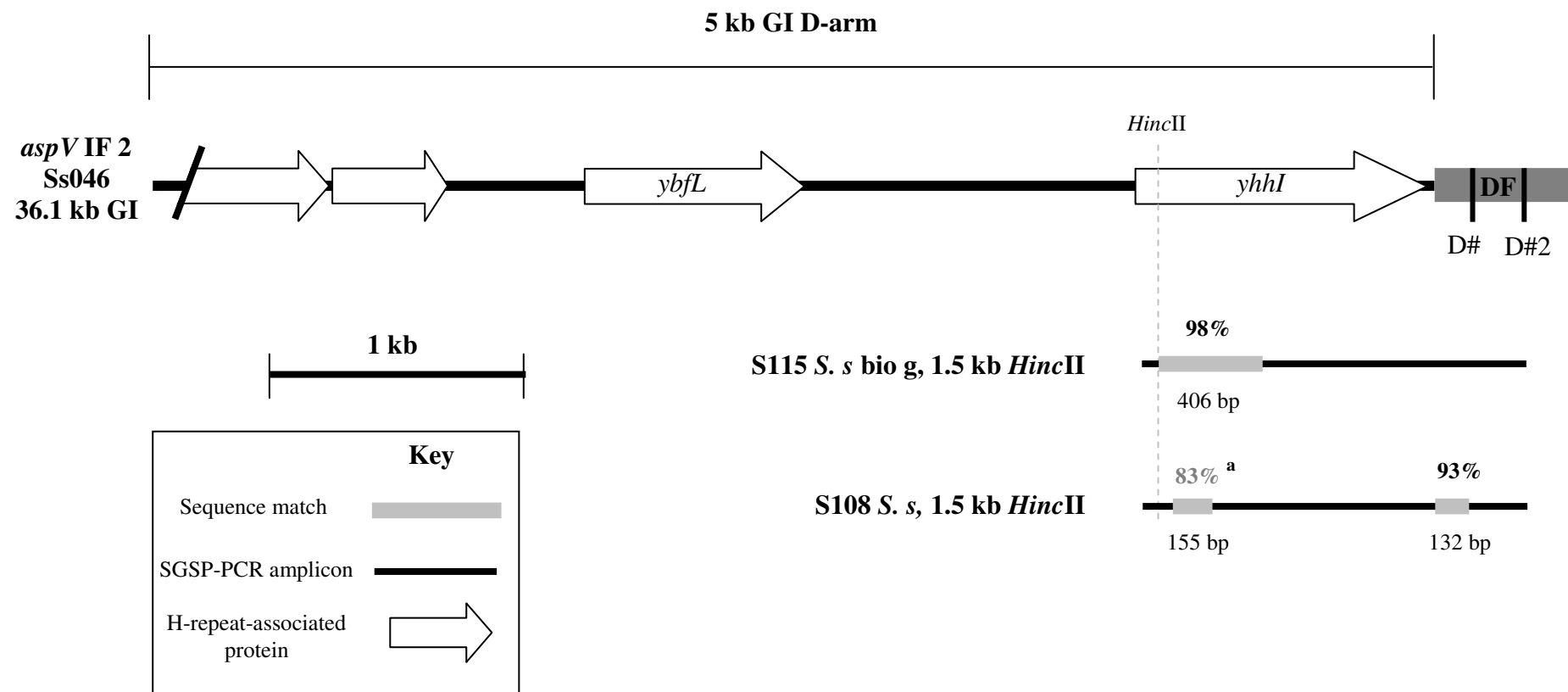


Figure 7.27. *aspV* D-arm SGSP-PCR results for the *S. sonnei* strains belonging to *aspV* island family 2

^a Sequence run failed, however the unclipped (low quality) sequences hit to the regions indicated

7.9.8 The *aspV* associated GIs are mosaic elements

The results from this study further highlight that this region is a hotspot for *Rhs* elements, as all EHEC O157, around half of the EHEC non-O157, EPEC, EAEC, ETEC (Hayashi *et al.*, 2001, Morabito *et al.*, 2003) and *S. enterica* subspecies I (Folkesson *et al.*, 2002) strains also harbour *Rhs* elements at the *aspV* locus.

Interestingly, *Rhs*-like elements are also associated with genes encoding Sci-like proteins in *S. flexneri*, *S. enterica* subspecies I, EDL933, *P. aeruginosa* and *Y. pestis* (Folkesson *et al.* 2002). However, both elements are found in different arrangements from species to species and both are only found associated with *aspV* in *S. flexneri*, *S. enterica* subspecies I and EDL933. Along with the presence of the *tau* gene cluster in Sf301, these results indicate that the *aspV* associated GIs are mosaic entities comprising a number of segments that have been acquired from different sources at different times and are most likely bacteriophage derived.

This mosaicism may be beneficial to the host strain as it may result in the generation of novel Sci-like effectors, which in turn may play an important role in the virulence of the host organism.

7.10 *thrW*

Table 7.5. SGSP-PCR results of the *thrW* tRIP negative strain-tRNA loci

<i>thrW</i> U# - T7# SGSP-PCR results									
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)							Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	<i>Eco</i> RV	<i>Hinc</i> II	
K12 MG1655					~0.4				
<i>S. dysenteriae</i> 3	S101	N ^a	N	N	N	N	~2.5 F ^b	N	
<i>S. dysenteriae</i> 9	S102	~0.9 ^c	~0.6 F	~1.9 F	N	N			769 [U#]
<i>S. dysenteriae</i> 6	S104	N	N	N	N	N			
<i>S. flexneri</i> 1a	S104	~0.9	N	~1.5	N	N			244 [SK#], 443 [U#]
<i>S. flexneri</i> 1b	S105	~0.9		~1.5					
<i>S. flexneri</i> 2a	S106	~2.1		~0.9					772 [SK#]
<i>S. flexneri</i> 2b	S107	N		~2.5					671 [SK#], 270 [U#]
<i>S. sonnei</i>	S108	~0.9	~3.5 F	~1.9	N	N			775 [SK#]
<i>S. flexneri</i> 6	S110	N	N	~2.5	N	N			461 [U#]
<i>S. flexneri</i> X	S111	N		~1.9					751 [SK#]
<i>S. flexneri</i> Y	S112	~0.9		~1.5					
<i>S. sonnei</i>	S113	N	~2.5 F	~1.2	N	N			227 [SK#], 1643 [U#] UC ^d
<i>S. sonnei</i> bio a	S114			N					
<i>S. sonnei</i> bio g	S115			~1.9					903 [U#]
<i>S. boydii</i> 1	S116	~0.9	N	~3.0 F	~1.6	N			192 [SK#]
<i>S. boydii</i> 2	S117			~3.0 F	~1.6				
<i>S. boydii</i> 3	S118			~1.0	~2.6				708 [SK#]
<i>S. boydii</i> 4	S119			~3.0 F	~1.6				
<i>S. boydii</i> 7	S120	N	~2.5 F	~1.2	N	N			791 [U#]

^a Indicates that no amplicon was generated^b The addition of 'F' after the text indicates that the amplicon was faint

^c Text highlighted in bold indicates that the amplicon was sequenced

^d The addition of 'UC' after the text indicates that the unclipped sequence was used in the analysis. The sequence run failed by the Phred20 (Ewing and Green, 1998) standards used by the sequencing company; however the low quality sequence still provided some meaningful information.

7.10.1 Diverse O-serotype converting bacteriophages are associated with *thrW* in *Shigella*

The *thrW* SGSP-PCR results show that all of the characterised *Shigella* strains harbour sequences found present on O-serotype converting bacteriophages, which have previously been reported to be associated with *thrW* in *Shigella*, also their functions are well characterised and a number of them have been completely sequenced (Adhikari *et al.*, 1999, Allison *et al.*, 2002, Allison and Verma, 2000, Casjens *et al.*, 2004, Clark *et al.*, 1991, Guan *et al.*, 1999, Mavris *et al.*, 1997) Table 5.1, Figure 7.28, Figure 7.29 and Figure 7.30 show that the GIs in the majority of the *S. flexneri* strains are most similar to the Sf301 and Sf2457T *thrW* associated prophages (*thrW*-IF1) and that the elements in most of the *S. boydii* strains, two of the *S. sonnei* strains and the *S. dysenteriae* 9 strain are most similar to the Sb227 *thrW* associated prophage (*thrW*-IF2). S113 (*S. sonnei*) and S120 (*S. boydii* 7) harboured sequences that were the most similar to the Sb227 prophage-like integrase gene, however, as they had a distinct restriction pattern (RP) and the integrase gene was truncated by mosaic IS elements, they were assigned to a separate island family – *thrW*-IF3 (see Figure 7.31).

S110 (*S. flexneri* 6 strain) had a distinct restriction pattern (see Figure 7.29), which was most similar to Sf2457T (*thrW*-IF1), however the sequence data had the highest nucleotide identity to Sb227. S110 was therefore overall designated as *thrW*-IF 2; however, the U-arm was given a split assignment (see Table 5.1). S110 is an interesting strain as it is more *S. boydii*-like in its tRIP profile and island content at most of the other tRNA loci than *S. flexneri*-like (see Table 5.1 and the corresponding details in each of the tRNA locus detailed sections). *S. flexneri* 6 strains have been shown previously to be *S. boydii*-like by analysis of their core DNA (Lan *et al.*, 2004) and there have been proposals for *S. flexneri* 6 to be re-classified as *S. boydii* (Petrovskaya and Bondarenko, 1977 and Petrovskaya and Khomenko, 1979).

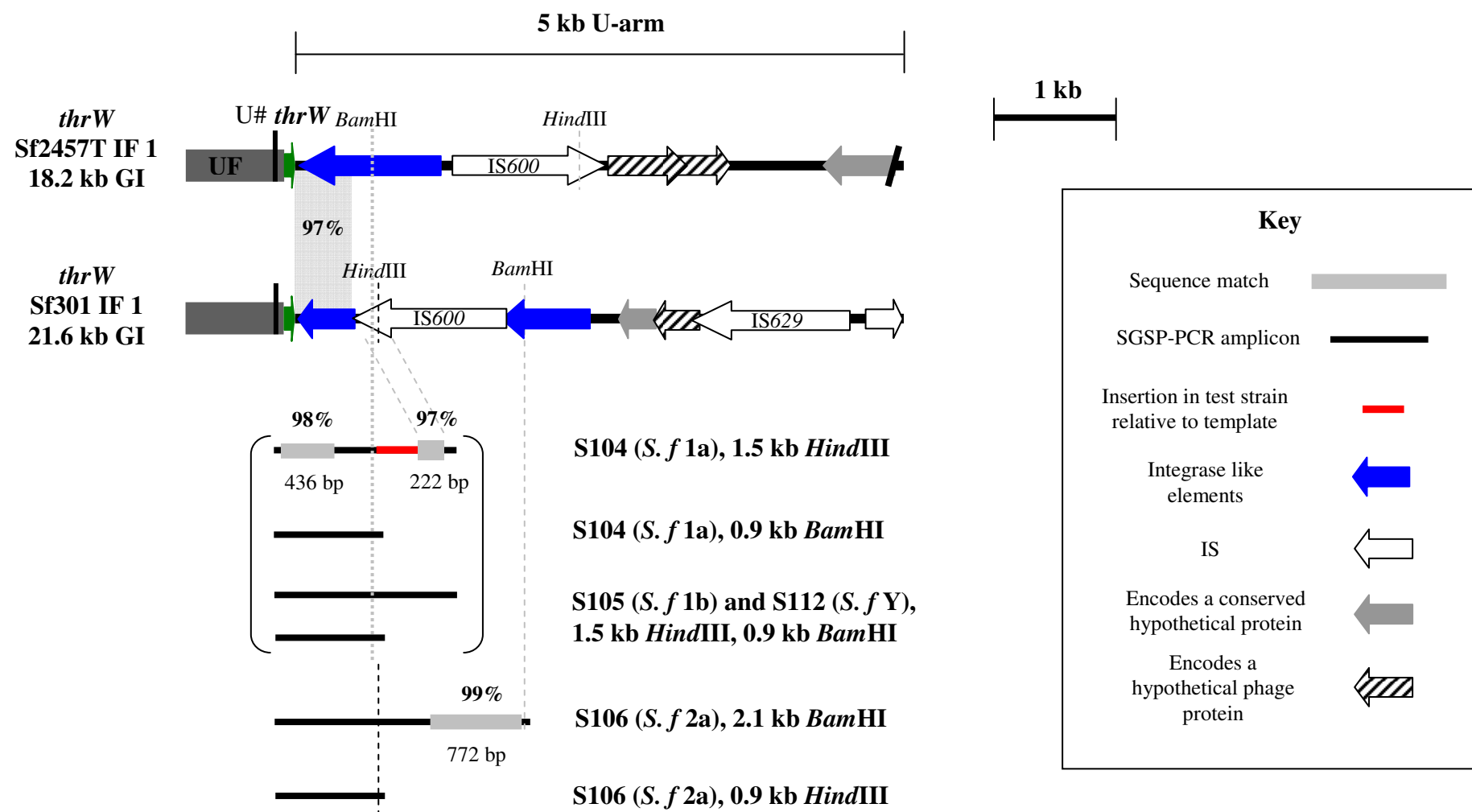


Figure 7.28. *thrW* U-arm SGSP-PCR results for the *S. flexneri* strains belonging to *thrW* island family 1

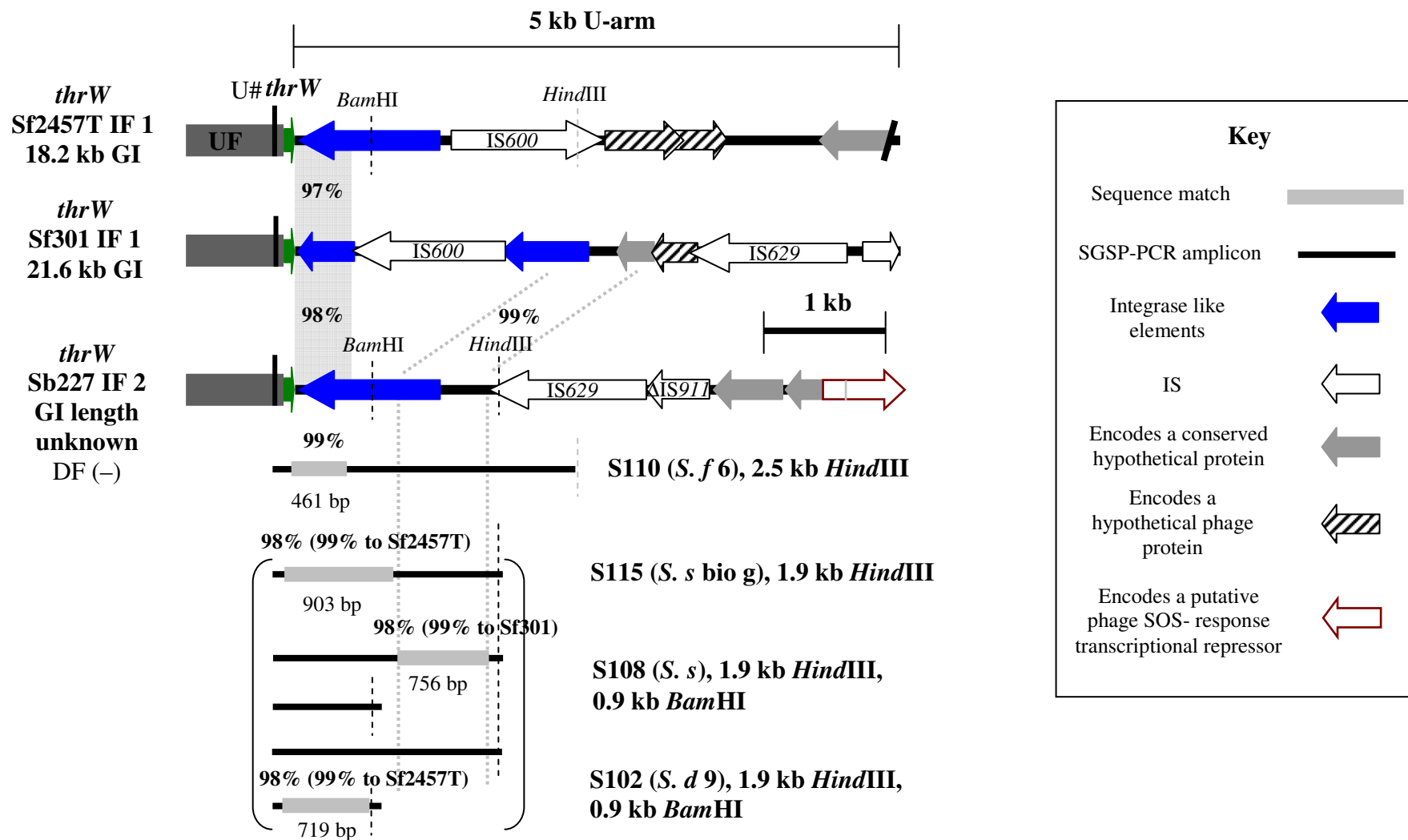


Figure 7.29. *thrW* U-arm SGSP-PCR results for the *Shigella* strains belonging to *thrW* island family 2

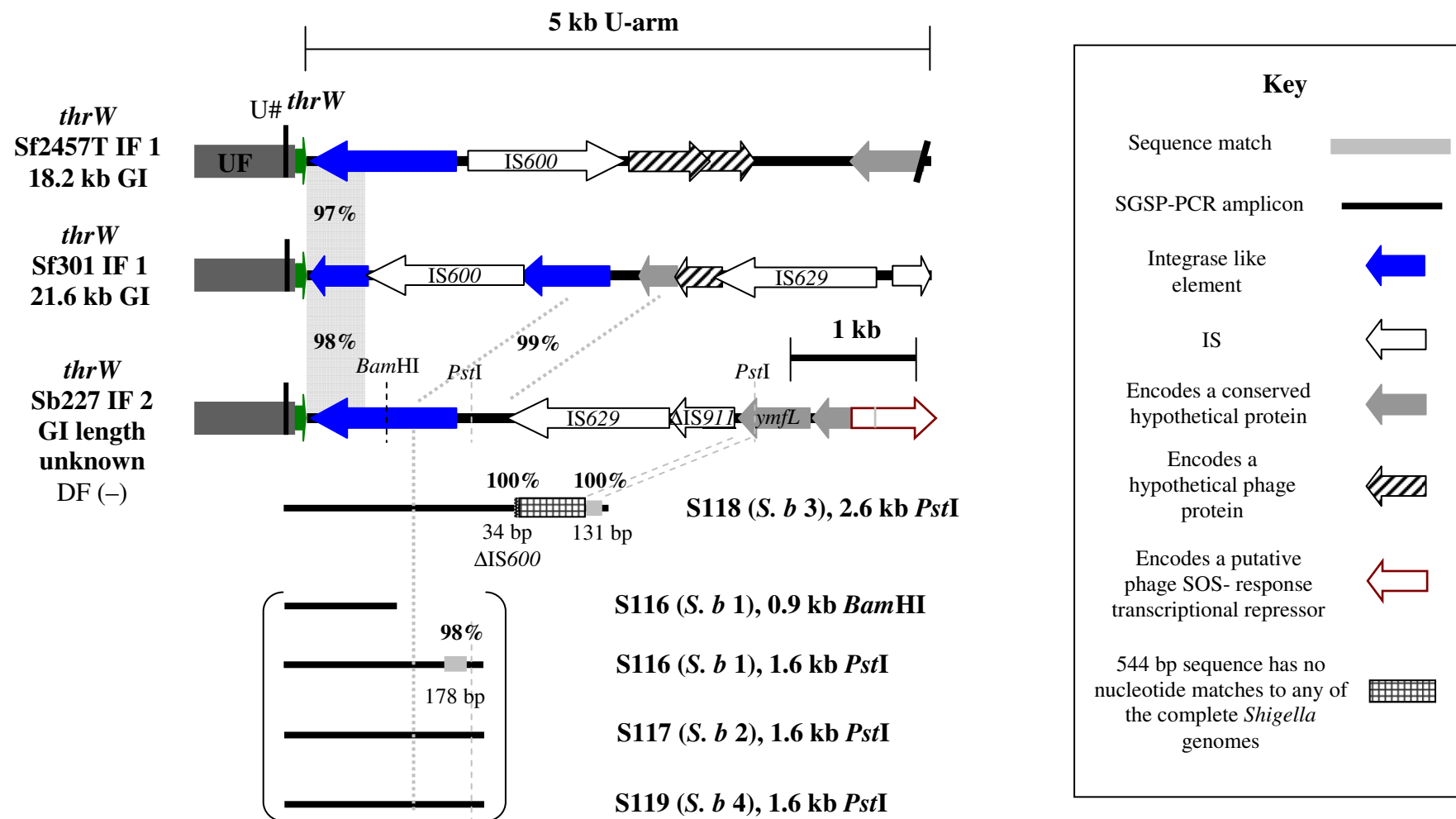


Figure 7.30. *thrW* U-arm SGSP-PCR results for the *S. boydii* strains belonging to *thrW* island family 2

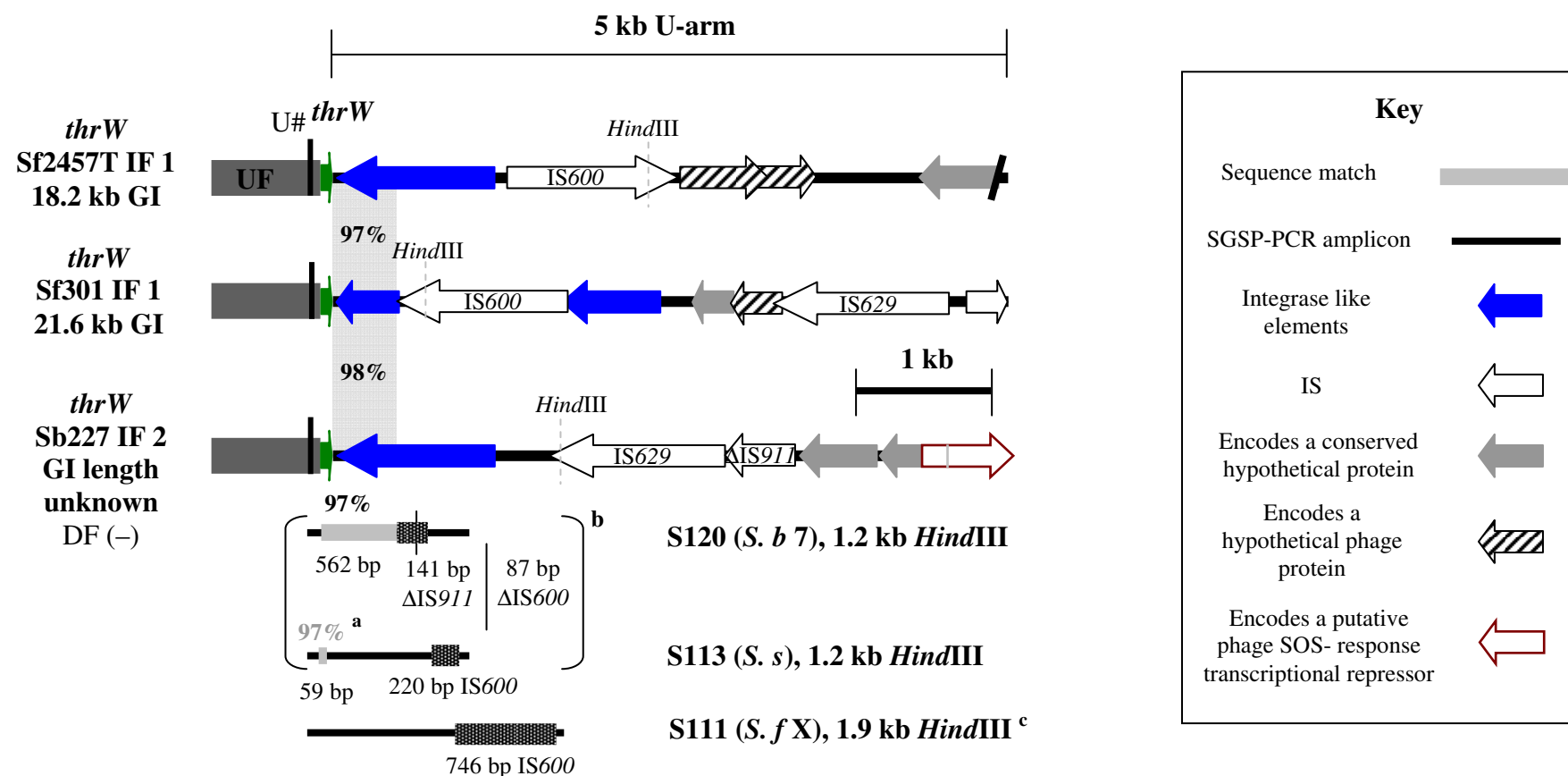


Figure 7.31. *thrW* U-arm SGSP-PCR results for the *Shigella* strains designated as *thrW* island family 3

^a Sequence run failed, however the unclipped (low quality) sequence hit to the region indicated

^b Designated as *thrW*-IF3

^c S111 was designated at 'unclassifiable' as only IS sequence data was obtained, however the restriction pattern (RP) and sequence data shows that it is similar to the *thrW*-IF3 strains.

7.10.2 The S118 (*S. boydii* 3) *thrW* prophage

Figure 7.30 shows that S118 harbours some sequence that has no nucleotide matches to any of the completely sequenced *Shigella* genomes. Further analysis indicated that this is phage-like sequence only found present in five other strains – four *E. coli* strains (K12 MG1655, W3110, CFT073 (UPEC) and B171 (EPEC [unfinished, GenBank accession number NZ_AAJX000000000]) and the unfinished *S. boydii* 18 BS512 strain (GenBank accession number NZ_AAKA000000000). In K12 MG1655 the sequence is found present on the e14 defective lambdoid prophage (see Figure 7.32), which is not associated with any tRNA locus, but interestingly is known to harbour sequences similar to the *thrW* associated *S. flexneri* phage SfV (Mehta *et al.*, 2004).

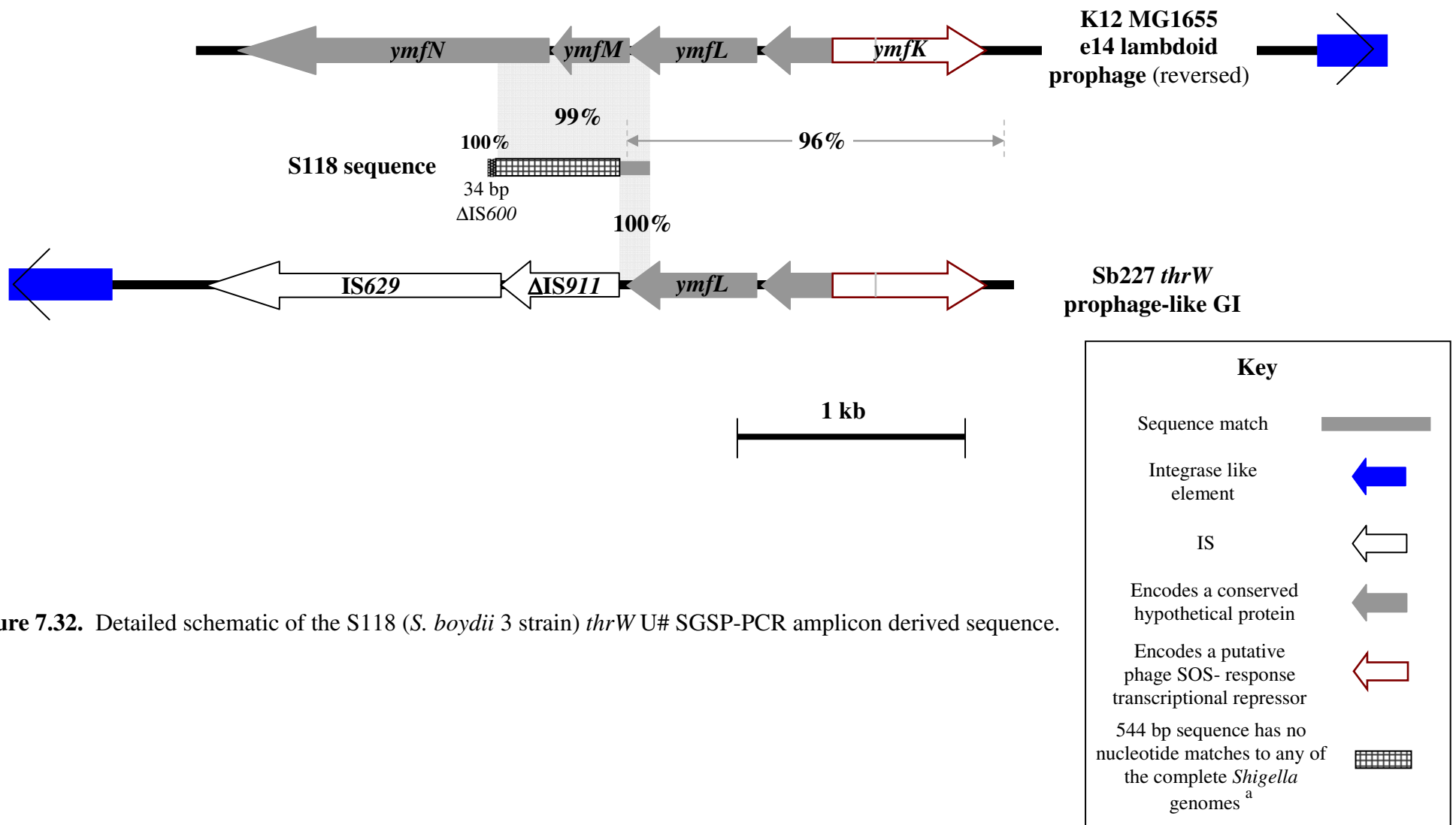


Figure 7.32. Detailed schematic of the S118 (*S. boydii* 3 strain) *thrW* U# SGSP-PCR amplicon derived sequence.

7.10.3 The S107 (*S. flexneri* 2b strain) *thrW* associated mosaic prophage

Figure 7.33 shows the sequence data obtained from the S107 *thrW* GI, the results indicate that this is a novel prophage-like element distinct to all of the other characterised strains, with a P4 integrase-like gene that has 97% nucleotide identity to an integrase-like gene that is present 6.3 kb into the Sf301 *thrW* associated GI. In addition, 671 bp of sequence was acquired downstream of the integrase that is not found in any of the sequenced *Shigella* strains, but is a mosaic of sequences that are found in other Enterobacteria lambdoid-like bacteriophages, that in turn insert at other locations in their respective host chromosomes. This GI was assigned to a unique family – *thrW*-IF4.

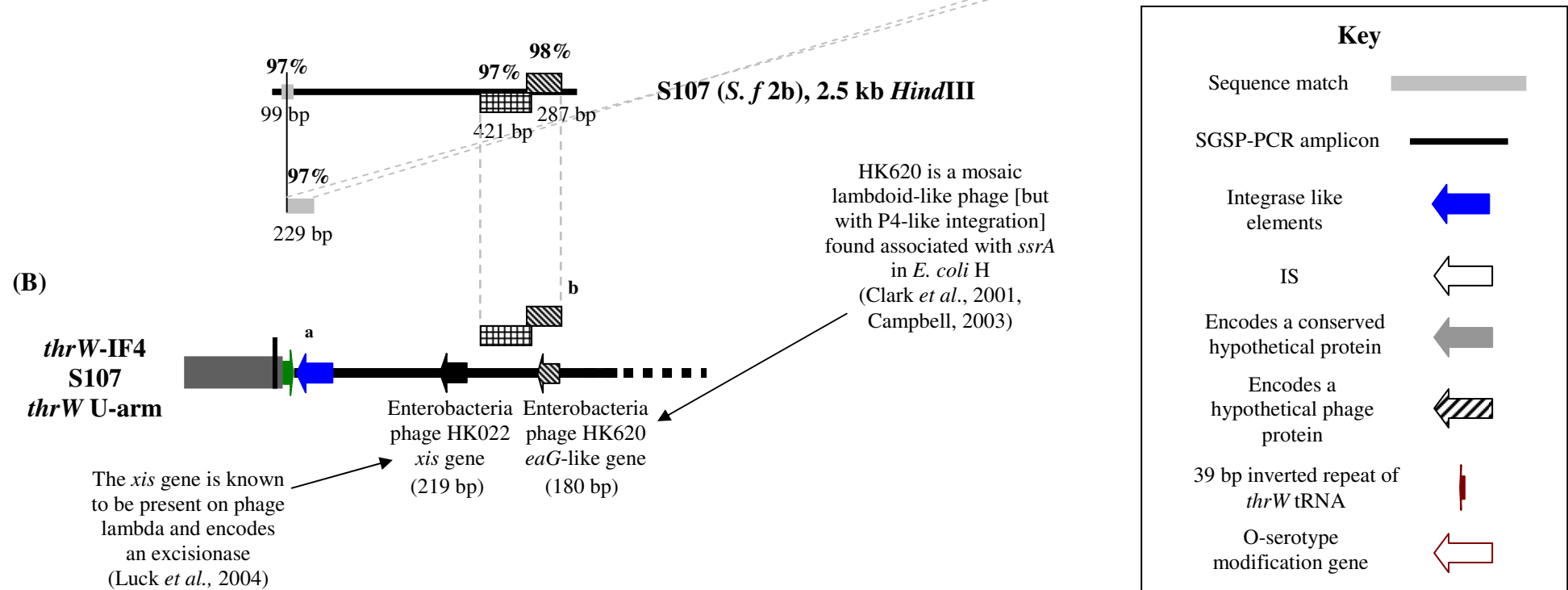
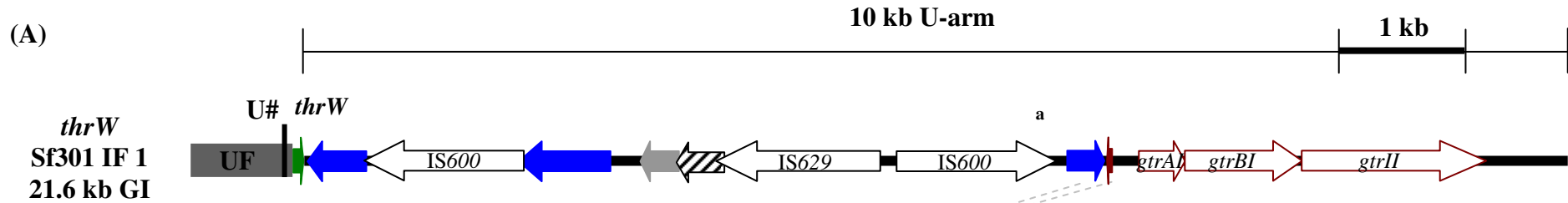


Figure 7.33. The S107 (*S. flexneri* 2b strain) *thrW* U# SGSP-PCR amplicon derived mosaic phage-like sequences

A. Alignment of the S107 U# generated SGSP-PCR amplicon and its corresponding sequence with the Sf301 genome

B. An Inferred diagram of the S107 *thrW* U-arm from the sequence data acquired.

^a The integrase gene fragment found 6.3 kb into the Sf301 *thrW* GI, adjacent to a 39 bp inverted repeat of the *thrW* 3' terminus, is found directly downstream of the *thrW* locus in S107 in the opposite orientation.

^b These sequences are not found in any of the sequenced *Shigella* genomes.

7.10.4 The *Shigella thrW* associated integrase genes are found in the opposite orientation

The integrase genes found associated with all of the tRNA loci in this study are P4-like. Previous studies have shown that in most cases, P4-like phage integrase genes are orientated in the same direction as the tRNA gene; this is due to the insertion point of the phage DNA in the tRNA gene (Campbell *et al.*, 2003); this pattern was also seen across all of the strain-tRNA loci in this study apart from at *thrW*. The *Shigella thrW* prophages are distinct in that all of the integrase genes found directly adjacent to the *thrW* gene are present in the opposite orientation to the tRNA locus (Figure 7.28 to Figure 7.33); this is the orientation that lambdoid-like integrase genes are usually found in (Campbell *et al.*, 2003). The *Shigella* O-serotype converting prophages have been previously shown to be mosaics of P4 and lambdoid-like elements (Allison *et al.*, 2002), with the results of this study further highlighting their diversity. However, the reason for the P4-like *int* genes being orientated in the opposite direction is yet to be determined.

These results indicate that the *thrW* locus in *Shigella* is one of the major hotspots for prophage activity; the associated prophages are a highly variable, mosaic family of elements, consisting of sequences found in both P4 and lambdoid-like bacteriophages found across a range of Enterobacterial hosts. This mosaicism is likely to shuffle and disrupt the O-serotype conversion genes found on these *thrW* specific phages; these events may promote the generation and expression of new O-serotypes in *Shigella* and could therefore be a key factor in the virulence of *Shigella*.

8.0 Discussion

8.1 tRIP and SGSP-PCR as strategies for GI detection and characterisation

The results of this study show that tRIP is a powerful tool to detect tRNA associated GIs across *Shigella* strains, as all of the characterised tRIP negative strain-tRNA loci were found to harbour island DNA (see section 8.2). So far, this indicates that the chance of a negative tRIP result at any given strain-tRNA locus being due to the presence of island DNA is 100%. The predictive power of the tRIP screen across *E. coli* strains is therefore very strong.

Using SGSP-PCR, I was able to characterise 81% of the putatively GI occupied strain-tRNA loci, indicating that this is a robust, cost effective strategy for island characterisation and that tRIP screening followed by SGSP-PCR should be applicable to a broad range of microorganisms.

8.2 GI diversity across *Shigella*

39% of the known elements that were characterised across *Shigella* match to islands present in the Islander database (see section 1.7), indicating that they are archetypal, prophage derived integrative elements.

In addition, at least 54% of the elements identified harbour phage-like integrase genes (see Figure 5.2), this strongly supports the notion that many of these elements arose following acquisition of horizontally transferred integrative GIs. The frequent presence of integrase genes also highlights the potential role of bacteriophages in the original and/or ongoing dissemination of island DNA in *Shigella*.

Through analysis of their GC contents compared with the surrounding flanking core DNA; elucidation of the putative functions of their encoded genes and analysis of the context of the sequences relative to other known GIs and relative to the surrounding DNA, the remaining tRIP-defined elements (see Table A2. 6) were all confirmed to be genomic islands/islets.

Many of the elements also contained IS elements and phage-like sequences, which are frequently associated with island DNA.

A total of thirty five distinct island families were found across the strains tested in this study, fifteen are Islander defined elements and twenty were defined by tRIP only (see Table 5.2). This indicates that many of the GIs found across *Shigella* may be more ancient, mosaic elements that are likely to be locked into the chromosome.

However, certain loci stand out as being hotspots for prophage activity; these are *argW*, *thrW* and *leuX* and the elements found associated with these loci exhibit considerable diversity across *Shigella* compared to other sites, indicating that these elements are drivers of genome plasticity in *Shigella*. This also coincides well with the data generated by (Fouts, 2006) in a study across 302 complete bacterial genomes, who found that *arg*, *leu*, *ser* and *thr* loci are the most popular tRNA sites for phage insertion. Future GI discovery experiments across *E. coli* should therefore be targeted primarily at these loci.

One completely novel GI was discovered, the *argW* associated prophage-like GI in S116 (*S. boydii* 1 strain) (see section 6.2), indicating that *Shigella* as a group has an open ‘pan-genome’ (Medini *et al.*, 2005), meaning that as more strains are characterised, more novel DNA is discovered. In the future, as more and more strains are characterised, the amount of novel DNA discovered may begin to plateau, if this was to occur, it would indicate the relative size of the *Shigella* pan genome and may even show that its pan genome is indeed finite. However, so far, the evidence from collective sequencing of *Shigella* strains indicates that its pan genome is open. This result also further highlights the role bacteriophages play in the the generation of diversity across *Shigella*.

Even so, the diversity across *Shigella* is relatively low compared to other Enterobacteriaceae; as a similar study across only ten *E. coli* human bloodstream infection isolates found at least seven GIs that contained sequences novel to *E. coli*, with six harbouring stretches of sequence

without any matches in the entire DNA database (K. Rajakumar, unpublished data). This shows that while *Shigella* has a plastic genome, it is a highly specialised human pathogen that has undergone considerable pathoadaptive genome reduction. The location of IS elements within many of the *Shigella* islands, when compared to similar elements in other *E. coli* strains, suggests that IS are playing a key role in the deletion of redundant island DNA in *Shigella*.

The diversity of islands across the *S. flexneri* and *S. sonnei* strains indicates that they are clonal pathogens, whereas the *S. boydii* and *S. dysenteriae* strains are more diverse, indicating that they have arisen independently from multiple origins; in particular S120 (*S. boydii* 7 strain), which has a distinct tRIP profile and island complement to the other *S. boydii* strains (see Table 3.2 and Table 5.1). Interestingly the *S. dysenteriae* strains were *S. boydii*-like in their GI content and tRIP profiles (see Table 3.2 and Table 5.1), this coincides with the findings by Ruiting Lan and colleagues on the evolutionary relationships of *Shigella* based on their core DNA and virulence plasmid forms, which shows that *S. boydii* (including *S. flexneri* 6) and *S. dysenteriae* strains are often found in the same cluster, whereas *S. dysenteriae* 1 is an outlier, being of a separate lineage (Lan *et al.*, 2004).

8.3 Key genomic islands identified across *Shigella* in this study

In addition to the previously well characterised GIs found across the chromosome of *Shigella* (see Table 1.3) there were some key islands identified in this study that encode genes that I believe could be playing a significant role in the virulence and/or transmission of virulence properties across *Shigella*, these are described in detail in section 7.0 and are summarised below in Table 8.1.

Table 8.1. Key *Shigella* islands identified in this study

tRNA locus	Island family	Associated strains	Key encoded genes	Functions	Details
<i>serU</i>	<i>serU</i> -IF1, <i>serU</i> -IF2	<i>S. flexneri</i> , <i>S. boydii</i> , <i>S. dysenteriae</i>	<i>ipaH</i>	Encodes an IpaH protein that along with other chromosomally encoded IpaH proteins are secreted by the virulence plasmid encoded T3SS and modulates the host inflammatory response. It is also an independently mobile gene.	Section 7.5 and section 7.7
<i>aspV</i>	<i>aspV</i> -IF1	<i>S. flexneri</i>	<i>sci</i> operon	May encode a novel secretion system and effector molecules	Section 7.9.3
<i>leuX</i>	<i>leuX</i> -IF3	<i>S. sonnei</i>	<i>fec</i> operon	Encodes a ferric dicitrate uptake system (siderophore). It is also an independently mobile operon that was previously identified on the SRL PAI that is associated with <i>serW/serX</i> across all four <i>Shigella</i> species	Section 7.1.3
	<i>leuX</i> -IF1	<i>S. boydii</i> , <i>S. dysenteriae</i>	<i>sigA</i>	Encodes a serine protease that is cytopathic and enterotoxic (member of SPATE family). It is also an independently mobile gene that was previously identified on the <i>she</i> PAI that is associated with <i>pheV</i> in <i>S. flexneri</i> .	Section 7.2

8.4 Independently mobile virulence determinants in *Shigella*

Table 8.1 shows that the GI encoded *ipaH* gene(s), *fec* operon and *sigA* gene have been found harboured in different locations in the chromosome of *Shigella*, suggesting that they are independently mobile and can move within the chromosome of the host bacteria. Each of these virulence determinants are flanked by IS elements, indicating that IS are responsible for their independent mobility.

The *fec* operon and *sigA* gene are of particular interest because they are both harboured on distinct GIs that are associated with distinct tRNA loci, therefore increasing their likelihood of transmission across *Shigella* strains via HGT. This also shows how mobile genetic elements like IS are potentially hugely influential in the transmission of virulence factors, as they can mobilise regions of DNA from one GI to another. If the recipient GI excises at a high frequency and/or it has broad host range, the corresponding virulence genes could be transmitted rapidly across the population and even between species. As the *fec* operon and *sigA* are both key virulence factors of *Shigella* (see Table 1.3), knowledge on their distribution across many more *Shigella* strains could be very useful in determining how ‘promiscuous’ they are, what chromosomal locations they preferentially target and in pinpointing possible emergent endemic strains.

In future experiments, DNA Probes specific to the *sigA* gene and genes of the *fec* operon could be used in screening Southern hybridisations on the genomic DNA of hundreds of strains to find out the distribution and copy number of these two elements across *Shigella*. This could be followed up by an SGSP-PCR-like strategy, using specific primers internal to these elements to walk outwards into the surrounding DNA; followed by sequencing to determine the context and location of the elements in the *Shigella* chromosome.

8.5 The *aspV* associated ‘*sci*’ island could encode novel virulence factors

The *sci* operon, harboured on the *aspV* associated GI, which is found in most of the *S. flexneri* strains probed in this study (*aspV*-IF1, see Table 5.1), could encode novel species-specific virulence determinants. This is because the *sci* genes encode putative proteins with homologies to either known or putative virulence factors that are found in various other Gram-negative pathogens that can also manipulate or invade eukaryotic cells (see section 7.9.3 for details).

It would therefore be interesting to see if deletion of this operon or indeed the entire GI affects the virulence of *S. flexneri*, such as reducing its ability invade the epithelial cells of colon. This could be tested *in vitro* by deletion of the GI via allelic exchange and the invasiveness of the mutant *Shigella* strain compared with the wild-type by its ability to enter HeLa cells (cytotoxicity assay) or by using the Sereny test. The rabbit ileal loop test could also be utilised to compare the enteropathogenicity of cell cultures, cell extracts and the culture supernatant from the mutant and wild-type strains; to determine whether the *sci* genes encode secretion apparatus and/or secreted toxins.

8.6 Improvement of the tRIP screen

The tRIP screen could be improved by incorporating a multiplex PCR approach, using a selection of the tRIP primer pairs in a single PCR. This would reduce costs and speed up the screening process so that hundreds of strains could be screened in a short period of time. The multiplex tRIP screen could be ‘tailor made’ to focus on tRNA loci of particular interest and would be useful to assess the prevalence of specific GIs across a panel of strains. It could also be used to quickly classify a strain, based on its putative GI content; as analysis of the tRIP profiles across *Shigella* indicated that the profiles correlated strongly with the respective ‘species’ of each *Shigella* strain (see Table 3.1). This classification method however, may

only be applicable to highly similar strains such as *Shigella*, as diverse groups of organisms are likely to yield more diverse tRIP profiles.

8.7 The power of tRIP when used with complementary *in silico* GI discovery methods

All of the known elements characterised by SGSP-PCR are also defined as GIs by a software tool called tRNAcc (for tRNA gene content and context analysis) developed by our group. tRNAcc is Mauve-facilitated multigenome comparative strategy, where a GI is described as the anomalous segment between the 3' end of the tRNA gene and the 5' end of the corresponding conserved downstream flanking region. tRNAcc selects all strain-specific segments of greater than 1 kb and therefore includes islands with missing or damaged integrase genes, whereas Islander excludes GIs of this nature. All of the elements identified by tRNAcc had many strain-specific sequences, significant dinucleotide biases and/or anomalous GC contents, indicating that they are horizontally acquired. tRNAcc was utilised to detect GIs across a number of sequenced *E. coli*, *Shigella*, *S. enterica* and *P. aeruginosa* strains and is therefore a powerful tool for the rapid *in silico* discovery of GIs across sequenced strains; also the software can be used to select UF and DF specific primers for the interrogation of test strains using tRIP. tRNAcc therefore enables the early and high-throughput discovery of GIs across large numbers of bacterial strains (Ou *et al.*, 2006).

The tRNAcc method has since been developed to integrate the ArrayOme software package (Ou *et al.*, 2005), which is used to compare the microarray-visualised genome (MVG) size of a strain with its pulsed-field gel electrophoresis (PFGE) measured chromosome size (an accurate measure of its actual chromosome size). The difference between the PFGE chromosome size and MVG is therefore the size of the novel mobile genome (or 'mobilome') associated with that strain. Strains that are found to have a potentially large mobilome can then be screened by tRIP following the design of specific UF and DF primers using tRNAcc.

The integrated package known as MobilomeFINDER is now available as an online GI discovery tool (<http://mml.sjtu.edu.cn/MobilomeFINDER>). MobilomeFINDER was used to compare the genomes of various species including nine *Escherichia coli* genomes, four *Salmonella enterica* genomes, two *Klebsiella pneumoniae* genomes and two *Streptococcus suis* genomes. The tool aids in the identification of mobilome rich strains and facilitates the high-throughput identification and characterisation of GIs across a broad range of bacterial strains (Ou *et al.*, 2007).

8.8 Future work

- Southern hybridisation and primer walking experiments on hundreds of strains to determine the prevalence and distribution of the independently mobile virulence factors *sigA* and the *fec* operon across *Shigella*.
- Deletion of the *S. flexneri sci* island to determine its effects on the virulence of *Shigella*.
- Improvement of the tRIP screen by using a multiplex PCR approach, followed by the screening and SGSP-PCR characterisation of hundreds of strains, focussing initially on the key hotspots, *serU*, *leuX*, *argW* and *thrW*.
- Complete sequencing of the *argW* associated novel *S. boydii* 1 prophage-like GI to determine its putative functions.
- Characterisation of the remaining uncharacterised strain-tRNA loci across the strains screened in this study.

8.9 Conclusion

The major development from this study is the evidence that a number of key *Shigella* virulence determinants are independently mobile and not only localised to single families of islands; this significantly increases their potential to spread by HGT across *Shigella* and could contribute to the rapid emergence of new endemic strains. The probing of many more *Shigella* strains using high-throughput tRIP-like approaches will contribute significantly to our knowledge of the mechanisms behind the virulence and the spread of virulence determinants across this pathogen and will hopefully unveil more putative virulence factors; which may in turn be viable drug targets or will facilitate the construction of improved vaccine candidates.

A1.0 Appendix 1

All media were purchased from Sigma unless otherwise stated in parentheses.

LA

5 g Bacto-tryptone (Difco)

2.5g Bacto-yeast extract (Difco)

5 g NaCl

7.5 g Bacto-agar (Difco)

Made up to 500 ml in ddH₂O

Autoclaved at 121 °C at 15 psi for 15 min

LB

5 g Bacto-tryptone

2.5g Bacto-yeast extract

5 g NaCl

Made up to 500 ml in ddH₂O

Autoclaved at 121 °C at 15 psi for 15 min

TE buffer

10 mM Tris-HCl

1 mM EDTA

Made from 1 M stock of Tris-HCl (pH 7.5) and 500 mM stock of EDTA (pH 8.0) which were both autoclaved at 121 °C at 15 psi for 15 min

CTAB/NaCl solution

0.7 M NaCl

10% (w/v) CTAB in sterile nH₂O

6 x loading dye

11 mM EDTA

3.3 mM Tris-HCL

2.5% (w/v) Ficoll 400

0.0017 % SDS

0.15 % Orange G

pH 8.0

SOC

4 g Bacto-tryptone

1 g Bacto-yeast extract

0.1 g NaCl

Made up to 200 ml in ddH₂O

Autoclaved at 121 °C at 15 psi for 15 min

After autoclaving add to 10 ml of medium:

50 µl of filter sterilised 2 M MgCl₂

200 µl of filter sterilised 1 M glucose

6% sucrose agar

5 g Bacto-tryptone (Difco)

2.5g Bacto-yeast extract (Difco)

30 g sucrose

7.5 g Bacto-agar

Made up to 500 ml in ddH₂O

Autoclaved at 121 °C at 15 psi for 15 min

6% sucrose broth

5 g Bacto-tryptone (Difco)

2.5g Bacto-yeast extract (Difco)

30 g sucrose

Made up to 500 ml in ddH₂O

Autoclaved at 121 °C at 15 psi for 15 min

50 x TAE buffer

2 M Tris-HCl

2 M Acetic Acid

5 0mM EDTA

2 x wash solution

2 x SSC (1:10 dilution of stock 20 x SSC)

0.1% (v/v) SDS (1:100 dilution of stock 10% SDS)

0.5 x wash solution

2 x SSC (1:40 dilution of stock 20 x SSC)

0.1% (v/v) SDS (1:100 dilution of stock 10% SDS)

20 x SSC

3 M NaCl

0.3 M sodium citrate

pH 7.0

Autoclaved at 121 °C at 15 psi for 15 min

Blocking solution stock (Roche)

10% (w/v) in maleic acid buffer

Autoclaved at 121 °C at 15 psi for 15 min

Blocking solution

1% (v/v) Blocking Solution in maleic acid buffer (1:10 dilution of stock 10% Blocking Solution)

Detection buffer

100 mM Tris-HCl

100 mM NaCl

pH 9.5

Autoclaved at 121 °C at 15 psi for 15 min

Maleic acid buffer

0.3 M maleic acid

150 mM NaCl

pH 7.5

Autoclaved at 121 °C at 15 psi for 15 min

N-lauroylsarcosine stock

10% (w/v) in nH₂O

Filter sterilised

Prehybridisation solution

5x SSC

0.1% (v/v) N-lauroylsarcosine (1: 100 dilution of stock 10% N-lauroylsarcosine)

0.02% (v/v) SDS (1:500 dilution of stock 10% SDS)

1% (v/v) Blocking Solution (1:10 dilution of stock 10% Blocking Solution)

SDS stock

10% (w/v) in nH₂O

Filter sterilised

Washing buffer

0.3% (w/v) Tween 20 in maleic acid buffer

A2.0 Appendix 2

A2.1 Primers used in this study

Table A2. 1. tRIP and SGSP-PCR Primers

Primer name	Sequence (5'-3')	T _m (°C)	U-D tRIP screen annealing temperature (°C)	Length of tRIP product in control strain ^a	SGSP with T7# start annealing temp (°C) ^b	Notes
<i>serU</i> U	TCCAGGGCCACTTAATCATCGTT	58.4			68.4	
<i>serU</i> D	TTGCACCACGAAAATCATCTCAT	56	55.5	1790 bp	66	
<i>aspV</i> U	TTGCGGTGGCGAGGAAAATGTT	62.2			68.4	
<i>aspV</i> D	GGTGACAGCCGGGTGATTA	53	54.6	3120 bp	63	
<i>aspV</i> U 2	GCTTAAGCGCGATATTCCGAAGAC	59.3			68.4	
<i>aspV</i> D 2	GCCGCTGGTGTGCTACGACTTAC	59.7	56.9	3375 bp	68.4	Used for SGSP-PCR only
<i>asnV</i> U 1	GCCCCGGCATAACAAATAATAAAAA	55.6			65.6	
<i>asnV</i> D 1	CGAGAAACCCCGCGTAACTGG	60.3	57.1	1978 bp	68.4	
<i>asnV</i> U 2	AAGTGCCGCCATTACTTACAACCAG	59.4			68.4	
<i>asnV</i> D 2	CCATTGCCGGTAACCCCATCTTT	61.6	56.6	564 bp	68.4	

<i>serT</i> U	GCACTTTTGGCTGTTTTTCA	50.6	54.1	306 bp	60.6
<i>serT</i> D	TTTACCCATCTTTACGCATTTG	51.6			61.6
<i>serW</i> U	GGAGTAATGTGCCGAACCTGT	53.1	56.1	1422 bp	63.1
<i>serW</i> D	CACCGATGCGATGGAAGAGAT	56.3			66.3
<i>metV</i> U	TAAGGCGCAACGAAGATAACAAAC	56.6	58.6	1379 bp	66.6
<i>metV</i> D	CCGGCCAATGCACAGGATA	56.5			66.5
<i>pheU</i> U	GAAACGCAAACCGCCGAACAAAA	63.6	60.2	749 bp	68.4
<i>pheU</i> D	CACGGGGCCGCACGACATT	63.3			68.4
<i>glyU</i> U	ATGGCGAATTAATCAGCAGTCAGC	58.4	55.8	838 bp (CFT073)	68.4
<i>glyU</i> D	TCCGGGATTATTGTCTGCAGTAGTT	57.8			67.8
<i>argW</i> U	TCTGGCCCTTCGCACTACCTACTT	59.3	58.3	1774 bp (Sb227) ^c	68.4
<i>argW</i> D	GCCCGGGCATCAGCAGACATA	61.6			68.4
<i>thrW</i> U	TGACGCATCGCCCGGTTAGTTT	62.2	54.9	> 5 kb ^d	68.4
<i>thrW</i> D	ACGTCTGCGGTTYGGTGGAGTTT	62.6			68.4
<i>serX</i> U	CAAAGRCCACCAGCATAACAAATC	58.5	59.5	903 bp	68.4
<i>serX</i> D	TTCCCCTCGCCCWAACAGACG	59.9			68.4
<i>asnT</i> U	AGGTTGCTGGCTGGGAACACGAT	62.6	56.9	> 5 kb ^d	68.4
<i>asnT</i> D	ACTGGCAACCTGATAACCGACTCCA	61.6			68.4

<i>asnT</i> U 3	GCCCCAGAACTTTTTGCTCCTCG	61.9	56	> 5 kb ^d	68.4	used for SGSP-PCR only
<i>asnT</i> D	as above	as above			as above	
<i>pheV</i> U	CCGGATTACGCATCTGTGGCATT	61.6	60.2	> 5 kb ^d	68.4	
<i>pheV</i> D	GCGGCGCGTTTTATTCACTGGT	61.9			68.4	
<i>selC</i> U	CCTTGATGCTATAGGGGTGCTGAGA	59.5	57.2	2850 bp	68.4	
<i>selC</i> D	CAATYAGCGTTGAGGGATAGGTGGT	58.5			68.4	
<i>leuX</i> U	CACCACTTTATCGGCACCCATCG	62	60.1	4.5 kb	68.4	
<i>leuX</i> D	GGAGGCCCGYCATGTCACCTT	61.8		(S110)	68.4	
<i>ssrA</i> U	CCGTACCCGCAAGTTACTTCTCAA	58.4	57.5	4640 bp (Sf301)	68.4	
<i>ssrA</i> D	AGGGKTACTCGATGGCGGTCTATA	58.9			68.4	

^a Using K12 MG1655 as the template unless otherwise stated in parentheses

^b SGSP-PCRs were performed using hot-start ‘touchdown’ PCR, where the initial annealing temperature is reduced by 1 °C per cycle for 10 cycles, then run for a further 20 cycles at the touchdown annealing temperature (see section 2.8.2 for methodology)

^c Can also use S119 (*S. boydii* 4 strain) as the positive control at this locus, produces a tRIP amplicon of 1.6 kb

^d All known strains produce tRIP amplicons over 5 kb

Table A2. 2. Other primers used in this study

Primer name	Sequence (5'-3')	T _m (°C)	Restriction site incorporated
pBluescript KS II (+) primers			
M13 (-20) F	GTAAAACGACGGCCAGT	54	
M13 R	GGAAACAGCTATGACCATG	53.5	
T3	AATTAACCCTCACTAAAGGG	53.2	
T7	GTAATACGACTCACTATAGGGC	58.4	
SK	CGCTCTAGAAGTAGTGGATC	57.3	
KS	TCGAGGTCGACGGTATC	55.2	
<i>serU</i> novel sequence primers			
S102novelseqF	CGAGCTTGCGAAAACTCTGAGC	59.3	
S102novelseqR	TGCCACGTTGCCAATCAGAC	60.3	
<i>aspV</i> Southern hybridisation primers			
<i>aspV</i>Uprobe F	CGGCGCGGAGTTGGTGAT	59.1	
<i>aspV</i>UprobeR	AAAGCCATCGACGTTTGACCACC	60.9	

<i>int</i> -PCR primers			
P4I_F	GAATTACCGGACTGACCCTTC	54.4	
P4I_R	GAAGGGTCAGTCCGGTAATTC	54.4	
P4III_F	TCGCTCMAGTGAAGTGC	52.6	
P4III_R	CGCAGTTCAC ^T KGAGCGA	52.6	
<i>leuX</i> allelic exchange primers			
<i>leuX</i> UFF	CAGCTCTAGAGGCGTGCCGGTAGC	64.2	<i>Xba</i> I
<i>leuX</i> UFIF	GCCGAGAAGATGCATGCGGACTGG	62.5	<i>Nsi</i> I
<i>leuX</i> UFR	ACGTTCTAGATTATACCTGTGCGCACGC	59.7	<i>Xba</i> I
<i>leuX</i> UFIR	CCAGTCCGCATGCATCTTCTCGGC	62.5	<i>Nsi</i> I
<i>leuX</i> U 2	GGTTGTCGGCGCAACCTTGC	57.9	
<i>leuX</i> tRNArev	TCAACTGCGTCTACCGATTTCG	54.8	
<i>argW</i> allelic exchange primers			
<i>argW</i> UFF	TTGATCTAGACAGCGACGAAATTG	54	<i>Xba</i> I
<i>argW</i> UFR	CGCTTCTAGACTGCGGGGTAAGTACG	62.7	<i>Xba</i> I
<i>argW</i> U 2	CGCTACGCTTTAGCTATACG	51.8	
<i>argW</i> tRNArev	CTGCAATTAGCCCTTAGGAGG	54.4	

ori	AGGAACACTTAACGGCTGAC	51.8	
<i>argW</i> novel GI primers			
X106GIINTERNAL1	AAGCCGGAAATCACGAACGTAGTCG	59.3	
X106GIINTERNAL2	GATTTGAGGCAAAACACTTGTTGG	54	
X106intF	CTTACCCGCCTTGCCATTGAGC	58.6	
Tn5 Km ^r cassette primers			
CF#	TGGAATGCATGCGAACCGGAATTGC	59.3	NsiI
CR#	GGGAATGCATACTCATGAGATGCC	57.4	NsiI

A2.2 *In silico* SGSP-PCR results

Table A2. 3. Length of the *in silico* SGSP-PCR products with the U-T7 primers across the 16 tRNA loci in the *E. coli* K12 MG1655 chromosome

tRNA gene	<i>Bam</i> HI		<i>Eco</i> RI		<i>Hind</i> III		<i>Pst</i> I		<i>Sal</i> I		<i>Hinc</i> II		<i>Eco</i> RV	
<i>aspV</i>	6.3 ^a	6.5 ^b	0.7	1.6	1.0	2.5	22.9	25.4	12.6	15.9	0.5	3.9	0.8	0.8
<i>thrW</i>	5.3	18.2	6.5	10.5	4.2	28.7	0.4	3.1	22.1	22.7	3.7	3.8	1.4	2.8
<i>serW</i>	4.4	14.8	1.5	5.6	1.9	2.1	0.9	8.7	16.6	36.6	1.4	3.9	2.6	9.6
<i>serT</i>	5.1	5.6	1.4	1.3	27.5	45.3	0.9	13.6	19.0	19.7	4.9	5.5	3.2	5.7
<i>serX</i>	14.3	15.5	8.0	12.3	2.6	3.1	3.3	4.5	0.8	1.7	0.2	1.1	2.4	2.3
<i>serU</i>	14.8	20.4	2.0	10.6	0.4	5.9	7.7	11.4	16.0	25.3	1.6	2.3	2.1	2.3
<i>asnT</i>	5.2	20.4	8.2	10.6	5.1	5.9	3.4	11.4	9.1	25.3	0.5	2.3	3.7	3.8
<i>asnV</i>	3.0	5.3	6.4	11.6	7.2	7.4	4.2	9.5	15.2	22.3	2.2	3.8	1.8	2.7
<i>argW</i>	6.7	18.1	1.6	3.9	1.8	2.8	0.4	0.5	15.5	18.9	0.5	1.4	2.3	4.8
<i>ssrA</i>	10.8	15.3	1.4	13.1	1.2	5.7	1.9	2.7	5.8	12.9	1.1	1.7	0.3	1.2
<i>metV</i>	1.9	7.3	31.1	40.3	7.5	20.4	2.5	8.6	1.0	7.5	1.0	3.2	1.7	2.6
<i>glyU</i>	6.7	37.8	0.9	5.7	2.0	3.9	9.8	20.2	3.7	5.2	0.7	2.3	7.8	8.1
<i>pheV</i>	15.8	19.3	19.4	26.5	10.5	12.1	0.2	1.0	8.1	18.9	0.2	0.1	2.0	6.0
<i>selC</i>	18.9	32.6	1.3	16.6	15.7	23.8	4.4	15.1	9.1	10.0	2.2	3.1	0.2	0.8
<i>pheU</i>	13.6	22.9	5.4	9.9	14.6	27.0	4.1	7.3	0.4	15.0	0.3	1.8	0.4	7.1
<i>leuX</i>	25.3	34.5	10.5	12.4	3.2	4.0	2.6	15.2	18.1	30.1	1.1	1.1	0.3	0.7
<i>aspV</i> 2	6.4	6.5	0.8	1.6	1.0	2.5	22.9	25.4	12.6	15.9	0.6	3.9	0.9	0.8
<i>asnV</i> 2	2.0	5.3	5.4	11.6	6.2	7.4	3.2	9.5	14.2	22.3	1.2	3.8	0.8	2.7
<i>asnT</i> 3	5.1	20.4	8.1	10.6	5.0	5.9	3.2	11.4	8.9	25.3	0.3	2.3	3.6	3.8

Table A2. 4. Length of the *in silico* SGSP-PCR products with the D-T7 primers at the 16 tRNA loci in the *E. coli* K12 MG1655 chromosome

tRNA site	<i>Bam</i> HI		<i>Eco</i> RI		<i>Hind</i> III		<i>Pst</i> I		<i>Sal</i> I		<i>Hinc</i> II		<i>Eco</i> RV	
<i>aspV</i>	3.4	6.5	2.3	14.7	2.3	28.7	5.8	25.40	6.7	15.9	1.6	1.8	2.5	6.4
<i>thrW</i>	14.2	17.8	3.9	4.7	2.8	15.0	4.0	4.07	4.0	34.1	3.4	5.6	4.2	6.4
<i>serW</i>	12.0	14.8	0.1	5.9	1.8	2.1	0.6	16.52	21.6	36.6	0.2	2.5	8.6	9.6
<i>serT</i>	0.9	5.6	0.4	1.3	18.3	45.3	13.1	13.58	1.2	19.7	1.2	5.5	2.9	5.7
<i>serX</i>	2.2	15.5	5.3	12.3	1.6	3.1	2.3	4.52	0.3	0.5	0.3	0.2	1.1	2.3
<i>serU</i>	7.5	20.4	10.5	10.6	1.5	7.8	5.7	11.37	11.4	25.3	0.4	4.3	2.2	2.3
<i>asnT</i>	5.5	9.2	2.6	3.2	1.6	7.3	5.2	5.86	1.2	22.3	1.2	1.2	5.1	6.5
<i>asnV</i>	4.5	5.3	7.3	11.6	2.5	7.4	7.5	9.52	9.3	22.3	3.8	3.8	0.3	12.5
<i>argW</i>	6.4	17.8	3.4	20.3	1.6	4.5	3.3	12.82	17.4	18.9	1.4	2.4	1.5	5.5
<i>ssrA</i>	9.8	18.1	4.8	7.3	1.6	11.8	8.7	24.80	3.7	7.9	1.7	1.8	3.1	3.4
<i>metV</i>	6.9	7.3	10.7	40.3	14.5	20.4	7.6	8.59	0.6	3.0	0.6	1.7	2.5	2.6
<i>glyU</i>	5.9	9.7	1.4	9.7	2.0	5.3	2.8	3.97	9.0	11.2	0.4	1.6	0.7	0.8
<i>pheV</i>	13.8	19.3	17.4	26.5	11.9	12.1	0.7	12.97	2.1	8.9	2.1	6.1	5.4	10.2
<i>selC</i>	16.6	32.6	1.7	2.2	11.1	23.8	13.7	15.10	4.0	10.0	0.9	0.9	2.9	4.5
<i>pheU</i>	10.2	22.9	5.5	9.9	13.3	27.0	4.1	7.32	0.6	3.7	0.4	0.8	0.5	2.0
<i>leuX</i>	8.9	17.0	6.0	13.3	15.6	16.1	0.8	1.05	0.7	14.2	0.7	1.5	2.9	4.8
<i>aspV</i> 2	3.6	6.5	2.5	14.7	2.5	28.7	6.0	25.40	6.9	15.9	1.8	1.8	2.7	6.4
<i>asnV</i> 2	4.0	5.3	6.9	11.6	2.0	7.4	7.1	9.52	8.9	22.3	3.4	3.8	2.6	2.7
<i>asnT</i> 3	5.5	9.2	2.6	3.2	1.6	7.3	5.2	5.86	1.2	22.3	1.2	1.2	5.1	6.5

^a Length of SGSP-PCR amplicon (kb) indicated in bold.

^b Length of the restriction fragment that U#/D# is borne on (kb).

A2.3 Details of the different size tRIP amplicons generated

Table A2. 5. tRIP-positive amplicons of different length to the original six control *E. coli* and *Shigella* strains.

tRNA gene	Expected size in positive control (control strain)	Empty tRNA site	Strain	Strain code	tRIP amplicon length	Sequence data
<i>aspV</i>	3120 bp (K12 MG1655)	0.7 kb	<i>S. dysenteriae</i> 3	S101	1.7 kb	No
			<i>S. dysenteriae</i> 9	S102	1.7 kb	No
			<i>S. dysenteriae</i> 6	S103	1.7 kb	No
			<i>S. flexneri</i> 6	S110	1.7 kb	U# sequence has 99% ID to the Sb227 <i>aspV</i> UF and the associated islet
			<i>S. boydii</i> 1	S116	1.7 kb	No
			<i>S. boydii</i> 2	S117	1.7 kb	No
			<i>S. boydii</i> 3	S118	1.7 kb	No
			<i>S. boydii</i> 4	S119	1.7 kb	D# sequence has 99% ID to the Sb227 <i>aspV</i> DF and the associated islet
<i>serW</i>	1422 bp (K12 MG1655)	1.1 kb	<i>S. sonnei</i>	S113	1.5 kb	No (<i>in silico</i> tRIP in <i>S. sonnei</i> 046 is 1484 bp)
			<i>S. sonnei</i> bio a	S114	1.5 kb	No
			<i>S. sonnei</i> bio g	S115	1.5 kb	No
<i>metV</i>	1379 bp (K12 MG1655), 1270 bp in Sf301 and Sf2457T due to a missing <i>met</i> tRNA in the U flank. The locus is empty in the three <i>Shigella</i>	1.4 kb	<i>S. flexneri</i> 1a	S104	1.3 kb	No
			<i>S. flexneri</i> 1b	S105	1.3 kb	U# sequence has 99% ID to the Sf301 <i>metV</i> UF and DF, and confirms the missing <i>met</i> tRNA in this strain
			<i>S. flexneri</i> 2a	S106	1.3 kb	No
			<i>S. flexneri</i> 2b	S107	1.3 kb	No
			<i>S. flexneri</i> X	S111	1.3 kb	U# sequence has 99% ID to Sf301 and Sf2457T <i>metV</i> UF and DF, and confirms the missing <i>met</i> tRNA in this strain

	strains		<i>S. flexneri</i> Y	S112	1.3 kb	No
<i>glyU</i>	838 bp (CFT073)	0.7 kb	<i>S. flexneri</i> 6	S110	1.2 kb	U# sequence has 98% ID to Sb227 <i>glyU</i> UF and walks 22 bp into the associated GI
			<i>S. flexneri</i> X	S111	3.2 kb	No
			<i>S. flexneri</i> Y	S112	3.2 kb	U# sequence has 97% ID to the UF and the U arm of the <i>glyU</i> GI in EDL933
			<i>S. boydii</i> 7	S120	4.4 kb	U# sequence has 98% ID to the UF and the U arm of the <i>glyU</i> GI in EDL933
<i>argW</i>	Initially there was no positive control strain with an insert under 3.0 kb	0.9 kb	<i>S. dysenteriae</i> 9	S102	2.6 kb	U# sequence has 99% ID to the Sb227 <i>argW</i> UF, walks 160 bp into the associated islet then the start of an ISS <i>bo6</i> , which is not present in any of the sequenced genomes in this location (see Figure 6.4)
			<i>S. flexneri</i> 6	S110	1.6 kb	U# sequence has 100% ID to the Sb227 <i>argW</i> UF and tRNA
			<i>S. boydii</i> 2	S117	1.6 kb	U# sequence has 99% ID to the Sb227 <i>argW</i> UF and associated islet
			<i>S. boydii</i> 3	S118	2.5 kb	U# sequence has 99% ID to the Sb227 <i>argW</i> UF and associated islet
			<i>S. boydii</i> 4	S119	1.6 kb	U# sequence has 100% ID to the Sb227 <i>argW</i> UF and associated islet
<i>selC</i>	2850 bp (K12 MG1655)	1.0 kb	<i>S. dysenteriae</i> 3	S101	3.3 b	U# sequence has 99% ID to Sb227 <i>selC</i> UF and islet
			<i>S. dysenteriae</i> 9	S102	3.3 kb	U# sequence has 99% ID to Sb227 <i>selC</i> UF and islet
			<i>S. dysenteriae</i> 6	S103	3.3 kb	No
			<i>S. flexneri</i> 6	S110	4.8 kb	Same as S116
			<i>S. boydii</i> 1	S116	4.8 kb	Confirmed by SGSP-PCR as containing the same sequence as the Sb227 <i>argW</i> islet
			<i>S. boydii</i> 2	S117	4.8 kb	No
			<i>S. boydii</i> 3	S118	5.0 kb	No

			<i>S. boydii</i> 4	S119	4.8 kb	Same as S116
leuX	Initially there was no positive control strain with an insert under 3.0 kb	1.8 kb	<i>S. flexneri</i> 1b	S105	6.0 kb	No
			<i>S. flexneri</i> 6	S110	4.5 kb	D# sequence has 99% ID to the Sb227 <i>leuX</i> DF
			<i>S. flexneri</i> Y	S112	6.0 kb	No
ssrA	Initially there was no positive control strain with an insert under 3.0 kb	1.0 kb	<i>S. flexneri</i> 1a	S104	4.8 kb	Same as S112
			<i>S. flexneri</i> 1b	S105	4.8 kb	Same as S112
			<i>S. flexneri</i> 2a	S106	4.8 kb	Same as S112
			<i>S. flexneri</i> 2b	S107	4.8 kb	Same as S112
			<i>S. flexneri</i> X	S111	4.8 kb	Same as S112
			<i>S. flexneri</i> Y	S112	4.8 kb	Confirmed by SGSP-PCR as containing the same sequence as the Sf301 islet

A2.4 Details SGSP-PCR derived island sequences that are defined by tRIP only.

Table A2. 6. Island sequences obtained by sequencing of SGSP-PCR amplicons that are defined as island DNA by the tRIP method only and their corresponding GC contents and nucleotide matches to the *E. coli* and *Shigella* genomes available in the NCBI database.

				Sequenced <i>Shigella</i> strains ^b							Sequenced <i>E. coli</i> strains													
Strain -tRNA locus	U/D amplicon	GC content ^a	No. of genomes the sequence is present in ^a	<i>S. b</i> BS512	<i>S. d</i> 1012	Sd197	Sf301	Sf2457T	Ss046	Sb227	K12 MG1655	CFT073	Sakai	EDL933	W3110	101-1	53638	B171	B7A	E110019	E22	E24377A	F11	HS
<i>aspV</i>																								
S104	D	38.3%	2				Y _c	Y																
S105	D	34.5%	2				Y	Y																
S106	D		2				RP _d	RP																
S107	D		2				RP	RP																
S108	D		15		RP	RP			RP		RP		RP	RP	RP	RP	RP	RP	RP	RP	RP	RP		RP
S111	D		2				RP	RP																
S112	D		2				RP	RP																
S115	D	40.8%	15		Y	Y			Y		Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y
<i>metV</i>																								
S101	U	45.6%	4	Y						Y		Y											Y	
S102	D		4	RP						RP		RP											RP	

S103	D	35.6%	4	Y			Y		93% e										93%
S110	D		4	RP			RP		RP										RP
S116	U	34.9%	4	Y			Y		91%										91%
	D	34.2%	4	Y			Y		91%										Y
S117	D		4	RP			RP		RP										RP
S118	D		4	RP			RP		RP										RP
S119	D		4	RP			RP		RP										RP
<i>glyU</i>																			
S101	D																		
S102	D	32.1%	16	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
S103	D	30.8%	16	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
S105	D		16	RP		RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP
S106	D		16	RP		RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP
S107	D		16	RP		RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP
S108	D	34.9%	16	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
S113	U	32.9%	19			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	D		13						RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP
S114	D		13						RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP
S115	D	35.2%	13					Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
<i>argW</i>																			
S108	D	51.2%	12			Y	Y	Y	Y					Y	Y	Y	Y	Y	Y
S114	D	50.6%	12			Y	Y	Y	Y					Y	Y	Y	Y	Y	Y
S115	D	45.8%	12			Y	Y	Y	Y					Y	Y	Y	Y	Y	Y

<i>serX</i>																					
S101	U		4	RP				RP	RP											RP	
S102	U		4	RP				RP	RP											RP	
S103	U		4	RP				RP	RP											RP	
S108	U	41.9%	4	Y				Y	Y											Y	
S113	U	40.4%	4	Y				Y	Y											Y	
S114	U		4	RP				RP	RP											RP	
S115	U		4	RP				RP	RP											RP	
S116	U		4	RP				RP	RP											RP	
S117	U		4	RP				RP	RP											RP	
S118	U		4	RP				RP	RP											RP	
S119	U	40.9%	4	Y				Y	Y											Y	
<i>asnT</i>																					
S101	D	47.3%	19	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
S102	U	42.1%	12	Y	Y			Y	Y	Y							Y		Y	Y	Y
S103	D	46.6%, 54.8%	19	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
S120	D	48.9%	16		Y			Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
<i>pheV</i>																					
S101	U		14	RP	RP				RP	RP	RP		RP	RP	RP		RP	RP	RP	RP	RP
S102	U		14	RP	RP				RP	RP			RP	RP	RP		RP	RP	RP	RP	RP
S104	U	45%	2			Y	Y														
S105	U		2			Y	Y														
S110	U		14	RP	RP				RP	RP	RP		RP	RP	RP		RP	RP	RP	RP	RP
S111	U		2			Y	Y														

S116	U		14	RP	RP		RP												
S119	U	60.3%	14	Y	Y		Y												
<i>leuX</i>																			
S101	U	41.9%	9	Y			Y	Y										Y	
S102	U	41.9%	9	Y			Y	Y										Y	
	D	42.4%	11					Y										Y	Y
<i>ssrA</i>																			
S101	U	50.9%	11	Y	Y	Y		Y					Y			Y		Y	
S102	U		11	RP	RP	RP		RP					RP			RP		RP	
S103	U		10	RP	RP	RP		RP					RP					RP	
S110	U	49.7%	11	Y	Y	Y		Y					Y			Y		Y	
S113	D	31.9%	10	Y			Y						Y			Y		Y	
S114	D	29.5%	10	Y			Y						Y			Y		Y	
S116	U		11	RP	RP	RP		RP					RP			RP		RP	
S117	U		10	RP	RP	RP		RP					RP					RP	
S118	U	50.8%	10	Y	Y	Y		Y					Y					Y	
S119	U		11	RP	RP	RP		RP					RP			RP		RP	

- ^a Excludes any IS elements.
- ^b Strains highlighted in bold are completely sequenced genomes, those not in bold are incomplete genomes.
- ^c The sequence obtained had $\geq 95\%$ nucleotide identity to the sequenced strain indicated.
- ^d The SGSP-PCR amplicon(s) generated indicated that this strain had the same restriction pattern (RP) in the GI U/D-arm to a sequenced representative, and therefore was very likely to harbour the same sequence.
- ^e The sequence obtained had less than 95% nucleotide identity to the sequenced strain indicated, as indicated by the actual percentage shown.

A2.5 Details of the generation of the *leuX* UF suicide constructs

A2.5.1 Insertion and orientation of the Km^r cassette in the *leuX* UF region

500 ng of the cleaned PCR amplified Tn5 derived Km^r cassette was digested with *Nsi*I and ligated into pJL1/*Nsi*I. The ligation was electroporated into *E. coli* DH5 α and the cells plated onto LA containing 50 μ g/ml Km and 100 μ g/ml Ap. As in the *argW* experiments, some of the Ap^r + Km^r transformants were screened to verify the orientation of the Km^r cassette relative to the surrounding *leuX* UF DNA (see 6.3.5).

The optimal conformation is shown in Figure A2. 1. This was performed by digesting candidate plasmids with the restriction enzymes *Sac*I (cuts in the Km^r cassette and MCS of pBluescript) and *Eco*RV (cuts in the *leuX* UF region and MCS of pBluescript, data not shown).

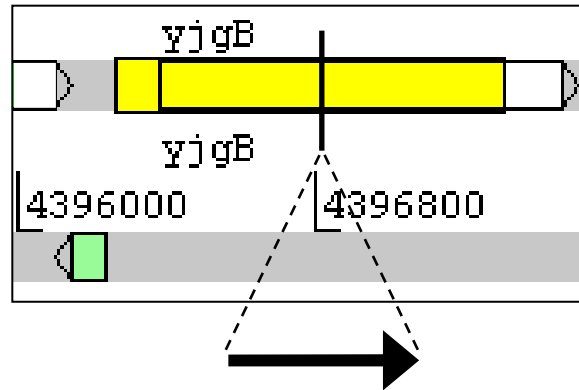


Figure A2. 1. View of Sf301 *leuX* UF region taken from Artemis.

The green box is the *leuX* tRNA gene; its orientation is shown by the associated arrowhead. The yellow area indicates the region in the UF where homologous recombination would take place. Within this region is the non-essential gene *yjgB* (Gerdes *et al.*, 2003), its orientation is shown by the associated arrowhead. The vertical black line indicates where the *yjgB* gene would be disrupted by the *Km^r* cassette, the black arrow (not to scale) indicates the preferred orientation of the *Km^r* cassette as it is colinear with *yjgB*.

The plasmid with the insert DNA in the optimal orientation was designated pJL2 (see Figure A2. 2)

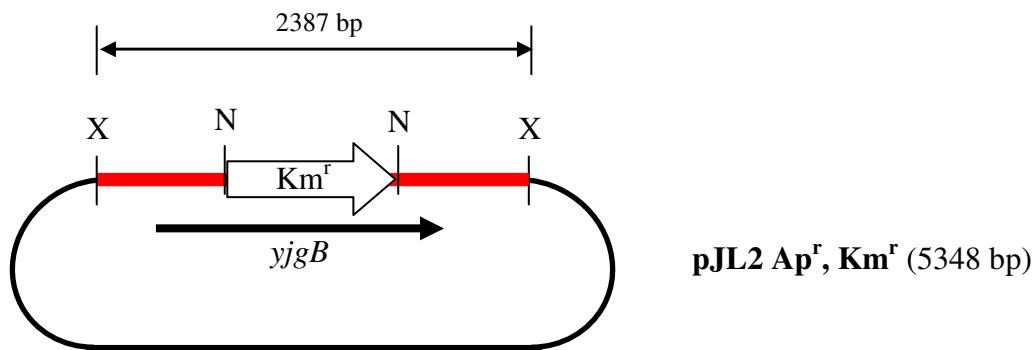
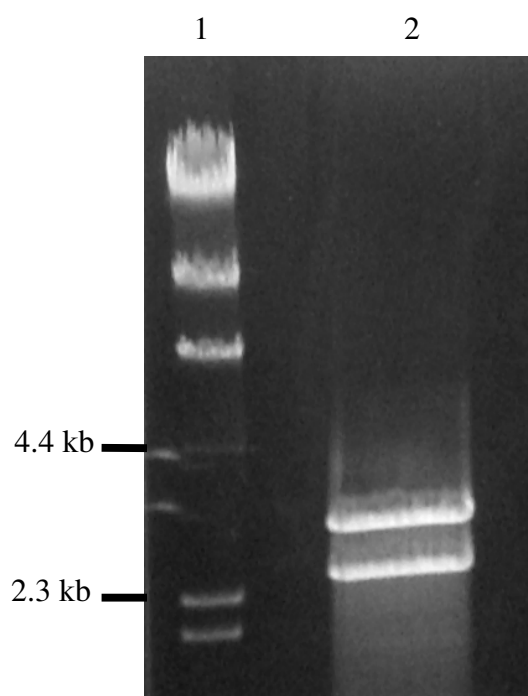


Figure A2. 2. pJL2.

The black segment represents pBluescript and the red segment represents the *leuX* UF region, the white arrow is the Km^r cassette and its orientation, the thick black arrow indicates the orientation of the *yjgB* gene. X and N stand for *Xba*I and *Nsi*I sites respectively. Drawings are not to scale.

A2.5.2 Insertion and orientation of the mutant *leuX* UF region in pDS132

1 µg of pJL2 was digested with *Xba*I to produce two fragments of 2961 bp (pBluescript) and 2387 bp (the *leuX* UF mutant construct). The entire digest was electrophoresed (see Figure A2. 3).



Lane number:

1. λ /HindIII ladder
2. pJL2/XbaI

Figure A2. 3. Agarose gel showing pJL2 digested with *Xba*I.

The smaller of the two bands is the 2387 bp *leuX* region containing the Km^r cassette.

The 2387 bp fragment was gel extracted, cleaned and ligated into pDS132/*Xba*I. The ligation was electroporated into *E. coli* CC118 λ *pir* as this strain supports the replication of the suicide plasmid to low copy-number. The cells were plated onto LA plus 50 μ g/ml Km and 30 μ g/ml Cm. Transformants were screened to verify the orientation of the Km^r cassette with regards to the *sacB* gene in pDS132. This was performed by digesting candidate plasmids with *EcoRV* (data not shown). Constructs with the insert in both orientations were obtained and these were designated pJL3 and pJL4 (see Figure A2. 4).

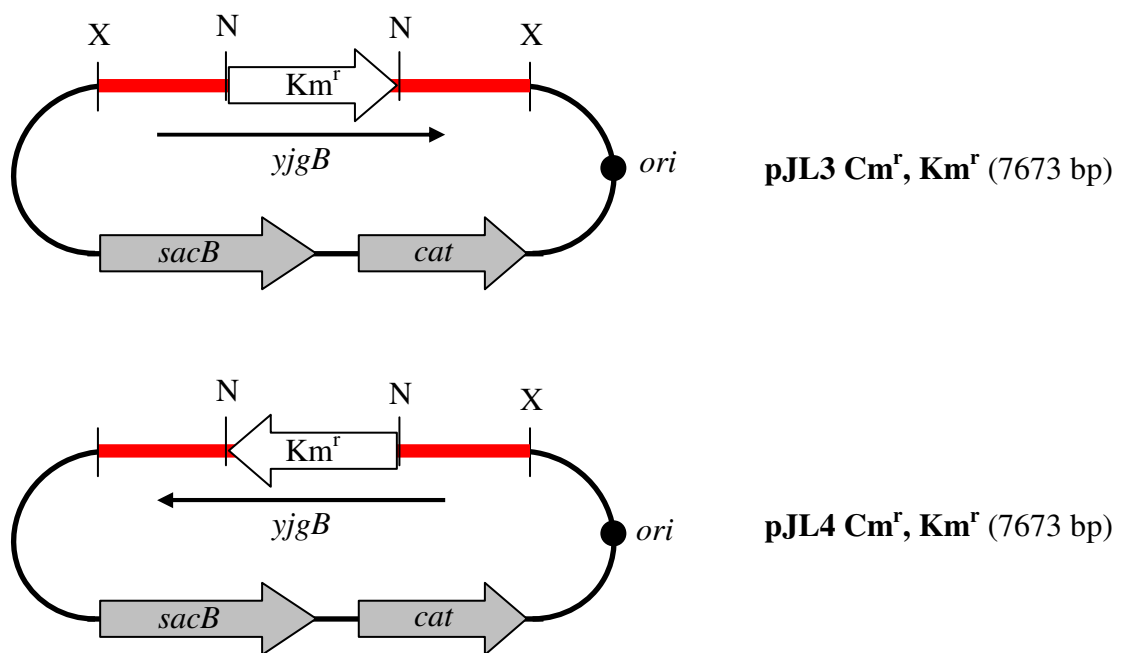


Figure A2. 4. pJL3 and pJL4: pDS132 derived suicide constructs used to deliver the mutant *leuX* UF region to S101.

pJL4 is the preferred choice as the *Km^r* cassette is colinear with the *sacB*, however, both constructs were used in the allelic exchange experiments to maximise the chances of generating recombinants.

These were then electroporated separately into *E. coli* SM10 λ *pir* and the cells plated onto LA plus 50 μ g/ml Km and 30 μ g/ml Cm. Transformants harbouring pJL3 and pJL4 (KR238 and KR239 respectively) were both used in conjugations with the recipient *Shigella* strain S101 (see section 2.16 for the protocol).

A2.6 Details of the island characterisation across all other tRNA loci known to be hotspots for GI insertion across *E. coli*

A2.6.1 *serW*

At the *serW* locus, S101 (*S. dysenteriae* 3 strain) was the only tRIP negative strain in this study and SGSP-PCR with the 5 original libraries from both the U and D primers did not produce any specific SGSP-PCR amplicons. As this strain is resistant to the antibiotics Ap, Cm, Sm, and Tc, I hypothesised that it harboured the SRL PAI (GenBank accession number AF326777) at *serW*, as the *serX* locus was found to be occupied with a different island by SGSP-PCR and sequence analysis (see Table 5.1). The SRL PAI is interesting in that it has an integrase gene at the distal end of the island, this is very rare, as with most GIs that harbour an integrase, it is located directly downstream of the tRNA gene. On further analysis of the *serW* and *serX* associated GIs in the four sequenced *E. coli* and *Shigella* genomes it was found that they also all harbour an integrase at the distal end of the GI (see Table A2. 7) and it is the last gene before a DR of the 3' terminus of the *ser* tRNA, indicating that they are all prophage-like islands. The integrase has 99% nucleotide identity across all of the strains, highlighting the evidence for the dissemination of a similar prophage like entity across these strains at *serW* and *serX*.

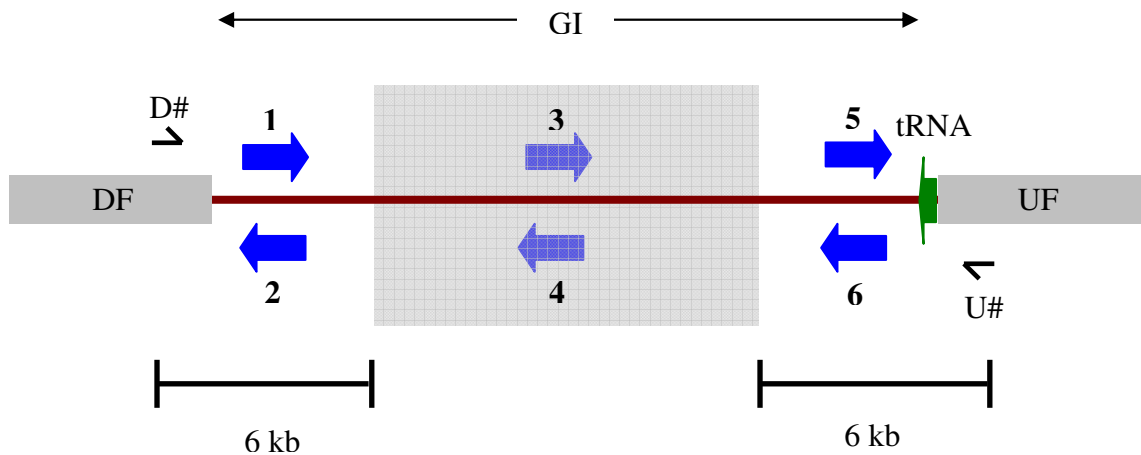
Table A2. 7. Occupancy of *serW* and *serX* in the original four complete *E. coli* genomes and the respective integrase locations of each island.

Strain	tRNA locus	Size of inserted element between UF and DF (kb)	Integrase present	Location of integrase
K12 MG1655	<i>serW</i>	0.3	No	N\A
	<i>serX</i>	0.5	No	N\A
EDL933	<i>serW</i>	87.9 ^a	Yes	D-arm
	<i>serX</i>	87.6	Yes	D-arm
CFT073	<i>serW</i>	0.3	No	N\A
	<i>serX</i>	113.8	Yes	D-arm
Sf301	<i>serW</i>	0.4	No	N\A
	<i>serX</i>	0	No	N\A

^a Bold text indicates strain-tRNA loci harbouring GIs.

Standard PCR using the *serW* D# and a primer that is conserved in the above integrase genes could be performed to prove the linkage of the integrase with the *serW* DF, I named this technique *int*-PCR. This sort of method has been used previously by Al-Hasani *et al.* 2001 to demonstrate the linkage of an integrase with corresponding chromosomal regions.

The integrase in the above GI positive strains is found in the same conformation, however, in an uncharacterised strain, there is the possibility that the integrase could be in any of six different conformations (see Figure A2. 5).



1. At distal end of GI, complementary to tRNA (as in SRL PAI, and GI harbouring strains in Table A2. 7)
2. At distal end of GI, colinear with tRNA
3. Anywhere within the GI that is over 6 kb from the U/D primer, complementary to tRNA
4. Anywhere within the GI that is over 6 kb from the U/D primer, colinear with tRNA
5. At proximal end of GI, complementary to tRNA
6. At proximal end of GI, colinear with tRNA

Figure A2. 5. Schematic showing the possible conformations that a GI harboured integrase gene could be in.

Blue arrows represent integrase genes. The grey area within the GI indicates the region that *int*-PCR would not be able to detect an integrase due to the limits of the standard PCR regime used. Figure is not to scale.

In *E. coli* and *Shigella*, conformations 5 and 6 are the most common; however 1 and 4 have also been reported.

To test for the presence of a similar integrase at positions 1, 2, 5 or 6 at the *serW* locus in S101, two conserved primers were designed to use in an *int*-PCR with the *serW* U and D

primers. The primers were placed towards the 5' end of the integrase gene as this is near the sequence that encodes the C-terminal domain of the integrase protein (Nunes-Duby *et al.*, 1998) and is likely to be more conserved in an uncharacterised strain. The primers were 21 mers long and the reverse complement of each other; named P4I_R# and P4I_F# (see Figure A2. 6 and Table A2. 2).

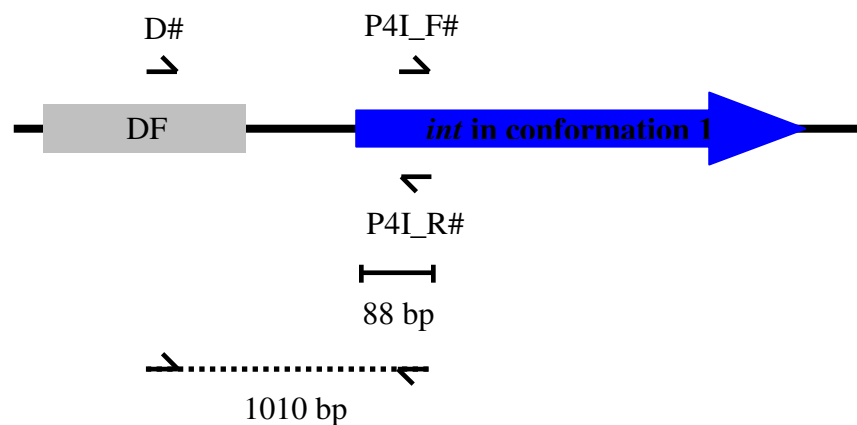
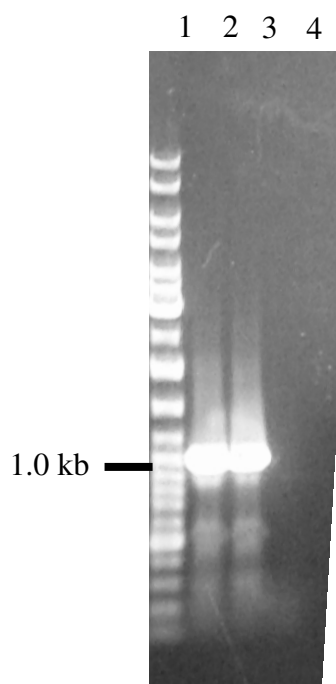


Figure A2. 6. Positions of the *int*-PCR primers on the *serX/serW* associated integrase gene in the sequenced genomes and SRL PAI.

Figure is not to scale. *In-silico* PCR using *serW* D# - P4I_R# with EDL933 produces a 1010 bp amplicon.

An *int*-PCR was performed using the tRIP stock genomic DNA as the template (see 3.1.1) using hot-start, touchdown PCR (see 2.8.2). The extension time for each cycle was 2.5 min.



Lane number:

1. GeneRuler™ Ladder (Fermentas)
2. *serW* D# - P4I_R#, 10 ng of EDL933 genomic DNA as template (positive control)
3. *serW* D# - P4I_R#, 10 ng of S101 genomic DNA as template
4. *serW* D# - P4I_F#, 10 ng of S101 genomic DNA as template

Figure A2. 7. Agarose gel of the *serW* *int*-PCR.

The results show that as expected, S101 has an integrase in the same conformation as with EDL933 (and all of the other GIs described above). To check the results, colony PCR was also performed to prove that this result was not due to cross contamination of S101 DNA with EDL933 DNA; the results were the same as above.

The rest of the PCR from lane 3 was gel extracted and sent for sequencing from the *serW* D#. The sequencing results show that the island DNA has the highest nucleotide identity to the SRL PAI (*serW*-IF1), with a 14 bp DR of the *serW* tRNA marking the distal end of the PAI. The flanking DNA has the highest nucleotide identity to Sb227 (see Figure A2. 8). This is the only case in the study where an integrase has been found at the distal end of the GI; however

it is likely that the *serX* associated SRL PAI-like elements found in the *S. flexneri* 2b and *S. flexneri* 6 strains also have an integrase in their D-arms (see *serX* results section A2.6.8).

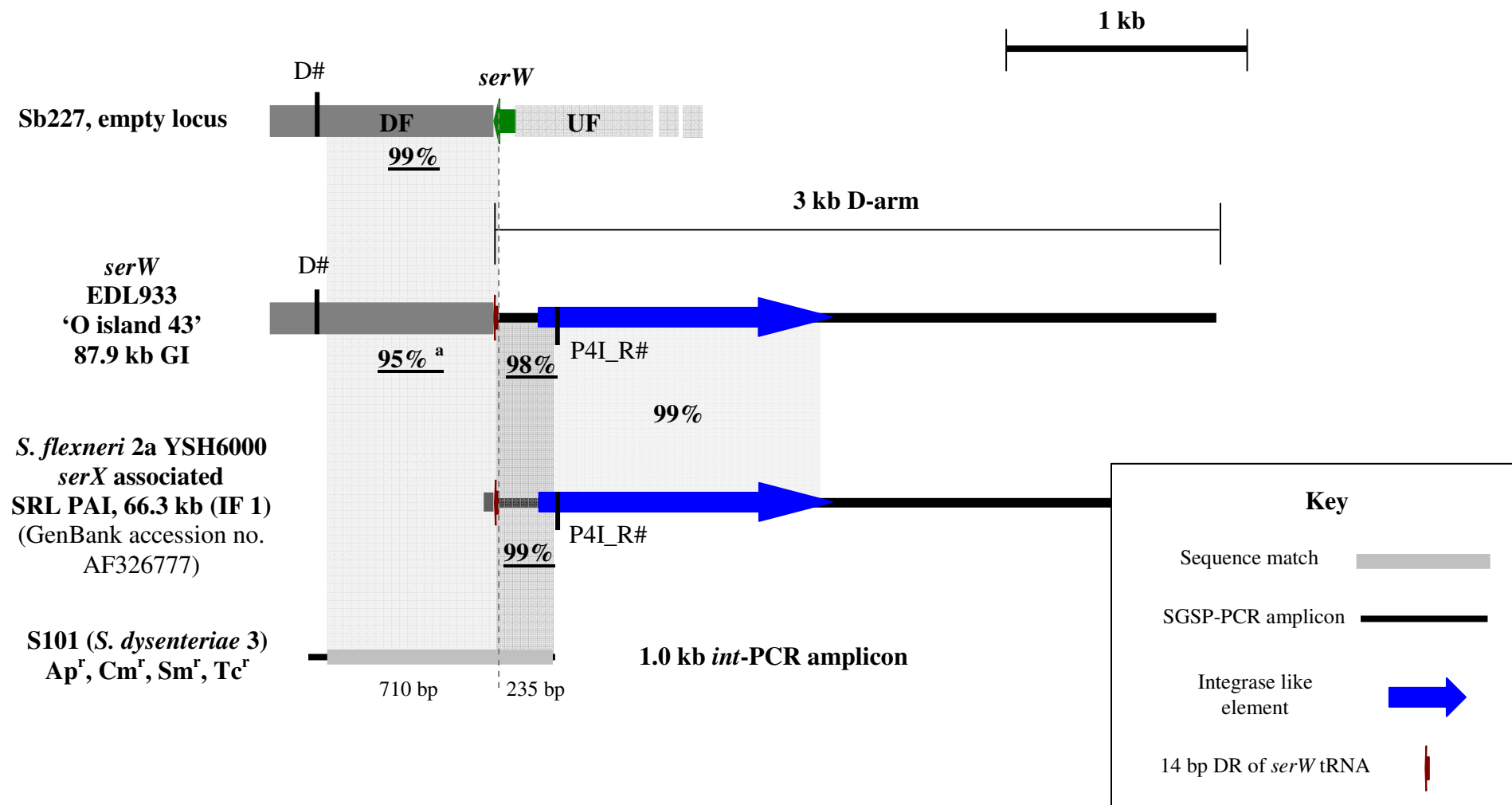


Figure A2. 8. *serW* D# - P4I_R *int*-PCR sequencing results for S101 (*S. dysenteriae* 3 strain).

^a Underlined values indicate the percentage nucleotide identity of the S101 sequence to the corresponding genomes

A2.6.2 *glyU*

Table A2. 8. SGSP-PCR results of the *glyU* tRIP-negative strain-tRNA loci

<i>glyU</i> U# - T7# SGSP-PCR results									
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)							Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	<i>Eco</i> RV	<i>Hinc</i> II	
K12 MG1655				~2.0					
<i>S. dysenteriae</i> 3	S101	N ^a	N	~2.0	N	N	N	N	
<i>S. dysenteriae</i> 9	S102	N	N	N	N	N			
<i>S. dysenteriae</i> 6	S103	N	N	~2.6 F ^b	~1.3 F	N			
<i>S. flexneri</i> 1a	S104	N	N	~ 1.0 ^c	N	N			392 [SK#], seq from U#
<i>S. flexneri</i> 1b	S105	N	N	~ 1.3 F	N	N			687 [U#] N/S ^d
<i>S. flexneri</i> 2a	S106	N	N	N	~2.0 F	N			
<i>S. flexneri</i> 2b	S107	N	N	~1.3 F	N	N			
<i>S. sonnei</i>	S108	N	~3.0 F	N	N	N			
<i>S. sonnei</i>	S113	~ 3.0 F	N	~2.6 F	N	N			274 [U#]
<i>S. sonnei</i> bio a	S114	N	N	N	N	N			
<i>S. sonnei</i> bio g	S115	N	N	N	N	N			
<i>S. boydii</i> 1	S116	N	N	~ 0.8 F	N	N	N	~2.5 F	510 [U#], 640 [SK#] N/S
<i>S. boydii</i> 2	S117	N	N	N	N	N			
<i>S. boydii</i> 3	S118	N	N	N	N	N			
<i>S. boydii</i> 4	S119	N	N	N	N	N	N	N	

<i>glyU</i> D# - T7# SGSP-PCR results									
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)							Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	<i>Eco</i> RV	<i>Hinc</i> II	
K12 MG1655				~2.0					
<i>S. dysenteriae</i> 3	S101	N	N	~ 1.0	~0.4	N			596 [SK#]
<i>S. dysenteriae</i> 9	S102	N	N	~ 1.4	~0.4	N			742 [SK#]
<i>S. dysenteriae</i> 6	S103	N	N	~ 1.4	N	N			873 [SK#]
<i>S. flexneri</i> 1a	S104		~2.3 F	~ 1.2	~2.0 F				774 [SK#]
<i>S. flexneri</i> 1b	S105		~2.3 F	~1.2	~2.0 F				
<i>S. flexneri</i> 2a	S106		~2.3 F	~1.2	~2.0 F				
<i>S. flexneri</i> 2b	S107		~2.3 F	~1.2	~2.0 F				
<i>S. sonnei</i>	S108	N	~ 1.4	~2.0	N	N			724(SK#)
<i>S. sonnei</i>	S113		~1.4	~2.0					
<i>S. sonnei</i> bio a	S114		~1.4	~2.0					
<i>S. sonnei</i> bio g	S115	N	~1.4	~ 2.0	N	N			861 [SK#]
<i>S. boydii</i> 1	S116	N	N	N	~0.4	N	N	~ 1.3	594 [SK#], 271 [D#] N/S
<i>S. boydii</i> 2	S117	N	N	~ 1.0	~0.4	N			340 [SK#] N/S
<i>S. boydii</i> 3	S118			~1.0 F					
<i>S. boydii</i> 4	S119	N	N	N	~0.4	N	N	~0.6 F	

^a Indicates that no amplicon was generated

^b The addition of 'F' after the text indicates that the amplicon was faint

^c Text highlighted in bold indicates that the amplicon was sequenced

^d Indicates that the sequenced amplicon was non-specific

S. sonnei

The D-arm results show that all of the *S. sonnei* strains harbour the sequence present in the D-arm of the Ss046 9.1 kb *glyU* associated GI (*glyU*-IF3, see Figure A2. 9). Even though the sequence walked into is present in over 10 of the *E. coli* genomes present on the NCBI database, it is likely to be island DNA, as it has a GC content of around 35%, 15% lower than the genome average, indicating that it was acquired from a foreign source (see Table A2. 6). The U-arm results show that only one of the *S. sonnei* strains was characterised (S113), as SGSP-PCR with the other three strains yielded no amplicons with all five libraries, indicating that the U-arm in these uncharacterised strains is different to S113. The U-arm sequence results for S113 indicated the presence of the same DNA as is found directly downstream of *glyU* in Ss046 (see Figure A2. 10). The DNA present on the Ss046 *glyU* island comprises IS elements and five ORFs that encode conserved hypothetical proteins which may be involved in transcriptional control and protein folding. One of the genes present in the D-arm, *yqeI*, which is disrupted by an *IS1*, would have previously encoded a putative sensory transducer, which is involved in chemotaxis. This suggests that this element may have originally been much larger and previously contained functional genes that have since been lost due to IS disruption.

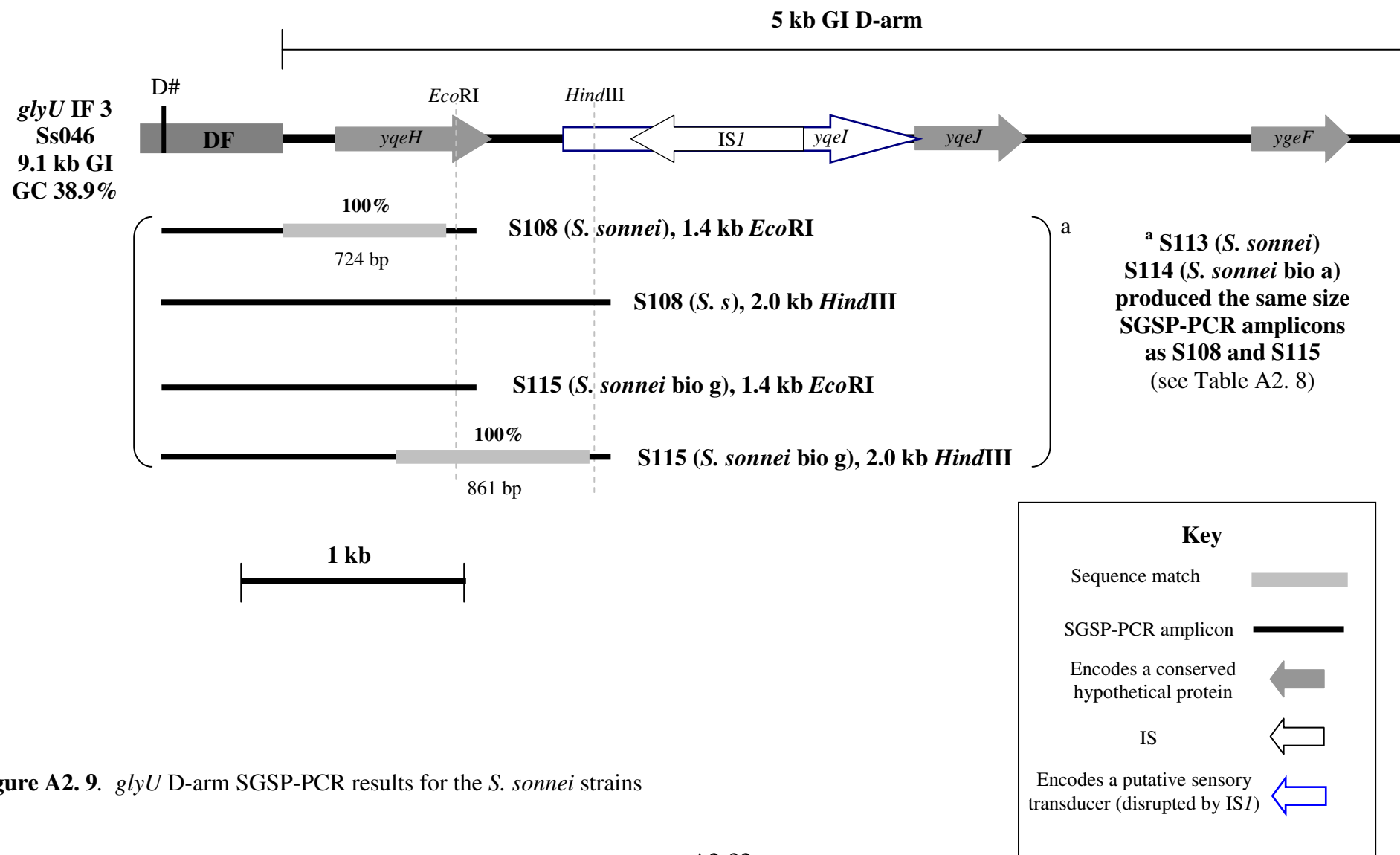


Figure A2. 9. *glyU* D-arm SGSP-PCR results for the *S. sonnei* strains

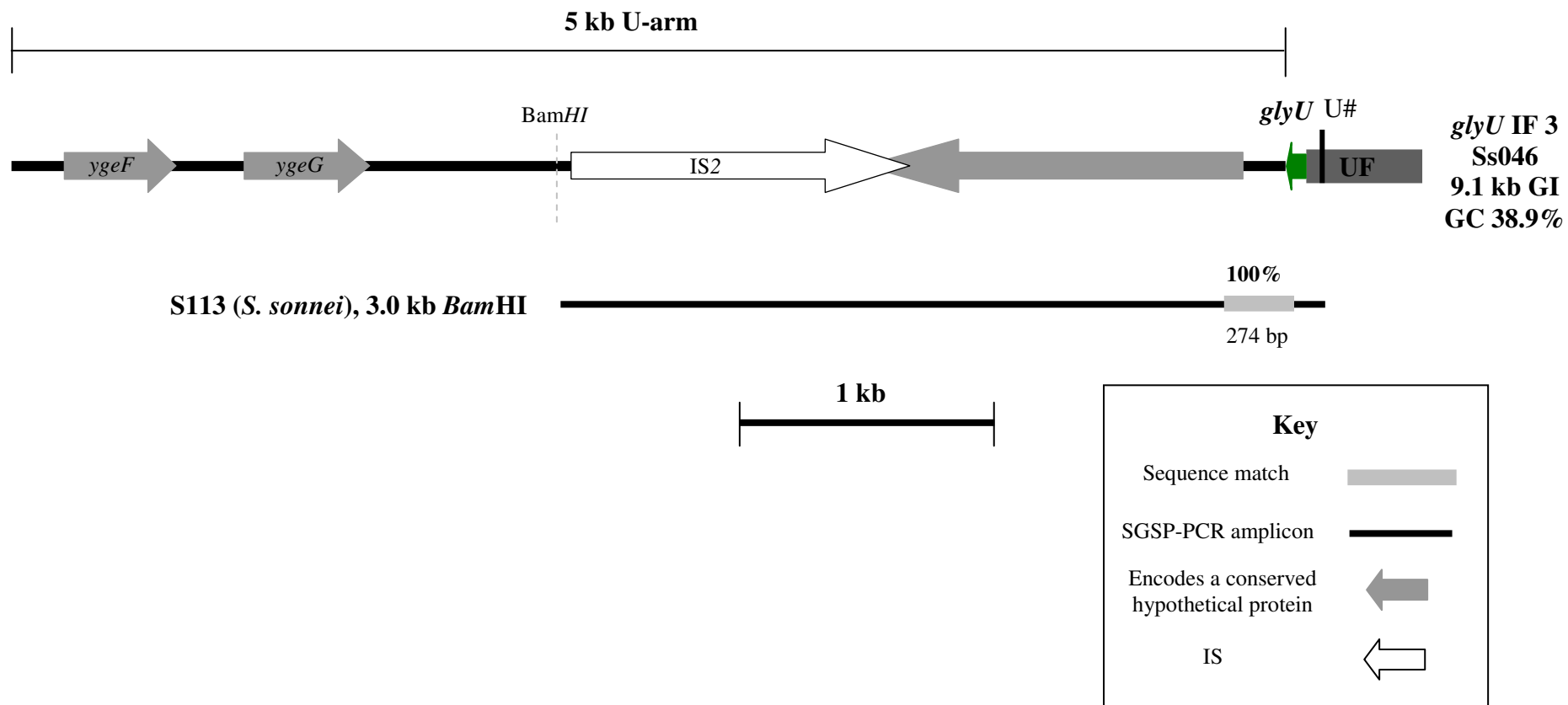


Figure A2. 10. *glyU* U-arm SGSP-PCR results for the *S. sonnei* strain S113

S. flexneri

The D-arm results show that all of the tRIP-negative *S. flexneri* strains harbour the same island DNA as is present in the D-arm of the Sf301 *glyU* associated 10.1 kb GI (see Figure A2. 11). The sequence walked into has regions of identity to the other *E. coli* and *Shigella* genomes, but after comparison with the other genomes, it is clear that these are just U and D-arm remnants of a previously inserted, larger element that has since been deleted, with the presence of an IS3 between the two remnants suggesting it has played a role in the deletion of the island DNA between these sequences (see Figure A2. 12). In Sf301 this island DNA is present as a 2.4 kb D-arm 'flanking islet' that was probably originally directly downstream of the *glyU* tRNA, but there has since been the insertion of another prophage-like element at the *glyU* tRNA, resulting in the formation of a 17 bp DR of the 3' terminus of *glyU* and the displacement of the former sequence further downstream. The 7.6 kb GI directly downstream of the *glyU* tRNA locus in Sf301 is also represented in the Islander database. Whether this prophage-like element is present in the *S. flexneri* strains S105, S106 and S107 is unknown, as SGSP-PCR from the U# in these strains was unsuccessful, whereas in S104, another prophage-like island is present in the U-arm, but it is different to the Sf301 *glyU* island (see A2.6.3 below).

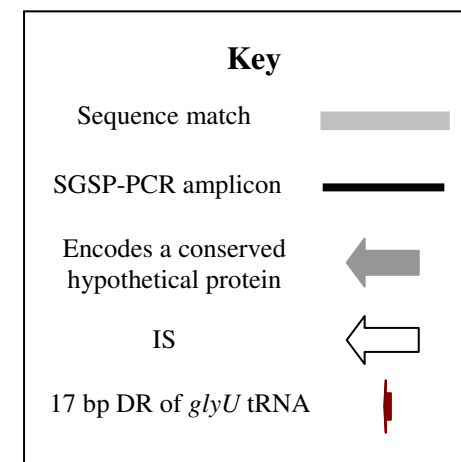
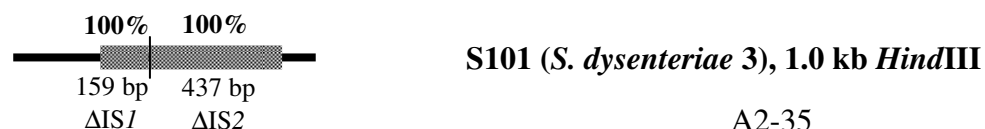
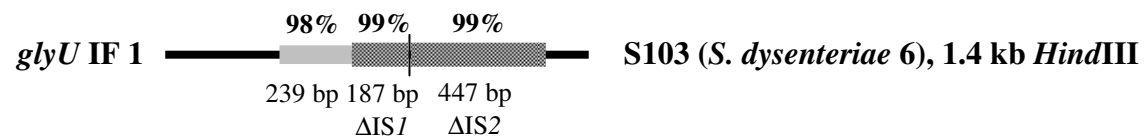
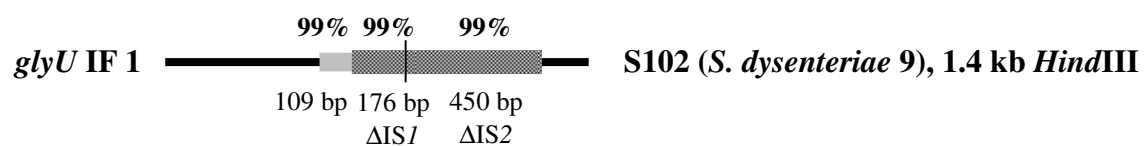
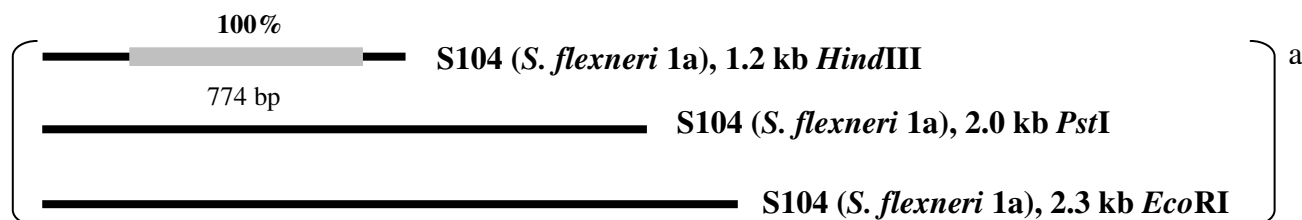
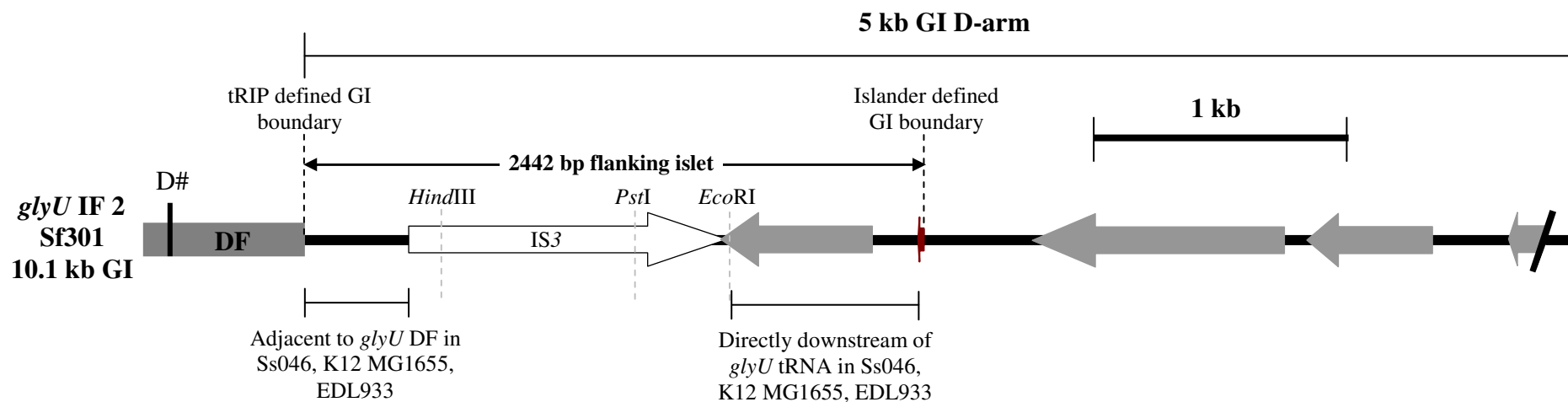


Figure A2. 11. *glyU* D-arm SGSP-PCR results for the *S. flexneri* strains S104, S105, S106 and S107 and the three *S. dysenteriae* strains.

^a S105 (*S. flexneri* 1b), S106 (*S. flexneri* 2a), S107 (*S. flexneri* 2b), produced the same size D# SGSP-PCR amplicons as S104 (see Table A2. 8).

Blastn analysis of the other complete *E. coli* and *Shigella* genomes indicated that K12 MG1655 and Sf301 harbour similar elements to Ss046 at *glyU* (see Table A2. 6), however Sf301 also has another prophage element inserted upstream, directly adjacent to *glyU* (see Figure A2. 12). The Ss046 *glyU* island had the most significant nucleotide matches to parts of the EDL933 *glyU* associated GI, which is defined by tRIP as being 27.7 kb long and has a GC content of 35.8%, very similar to that of the Ss046 GI which is 38.9%. On closer inspection it can be seen that a 21.0 kb region present in the EDL933 GI is absent from the Ss046 island and the other strains (see Figure A2. 12). This region is only found in EDL933 (and Sakai) and comprises genes which encode a type three secretion system (T3SS) and a putative invasin. The products of this T3SS have between 30-70% similarity to the products of the *Salmonella enterica* SPI-1 GI that harbours the *inv-spa* complex, which encodes a T3SS involved in the entry of the bacteria into the hosts epithelial cells and is essential for virulence. The amino acid content of some of the invasion genes are known to be diverse amongst subspecies of *S. enterica*; also the *inv-spa* genes in *S. enterica* are homologous to the invasion genes harboured on the *Shigella* virulence plasmid (see introduction). However, (Boyd *et al.*, 1997), suggested that the *Shigella* invasion genes were not acquired from *Salmonella*, but arose independently. The results of this study indicate that at some time, the ancestral *Shigella* strain harboured these genes on the *glyU* associated distal GI, but they have since been deleted, most likely after the acquisition of the virulence plasmid encoded T3SS (see below).

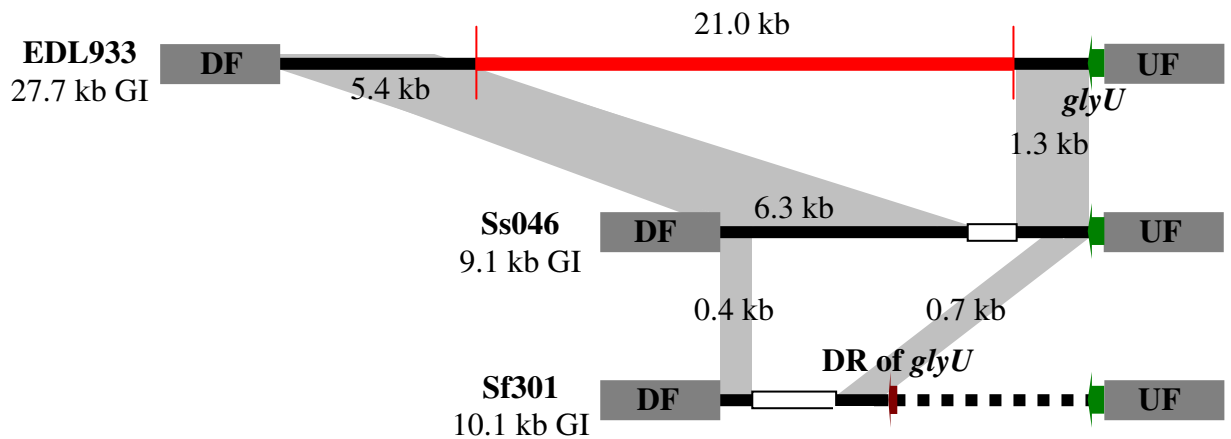


Figure A2. 12. Schematic showing a comparison between the island DNA at *glyU* in EDL933, Ss046 and Sf301.

Dark grey boxes containing ‘UF’ and ‘DF’ represent the upstream and downstream conserved flanking regions respectively, black lines indicate island DNA, the red line indicates the island DNA that harbours the *Salmonella*-like T3SS present in EDL933 but absent in the other strains. The broken black line represents the distinct 7.6 kb prophage-like element found adjacent to *glyU* in Sf301. Light grey areas indicate regions of nucleotide identity over 95% and white boxes indicate IS which could have played a role in the deletion of the EDL933 island DNA in Ss046 and Sf301. K12 MG1655 is not shown as it is very similar to the Ss046 GI. Drawing is not to scale.

These results indicate that the T3SS in EDL933 could be a diverse, ancient GI that was acquired by an ancestral ‘*E. coli*’, and may still be playing a role in the virulence of EHEC and could be ‘locked’ into the genome as it does not harbour an integrase. However, the majority of the island appears to have since been deleted in the other *E. coli* and *Shigella* strains. In *Shigella* this could be due to negative selection after acquisition of the virulence plasmid encoded T3SS, which may have provided a selective advantage.

The location of the IS elements in the Ss046, Sf301 and K12 MG1655 islands, suggests that these mobile elements may have played a role in the deletion of island DNA in this region. This again suggests that IS elements may play a role in the deletion/inactivation of island DNA which becomes redundant to the host organism.

A2.6.3 S104 (*S. flexneri* 1a strain) *glyU* U# Results

S104 was the only *S. flexneri* strain characterised from the U-arm, and the results indicated the presence of an element that had not been found previously downstream of *glyU* in any of the sequenced genomes. The sequence data shows that the amplicon walks into an integrase gene that is only found present in Sf2457T, also this integrase gene is found associated with another *gly* tRNA elsewhere in the genome, not *glyU*. In Sf2457T, this *gly* tRNA is found at the end of a string of three other tRNA loci, has only 69% nucleotide identity to *glyU* and is not found in any of the other *E. coli* or *Shigella* genomes (see Figure A2. 13). These are therefore likely to be non-essential tRNA genes that were acquired by horizontal gene transfer, harboured on a GI. Islands have been reported to sometimes contain tRNA genes and they are believed to help promote expression of the island DNA (K. Rajakumar personal communication). This *gly* associated island has been described previously in more detail as the T-2 GI by (Chen and Schneider, 2006), the authors describe it as a T7-like prophage that was acquired after the differentiation of the two sequenced *S. flexneri* 2a strains, because it is not found in Sf301. It has also has since been broken into at least two separate pieces, possibly by IS911 and has a total length of 3.8 kb. However, there is no mention of whether the entire region is part of a larger GI. The authors also show that there are another seven of these ‘T7’ islands associated with *gly* tRNA loci (including the *glyU* Islander defined GI, which is found in both Sf301 and Sf2457T) across the Enterobacteriaceae, ranging in length from 3.8 kb to 7.9 kb. They have similar organisation, they all encode integrases with high amino acid similarity to each other, they all insert into *gly* tRNA genes with the integrase gene orientated in the same direction, showing that they are P4-like prophages and they all have a nearly identical DR. This shows that these GIs are all part of the same group and the authors also provide evidence to show that some of the T7 islands may have undergone recombination with one another at some point. The role these islands are playing in virulence is yet to be determined.

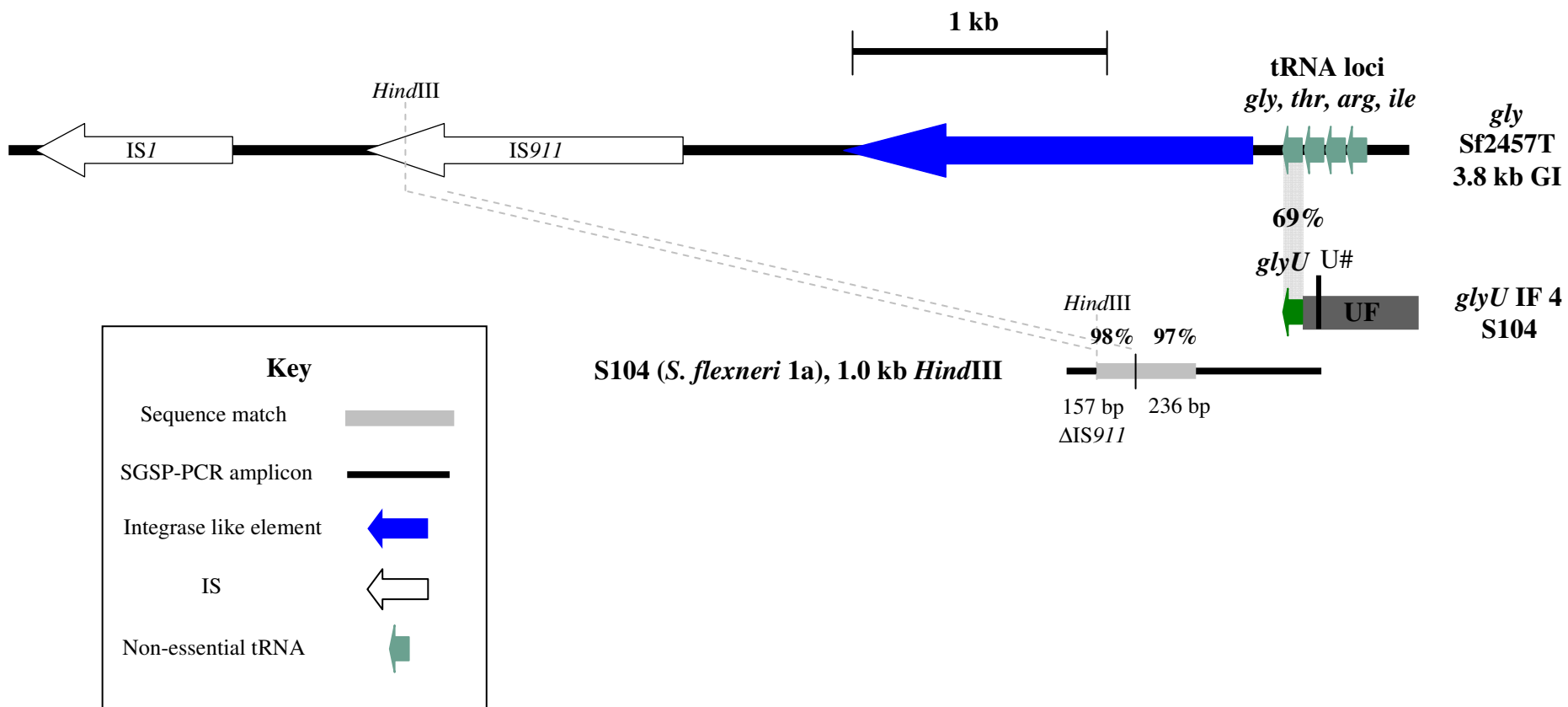


Figure A2. 13. S104 (*S. flexneri* 1a strain) *glyU* U-arm SGSP-PCR results.

These results are a good example of the ‘major codon hypothesis’ and highlight the way in which some prophages can insert into more than one member of a group of tRNA loci, even though some loci always appear to be the preferred choice. In the case of the T-2 island, these results suggest that in S104 the *glyU* site was not occupied by the other T7 island that is found in both Sf301 and Sf2457T, therefore the T-2 island was able to insert at the *glyU* locus preferentially.

In Sf2457T there is an IS911 downstream of the intact integrase gene, whereas in S104, the integrase gene has been truncated by part of an IS911 (see Figure A2. 13), the T-2 element may therefore be locked into the genome of S104.

S. dysenteriae

No sequence data was yielded from the U# SGSP-PCRs, so the three *S. dysenteriae* strains remain uncharacterised from the U-end of the GI. The D-arm results show that S102 and S103 both harbour a small part (78 bp) of the *glyU* island D-arm that is adjacent to the DF. This sequence is found in 16 of the *E. coli* and *Shigella* genomes available on the NCBI (see Table A2. 6), however it has the highest nucleotide identity to Sf301. The rest of the sequence obtained further into the GI is comprised of mosaic IS elements (see Figure A2. 11). Such a short length of specific island sequence makes it difficult to classify these GIs, therefore they have been given split assignments and put into a separate island family to the other *Shigella* strains – *glyU*-IF1 (see Table 5.1).

The sequence data from the S101 *glyU* D# amplicon indicated the presence of mosaic IS elements only, and no *glyU* DF sequence was obtained so this amplicon could not be confirmed as specific, therefore this strain-tRNA locus is still designated as uncharacterised. However, as the IS elements sequence data obtained was very similar to the S102 and S103 sequences, it is very likely that this amplicon is specific and there has been deletion of some

of the D-arm and DF DNA upstream of the D# sequence, for this reason I have included it in Figure A2. 11.

S. boydii

None of the tRIP-negative *S. boydii* strains produced any specific SGSP-PCR amplicons with the U or D primers, so they remain uncharacterised. Two of the strains were also tested using the additional enzyme libraries, but these also produced only non specific amplicons (see Table A2. 8). *In silico* analysis of the fully sequenced Sb227 genome indicated that the D# region is not present, so if this is also the case in the case in the strains tested in this study, this would explain the generation no specific SGSP-PCR amplicons from the D#.

Overall at this tRNA locus, five strain-tRNA loci remain uncharacterised and SGSP-PCR from the U# produced very few specific SGSP-PCR amplicons when compared to other tRNA sites. As the U# is very likely to be well conserved, this initially raised questions regarding the annealing of the *glyU* U# to the template. However, no problems were encountered with tRIP, or when sequencing from this primer, therefore a more likely hypothesis as to the generation of very few specific SGSP-PCR amplicons is that the GC content of the putative island DNA in the uncharacterised strains is likely to be relatively low. This hypothesis is based on the fact that the GC content of the GIs found at *glyU* in the characterised strains are generally low (35-40%, see Table A2. 6), and the recognition sequences of most of the six-base cutting restriction enzymes used for SGSP-PCR (*HindIII* and *EcoRV* are exceptions) have GC contents of 50% or more, so there are likely to be very few of these sites within the island DNA. In the characterised strains, most of the restriction sites found in the *glyU* associated islands' U and D-arms are found within the IS elements that disrupt the island DNA, and these have GC contents of 50-55%.

In future experiments, restriction enzymes with recognition sites of less than 50% may be more useful choices for SGSP-PCR, especially at tRNA loci where there is known to be significant prophage activity.

A2.6.4 *pheU* and *selC*

Table A2. 9. SGSP-PCR results of the *pheU* tRIP-negative strain-tRNA loci

<i>pheU</i> U# - T7# SGSP-PCR results							
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)					Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	
K12 MG1655						~0.4	
<i>S. dysenteriae</i> 3	S101	N ^a		N	~1.3 ^b	N	563 [SK#]
<i>S. dysenteriae</i> 9	S102	N		~1.5 F	~1.3	~0.4 F	
<i>S. dysenteriae</i> 6	S103	N		N	~1.3	~0.4	
<i>S. flexneri</i> 6	S110				~1.3		750 [SK#]
<i>S. boydii</i> 1	S116	~4.0		~2.6	~1.3	~0.4	1501 [SK#] UC ^c
<i>S. boydii</i> 2	S117	~4.0		~2.6	~1.3	~0.4	
<i>S. boydii</i> 3	S118	N		N	~1.3	~0.4	
<i>S. boydii</i> 4	S119				~1.3		693 [SK#]
<i>S. boydii</i> 7	S120				~1.3		124 [U#]

Table A2. 10. SGSP-PCR results of the *selC* tRIP-negative strain-tRNA loci

<i>selC</i> U# - T7# SGSP-PCR results							
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)					Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sa</i> II	
K12 MG1655			~1.3				
<i>S. flexneri</i> 1a	S104		~0.75	~0.85			
<i>S. flexneri</i> 1b	S105		~0.75	~0.85			
<i>S. flexneri</i> 2a	S106		~0.75	~0.85			
<i>S. flexneri</i> 2b	S107		~0.75	~0.85			
<i>S. sonnei</i>	S108		~0.75	~0.85			
<i>S. flexneri</i> X	S111		~0.75	~0.85			
<i>S. flexneri</i> Y	S112		~0.75	~0.85			
<i>S. sonnei</i>	S113		~0.75	~0.85			
<i>S. sonnei</i> bio a	S114		~0.75	~0.85			
<i>S. sonnei</i> bio g	S115		~0.75	~0.85			353 [SK#]

^a Indicates that no amplicon was generated

^b Text highlighted in bold indicates that the amplicon was sequenced

^c The addition of ‘UC’ indicates that the unclipped sequence was used in the analysis. The sequence run failed by the Phred20 (Ewing and Green, 1998) standard used by the sequencing company; however the unclipped (low-quality) sequence still provided some meaningful information.

SHI-3-like islands are associated with *pheU* in *S. boydii* and *S. dysenteriae*

The *pheU* U-arm SGSP-PCR results show that all of the *S. dysenteriae*, *S. boydii* strains and S110 (*S. flexneri* 6 strain) harbour an integrase gene that is most similar to the P4-like integrase gene associated with *pheU* in Sb227, CFT073 and *pheV* in Ss046 (*pheU*-IF1, see Figure A2. 14), indicating that elements with a similar integrase gene occupy both the *pheU* and *pheV* loci across *E. coli* and *Shigella*, again highlighting the role bacteriophage have played in the dissemination of prophage-like island DNA in the Enterobacteriaceae. Across the strains in this study, the *pheU* integrase gene is most likely to be harboured on a SHI-3-like element (see Table 1.3) which has been previously found associated with *pheU* in various *S. boydii* strains and an *S. dysenteriae* 2 strain (Purdy and Payne, 2001). The evidence for the presence of SHI-3-like elements in S116 (*S. boydii* 1) and S117 (*S. boydii* 2) is even stronger, as SGSP-PCR produced amplicons that walked up to 4.0 kb into the GI, and the sequence data corresponded to Sb227, which harbours a SHI-3 like element at *pheU* (see Figure A2. 14). The total size of the SHI-3-like GI present at *pheU* in the sequenced Sb227 chromosome cannot be defined by tRIP as the DF is missing in this strain.

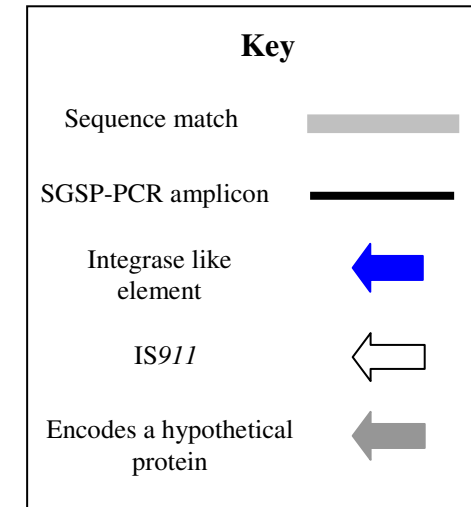
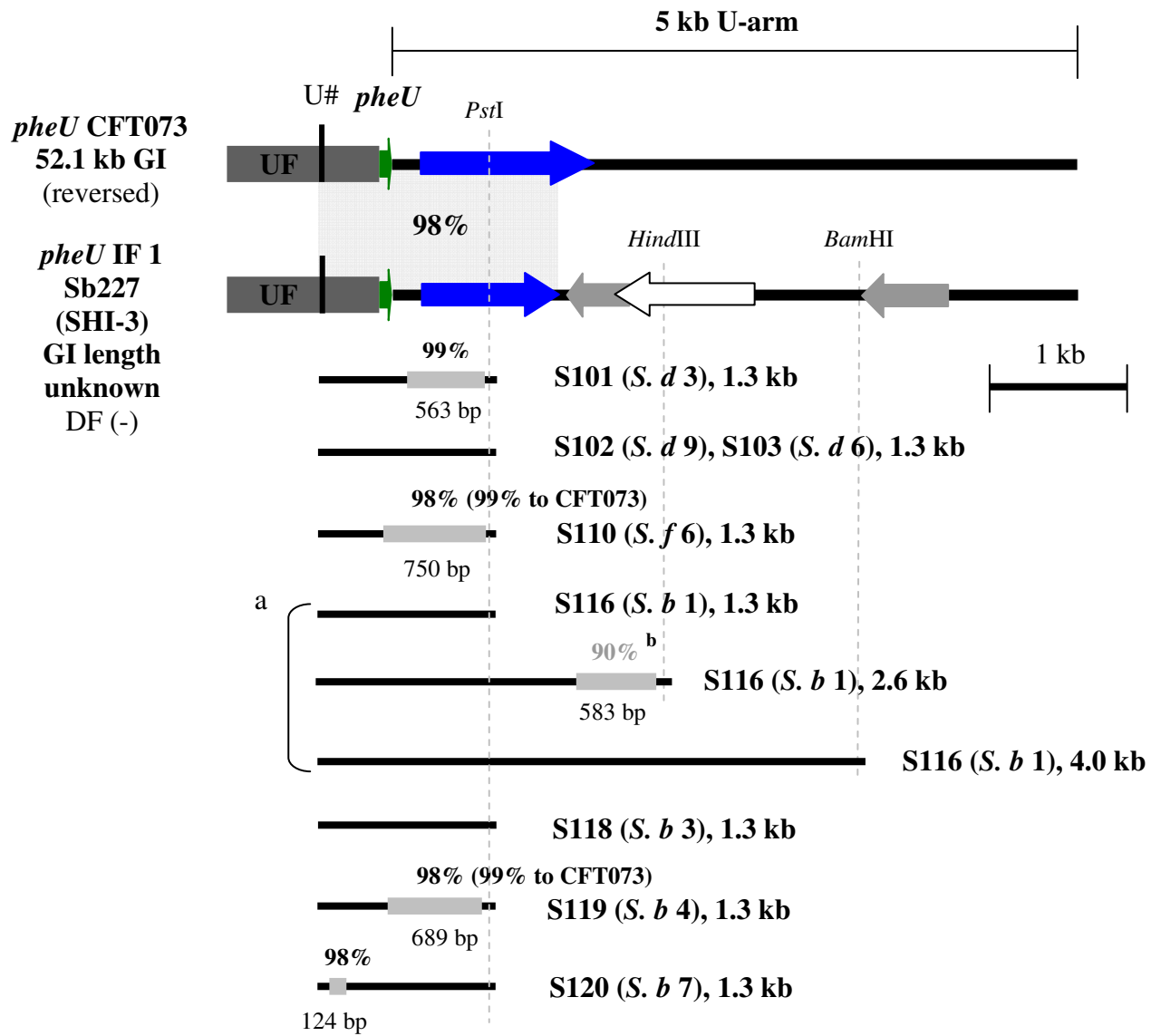


Figure A2. 14. *pheU* U-arm SGSP-PCR results for the *S. boydii* and *S. dysenteriae* strains.

^a S117 produced the same size U# SGSP-PCR amplicons as S116 (see Table A2. 9).

^b Sequence run failed, however the unclipped (low quality) sequence hit to the region indicated

SHI-2-like islands are harboured by *S. flexneri* and *S. sonnei* at *selC*

In all of the *S. flexneri* and *S. sonnei* strains, *pheU* was empty (apart from S110 [*S. flexneri* 6 strain] which is more *S. boydii*-like in its overall GI content, see Table 5.1); however, in these strains, SGSP-PCR indicated that the *selC* locus was occupied with an integrase gene that has the highest nucleotide identity to the integrase gene found at the start of the SHI-2 (*selC*-IF1), which has been previously found associated with *selC* in *S. flexneri* and *S. sonnei* (Moss *et al.*, 1999) (see Figure A2. 15). SHI-2 is similar to SHI-3 in that both islands are a similar size, both harbour identical aerobactin genes, and contain P4-like integrase genes; however, their structures and GC contents differ (see Table 1.3 also). The GC content of SHI-3 is more similar to the GC content of the *Shigella* chromosome, indicating that it may have been acquired at an earlier time than SHI-2 and has undergone amelioration (Purdy and Payne, 2001).

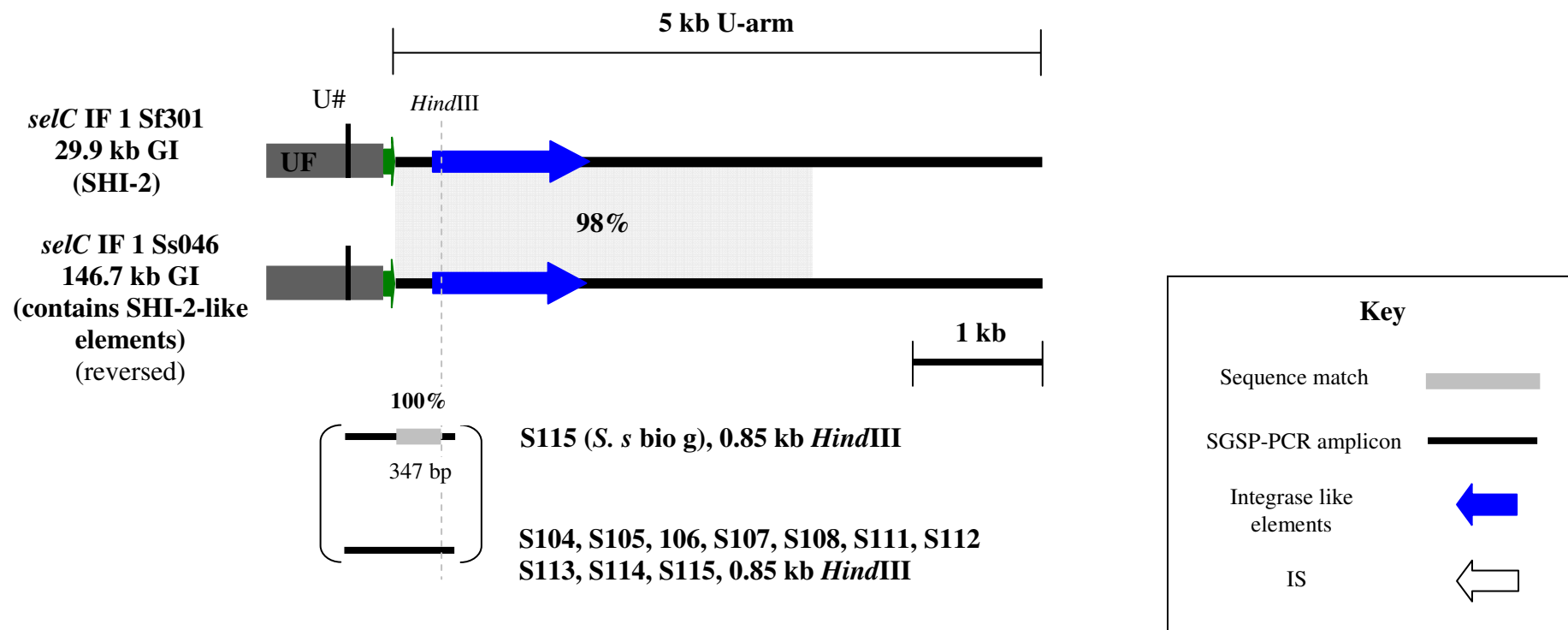


Figure A2. 15. *selC* U-arm SGSP-PCR results for the *S. flexneri* and *S. sonnei* strains

Therefore the tRIP and SGSP-PCR results for *selC* and *pheU* coincide with the literature, as SHI-2-like elements were found associated with *selC* in all of the typical *S. flexneri* and *S. sonnei* strains studied and *pheU* was found to be empty. Whereas at *selC*, *S. boydii* harbours two core DNA genes *yicK* and *yicL* downstream of *selC* which account for the presence of tRIP amplicons in the strains tested (see Table A2. 5 and Figure A2. 16), these genes are deleted in *S. flexneri* and *S. sonnei* (Moss *et al.*, 1999, Vokes *et al.*, 1999) and the *S. boydii* strains harbour SHI-3-like elements at *pheU* (Purdy and Payne, 2001).

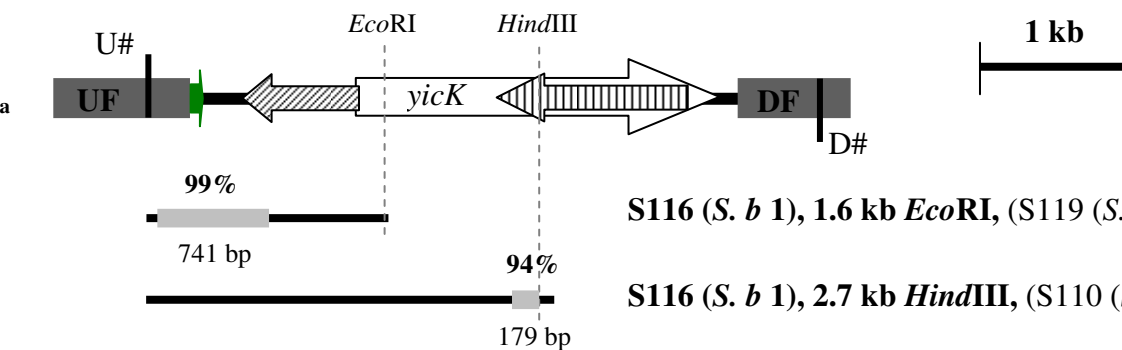
As the literature so far indicates that across *S. dysenteriae* the SHI-3 has only been found in *S. dysenteriae* 2, it would be worth probing the *pheU* locus across many more *S. dysenteriae* strains because SHI-3-like elements were found associated with *pheU* in all three of the *S. dysenteriae* strains screened in this study. This suggests that it may also be present in many more *S. dysenteriae* strains.

All of the *Shigella* strains tested have signatures that indicate the presence of an iron transport like island at either *selC* or *pheU*, but not at both, suggesting that the SHI-3 elements in *S. boydii* and *S. dysenteriae* (which were possibly acquired prior to SHI-2 ((Purdy and Payne, 2001), see page A2-49 also) could be acting in a selfish manner and preventing similar SHI-2 like elements from occupying the *selC* locus. This ‘selfish phage’ observation was also seen across the *S. flexneri* strains that harbour the *ipaH* prophage-like GI at the *serU* tRNA locus, indicating that many of the prophages that become incorporated into the host cells chromosome may act in a truly selfish way.

It was also interesting to see that the size of the *selC* tRIP amplicons produced by the *S. dysenteriae* strains only, were 3.3 kb, whereas the *S. boydii* strains and *S. flexneri* 6 strain produced different size amplicons (see Figure A2. 16, Table 3.2 and Table A2. 5). Therefore

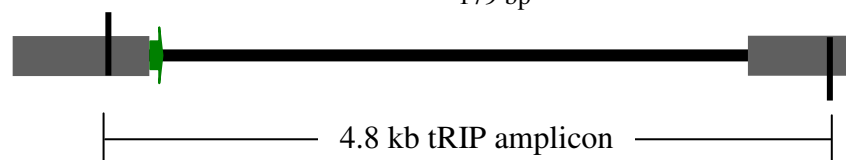
tRIP PCR at this site along with other could be used to quickly identify a *Shigella* isolate as either a *S. boydii* or a *S. dysenteriae*, which may be a useful diagnostic tool. It would therefore be interesting to see if this observation is consistent when extended across many strains of each species.

selC
Sb227
4.5 kb 'islet' ^a
(reversed)

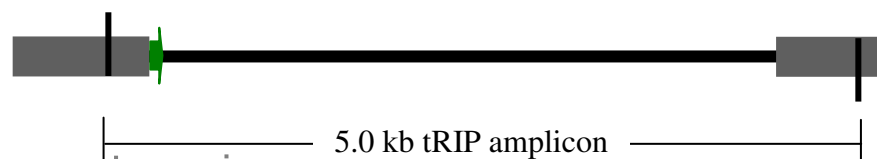


S116 (*S. b* 1), 1.6 kb *EcoRI*, (S119 (*S. b* 4) and S110 (*S. f* 6) had the same RP)

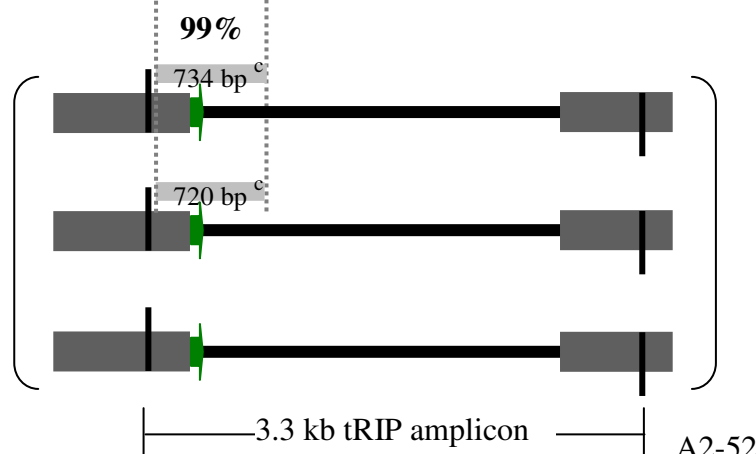
S116 (*S. b* 1), 2.7 kb *HindIII*, (S110 (*S. f* 6) had the same RP)



S116 (*S. b* 1) ^b, S117 (*S. b* 2),
S119 (*S. b* 4) ^b, S110 (*S. f* 6) ^b



S118 (*S. b* 3)



S101 (*S. d* 3)

S102 (*S. d* 9)

S103 (*S. d* 6)

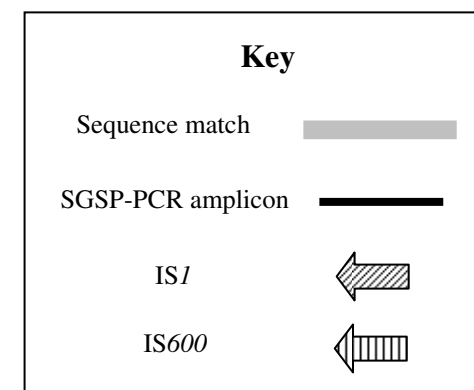


Figure A2. 16. *selC* sequence results of larger than expected tRIP amplicons

^a Core DNA disrupted by IS elements.

^b These larger than expected tRIP amplicons were characterised using SGSP-PCR, see amplicons produced and sequence data above.

^c tRIP amplicon sequence data.

A2.6.5 *pheV*

Table A2. 11. SGSP-PCR results of the *pheV* tRIP negative strain-tRNA loci

<i>pheV</i> U# - T7# SGSP-PCR results									
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)							Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	<i>Eco</i> RV	<i>Hinc</i> II	
K12 MG1655			~1.9						
<i>S. dysenteriae</i> 3	S101	N ^a	N	N	~0.25	N	~2.0	~0.15	
<i>S. dysenteriae</i> 9	S102	N	N	N	~0.25	N	~2.0	~0.15	
<i>S. dysenteriae</i> 6	S103	N	N	N	~0.25	N			
<i>S. flexneri</i> 1a	S104	N	N	1.7 ^b	~0.25	N			365 [U#]
<i>S. flexneri</i> 1b	S105	N	N	1.7	~0.25	N			186 [U#]
<i>S. flexneri</i> 2a	S106	~1.4			~0.25				738 [SK#]
<i>S. flexneri</i> 2b ^d	S107	N		N	~0.25				
<i>S. sonnei</i> ^d	S108	N	N	N	N				
<i>S. flexneri</i> 6	S110	N	N	N	~0.25	N	~2.0	~0.15	
<i>S. flexneri</i> X	S111	N	N	~1.7 F ^c	~0.25	N			
<i>S. flexneri</i> Y	S112	N	N	~1.5	~0.25	N			1696 UC ^e [SK#]
<i>S. sonnei</i> ^d	S113	N		N	N				
<i>S. sonnei</i> bio a ^d	S114	N		N	N				
<i>S. sonnei</i> bio g ^d	S115	N		N	N				
<i>S. boydii</i> 1	S116	N	N	N	~0.25	N	~2.0	~0.15	
<i>S. boydii</i> 2	S117	N	N	N	~0.25	N			
<i>S. boydii</i> 3	S118	N	N	N	~0.25	N			
<i>S. boydii</i> 4	S119	N	N	N	~0.25	N	~2.0	~0.15	365 [SK#]
<i>S. boydii</i> 7	S120	N	N	N	~0.25	N			

^a Indicates that no amplicon was generated

^b Text highlighted in bold indicates that the amplicon was sequenced

^c The addition of 'F' after the text indicates that the amplicon was faint

^d Strain was already characterised as harbouring *she* PAI-like elements at this locus by (Al-Hasani *et al.*, 2001a)

^e The addition of 'UC' after the text indicates that the unclipped sequence was used in the analysis. The sequence run failed by the Phred20 (Ewing and Green, 1998) standards used by the sequencing company; however the low quality sequence still provided some meaningful information.

S. flexneri

S107 (*S. flexneri* 2b strain) had already been characterised at the *pheV* locus by Al-Hasani *et al.*, 2001, and was found to harbour a *she* PAI-like element. S106 (*S. flexneri* 2a strain) was characterised by SGSP-PCR from the U# and the sequence derived from the vector primer was found to have 100% nucleotide identity to the corresponding region in the *she* PAI associated P4-like integrase gene found directly downstream of the *pheV* locus, indicating the presence of a *she* PAI-like island at this strain-tRNA locus. This is not surprising as nine other *S. flexneri* 2a strains characterised by Al-Hasani *et al.*, all harboured *she* PAI like elements. The *she* PAI has classic prophage-like features and is present in the Islander database, it also encodes two enterotoxins (see Table 1.3 for more details on the *she*-PAI).

However, the other *S. flexneri* strains had different elements associated with *pheV*. Directly downstream of the *pheV* gene, S104, S105 and S111 were found to harbour the DNA found the start of the tRIP defined *pheV* 8.4 kb distal extension (*pheV*-IF2) or 'flanking GI' that is found immediately downstream of the Islander defined *she* PAI in Sf301 (see Figure A2. 17).

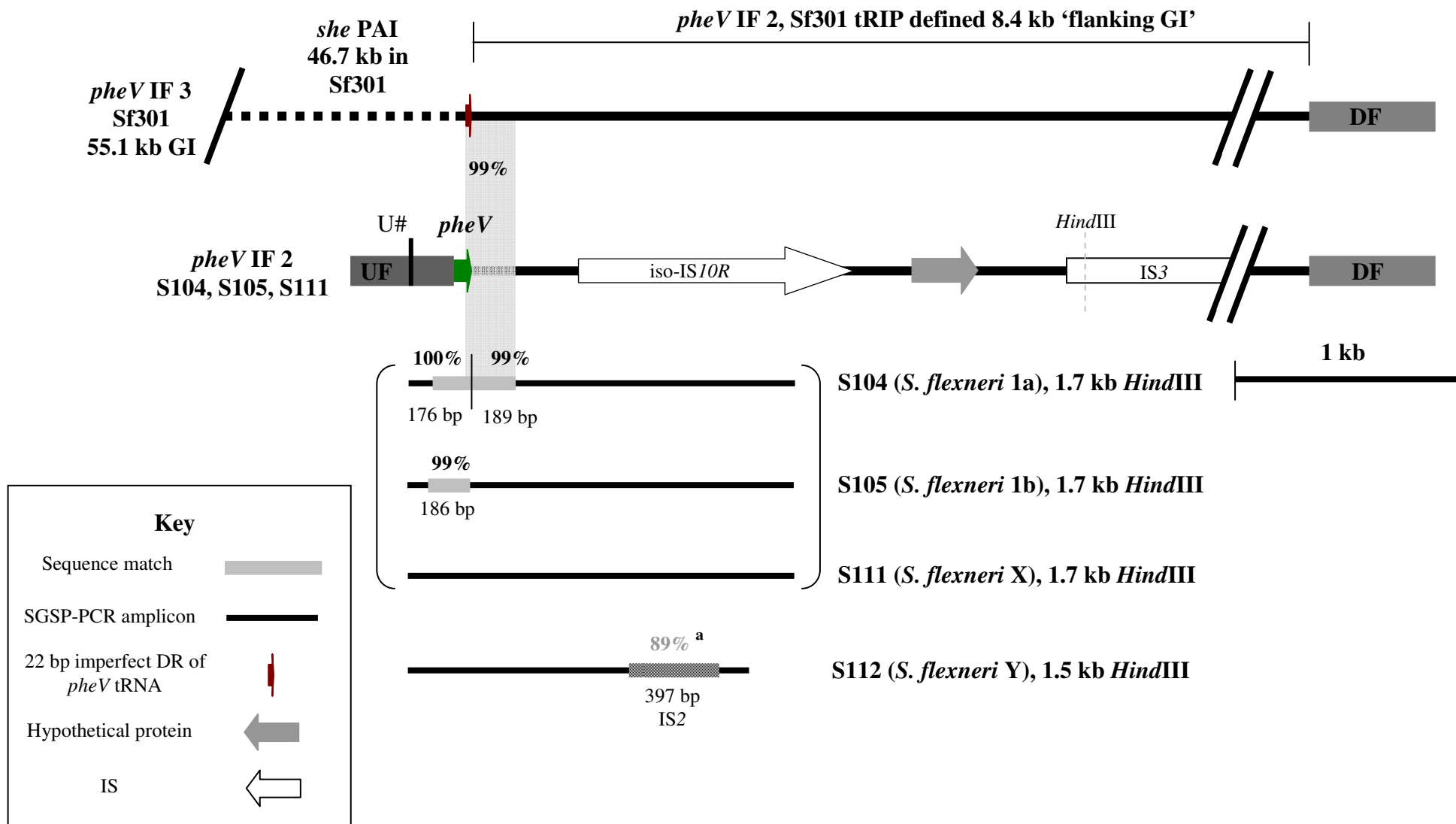


Figure A2. 17. *S. flexneri* strains that yielded U-arm amplicons that walked into the *pheV* associated Sf301 8.4 kb ‘flanking GI’ (*pheV* island family 2).

^a Sequence run failed, however the unclipped (low quality) sequence hit to IS2-like sequences not present in this region in any of the known genomes.

Therefore, in the above four strains the *pheV* associated *she* PAI has either deleted precisely or was never acquired. Even so, the tRNA^{acc} defined 8.4 kb flanking GI is very likely to be true island DNA as it has signatures of foreign origin: it contains DNA that is not found in the other sequenced *E. coli* genomes (see Table A2. 6 also), it has an overall GC content of 44.2% which is much lower than that of the Sf301 average (50.9%) and lower than the islander defined *she* PAI (49.1%), indicating that it is a distinct element. This GI is likely to be more ancient island DNA that was acquired by an ‘ancestral *S. flexneri*’ strain prior to its differentiation into the many *S. flexneri* strains. It is unlikely to be playing a direct role in the virulence of *Shigella* due to its small size and disruption by IS and is likely to be ‘locked’ into the chromosome of the strains that harbour it.

Interestingly the restriction profile of the element is different between Sf301 and the three *S. flexneri* strains, indicating that the iso-IS10R present in Sf301 may be absent in the other strains (see Figure A2. 17). This suggests that IS activity has possibly played a role in shaping the element and that it was originally a much larger island. The claim that mobile genetic elements are playing a role in this region, is this is further backed up by the evidence that the distal end of the *she* PAI is known to be less stable than the proximal end (Al-Hasani *et al.*, 2001).

S112 produced a *Hind*III amplicon of 1.5 kb, that when sequenced from the vector primer indicated the presence of IS2-like sequence downstream of *pheV*, suggesting that *she* PAI-like elements are not present at this site (see Figure A2. 17). However, IS elements are commonly found in *Shigella* genomes, as the amplicon was not sequenced from the U# and this sequence

has never been found in this location in any of the other genomes, it cannot be confirmed as specific to the *pheV* UF region, therefore the strain-tRNA locus has been classified as ‘uncharacterised’ (see Table 5.1). Even so, the *pheV* U primer had not produced any non-specific amplicons with previous SGSP-PCRs, and it was the only amplicon generated in the reaction, so it is very likely that this amplicon is specific.

S. sonnei

The *S. sonnei* strains used in this study were all found to harbour *she* PAI-like elements at *pheV* by Al-Hasani *et al.*, 2001).

S. boydii* and *S. dysenteriae

The island DNA found associated with *pheV* was the same in all of the characterised *S. boydii* and *S. dysenteriae* strains and S110 (*S. flexneri* 6 strain) which is more *S. boydii*-like in its island content. The island sequence obtained from the S119 amplicon had 100% nucleotide identity to the corresponding *pheV* island sequence in the Sb227 genome (*pheV*-IF1, see Figure A2. 18). The DNA walked into is also found associated with *pheV* in other *E. coli* and *Shigella* genomes (see Table A2. 6) but not in any of the sequenced *S. flexneri* or *S. sonnei* genomes, these are occupied with *she* PAI-like elements at *pheV*. The sequence corresponded to the *gspL* gene, which is one of the genes that encodes a type II secretion system (T2SS) known as the general secretion pathway (GSP). The GSP is responsible for the secretion of extracellular proteins such as proteases and toxins such as heat-labile toxins and aerolysins in many Gram-negative bacteria, however it is not usually found in *E. coli* or *Shigella* (Pugsley, 1993). More recently, studies have shown that there are homologs of the *gsp* genes found in other Gram-negatives and some of these are found associated with *pheV*, indicating that they were acquired by horizontal gene transfer (Francetic *et al.*, 1998). In Sb227 and Sd197 there is an 8.0 kb cluster of novel *gsp* genes that is likely to have been deleted in K12MG1655 and

is not found in any of the other *Shigella* genomes available (see Figure A2. 18). Their products have the highest similarity to the corresponding *gsp* genes products from enterotoxigenic *E. coli* (ETEC) that secrete heat-labile enterotoxin, one of its main virulence factors, which in turn has high homology to the *Vibrio cholerae* secretion system used to secrete cholera toxin, which is encoded by a prophage (Tauschek *et al.*, 2002, Yang *et al.*, 2005). This again provides further evidence to show that these genes were acquired by the above *Shigella* strains from a foreign source via horizontal gene transfer, possibly bacteriophage mediated and maybe prior to their divergence from an ancestral '*E. coli*'. The role these genes are playing in *Shigella* is yet to be determined; it would be interesting to find out if the *pheV* associated *gsp* genes in any of the *S. boydii* and *S. dysenteriae* strains do encode a functional T2SS and if so, does it also secrete a similar protein to the ETEC enterotoxin.

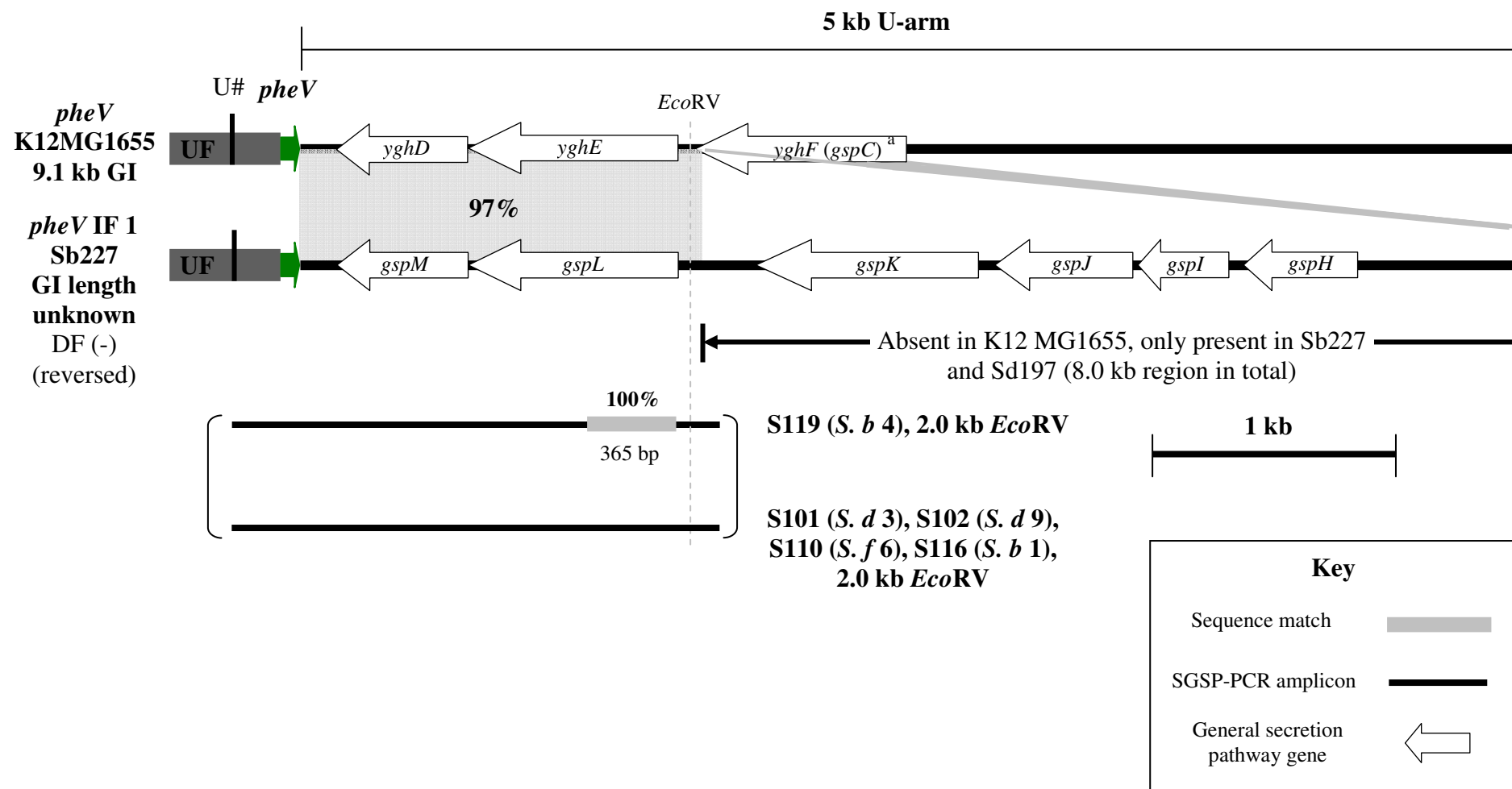


Figure A2. 18. *pheV* U-arm SGSP-PCR results for the *S. dysenteriae*, *S. boydii* strains and *S. flexneri* 6 strain.

Absence of *she*-PAI-like elements at *pheV* in *S. boydii* and *S. dysenteriae*

The results of the U# SGSP-PCRs show that *she* PAI-like elements are not associated with *pheV* in any of the *S. boydii* and *S. dysenteriae* strains characterised, maybe due to the presence of SHI-3-like elements at their cognate *pheU* loci, which were possibly acquired at an earlier stage (Purdy and Payne, 2001) as the GC content of the SHI-3 is 51%, closer to that of the core chromosome; also the *she* PAI and SHI-3 have a number of similar features such as the presence of a similar P4-like integrase gene and ORFs with high sequence similarity to prophage genes harboured on the EDL933 LEE PAI. Therefore the *pheU* associated SHI-3-like elements in these strains could be acting in a selfish manner, preventing the host chromosome from being occupied with another similar element.

These results suggest that the *S. boydii*, *S. dysenteriae* strains and *S. flexneri* 6 strain screened in this study have similar origins that are distinct from *S. flexneri* and *S. sonnei*, which could explain why the island DNA associated with *pheV* in these species is not found in any of the sequenced *S. flexneri* or *S. sonnei* genomes. This is also seen at other tRNA loci, suggesting that *S. boydii* and *S. dysenteriae* are more '*E. coli*-like' than *S. flexneri* and *S. sonnei* in their island content.

A2.6.6 *metV*

Table A2. 12. SGSP-PCR results of the *metV* tRIP negative strain-tRNA loci

<i>metV</i> U# - T7# SGSP-PCR results							
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)					Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	
K12 MG1655							
<i>S. dysenteriae</i> 3	S101	N ^a	N	N	~1.7 ^b	N	821 [T7#], 257 [U#]
<i>S. dysenteriae</i> 9	S102	N	N	N	N	N	
<i>S. dysenteriae</i> 6	S103	N	N	N	N	N	
<i>S. flexneri</i> 6	S110						
<i>S. boydii</i> 1	S116	N	N	N	~3.0 F ^c	N	189 [SK#]
<i>S. boydii</i> 2	S117	N	N	N	~3.0 F	N	
<i>S. boydii</i> 3	S118	N	N	N	N	N	
<i>S. boydii</i> 4	S119						
<i>S. boydii</i> 7	S120						

<i>metV</i> D# - T7# SGSP-PCR results							
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)					Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sa</i> I	
K12 MG1655							
<i>S. dysenteriae</i> 3	S101	~2.4		N	mult F ^a		
<i>S. dysenteriae</i> 9	S102	~2.4		N	N		
<i>S. dysenteriae</i> 6	S103	~1.5 F		N	~2.0		713 [T7#], 656 [D#]
<i>S. flexneri</i> 6	S110	~2.4					
<i>S. boydii</i> 1	S116	~2.4		N	N		675 [SK#]
<i>S. boydii</i> 2	S117	~2.4		mult F	N		
<i>S. boydii</i> 3	S118	~2.4		N	N		
<i>S. boydii</i> 4	S119	~2.4					
<i>S. boydii</i> 7	S120	~1.6			~3.0		

^a Indicates that no amplicon was generated

^b Text highlighted in bold indicates that the amplicon was sequenced

^c The addition of 'F' after the text indicates that the amplicon was faint

^d Indicates that multiple faint bands were produced, making it difficult to select the specific amplicon

Analysis

In all of the tRIP negative *Shigella* strains, the SGSP-PCR results show the presence of DNA with the closest nucleotide identity to the 6.3 kb islet associated with *metV* in Sb227 (see Figure A2. 19). The DNA walked into is not found in any element present in islander, however the GC content of the sequence is 10% lower than the Sb227 average (51.2%) and it is only found in four of the genomes present on the NCBI database; two *S. boydii* strains (Sb227, BS512 [unfinished]) and two pathogenic *E. coli* strains (CFT073 (UPEC) and F11 (ExPEC [unfinished], GenBank accession no. AAJU01000000), (see Table A2. 6 also). Therefore the evidence to show that this DNA is horizontally acquired is strong. The similar sequence present in CFT073 is an element at the distal end of the tRIP defined 32.7 kb *metV* GI. It is clearly a distinct element to the rest of the island as it is found between the conserved DF and a 14 bp DR of the 3' terminus of the *metV* tRNA locus. This indicates that in CFT073, the islet was possibly acquired prior to another insertion event which in turn has resulted in the formation of a DR, and the presence of a 27.1 kb GI directly downstream of the tRNA locus. Therefore in CFT073, the islet is regarded as a 'flanking GI' (see Figure A2. 20).

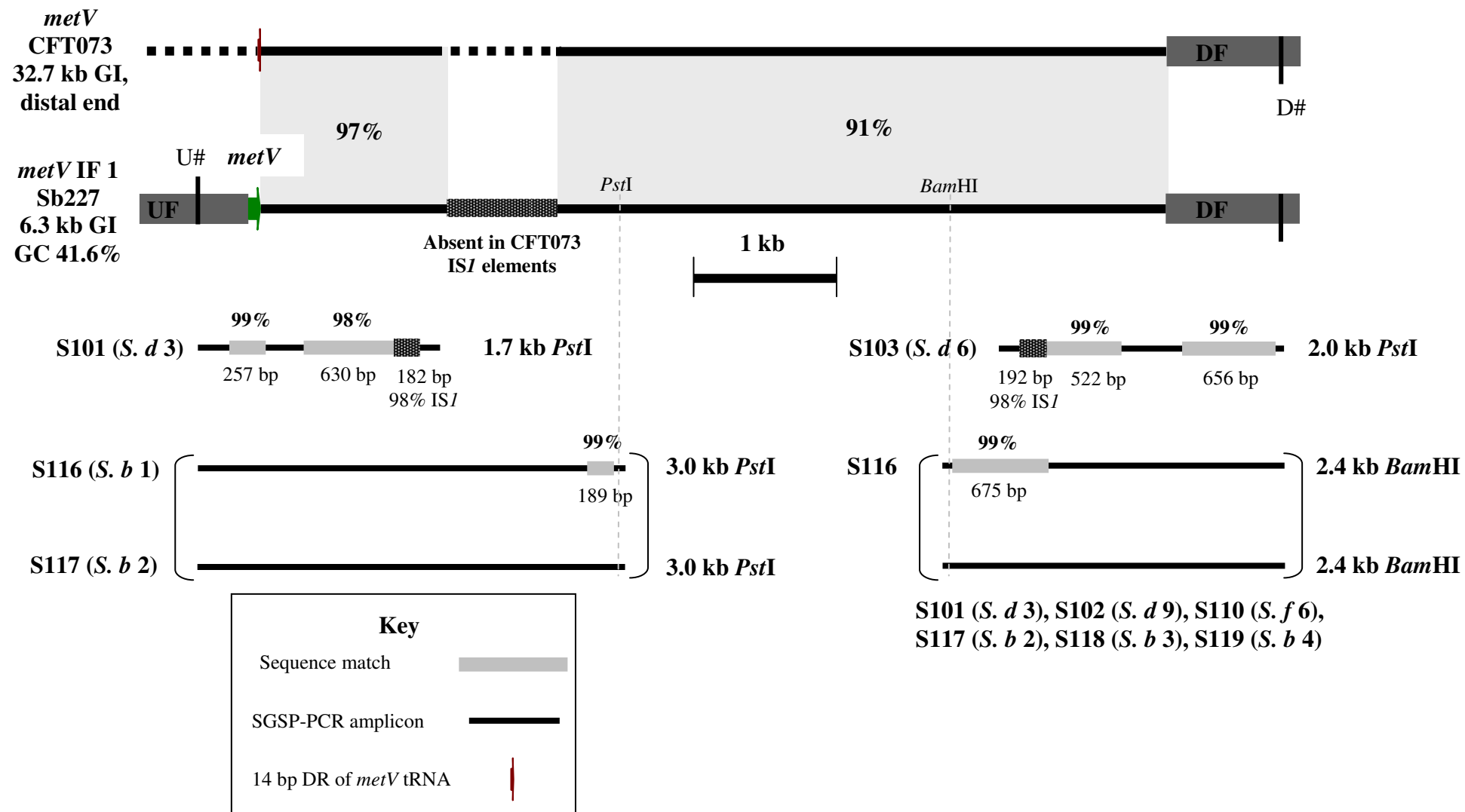


Figure A2. 19. *metV* SGSP-PCR results for the *S. boydii*, *S. dysenteriae* and *S. flexneri* 6 strains

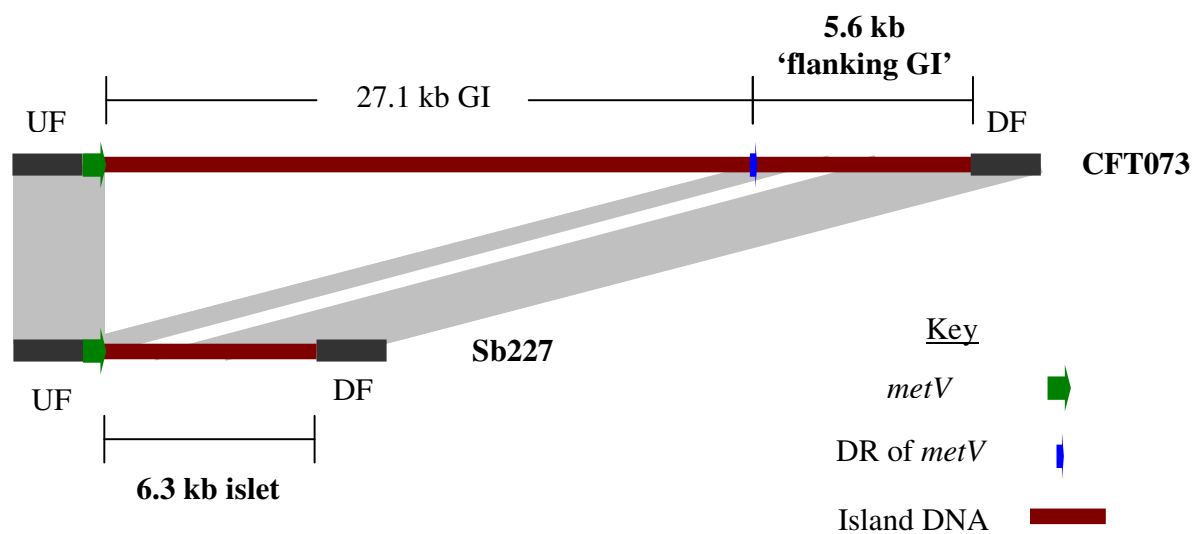


Figure A2. 20. Schematic showing a comparison of the island DNA at *metV* in CFT073 and Sb227. Figure is not to scale.

Light grey areas indicate regions of nucleotide identity over 90%. Figure is not to scale.

None of the genes present on the Sb227 islet seem to be associated with virulence, and it does not have any 'classic' PAI features apart from its relatively low GC content (41.6%). The genes present are involved in metabolism of amino acids (see Table A2. 13) and therefore may help improve the overall 'fitness' of the host organism, enabling it to survive in more nutrient limiting conditions. Traits of this nature when present in a pathogenic bacterium such as *Shigella* or UPEC, may help the organism colonise or persist in the host, in this context the island would be regarded as a PAI (Schmidt and Hensel, 2004).

The 'typical' *S. flexneri* and *S. sonnei* strains screened in this study, as well as the sequenced strains were all tRIP positive at the *metV* locus and confirmed as being 'empty' sites. Also, there are no signatures associated with *metV* in the *S. flexneri* and *S. sonnei* strains to indicate the previous presence of the islet, such as 'scar'-like remnants of the sequence, or IS elements.

Therefore the locus is truly ‘empty’. This shows that they are likely to have never acquired this DNA, possibly because they are likely to be of separate lineage to the *S. dysenteriae* and *S. boydii* strains.

These results also indicate that the *S. flexneri* 6 strain (S110) is *S. boydii*-like in its island content at the *metV* locus. This is also observed at other tRNA loci (see Table 5.1).

Table A2. 13. Contents of the tRIP defined 6347 bp Sb227 *metV* islet.

start	end	length (bp)	strand	synonym	gene	COG	product
2704227	2705174	948	-1	SB02700		COG0111	putative phosphoglycerate dehydrogenase
2705412	2705915	504	-1	SB02701	<i>insB</i>	COG1662	IS1 ORF 2
2705990	2706172	183	1	SB02702			putative IS1 encoded protein
2706143	2706619	477	-1	SB02703		COG0794	putative isomerase
2706588	2707796	1209	-1	SB02704		COG1168	cystathionine beta-lyase
2707796	2709376	1581	-1	SB02705			
2709408	2710055	648	-1	SB02706		COG3711	conserved hypothetical protein

A2.6.7 *ssrA*

Table A2. 14. SGSP-PCR results of the *ssrA* tRIP negative strain-tRNA loci

<i>ssrA</i> U# - T7# SGSP-PCR results									
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)							Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	<i>Eco</i> RV	<i>Hinc</i> II	
K12 MG1655				~1.2					
<i>S. dysenteriae</i> 3	S101	N	N	N ^a	~2.6 F ^b		~0.3	~1.5 ^c	574 [SK#]
<i>S. dysenteriae</i> 9	S102	N	N	N	N		~0.3	~1.5 F	
<i>S. dysenteriae</i> 6	S103	N	~2.6 F	N	N				
<i>S. sonnei</i>	S108	N	N	N	N		~0.3	N	
<i>S. flexneri</i> 6	S110	N	N	N	N		~0.3	~1.5	511 [SK#]
<i>S. sonnei</i>	S113	N	N	N	N				
<i>S. sonnei</i> bio a	S114	N	~4.0 F	N	N				
<i>S. sonnei</i> bio g	S115	N	N	~4.1 F	N				
<i>S. boydii</i> 1	S116	N	~2.6 F	~4.0 F	N		~0.3	~1.5	534 [U#]
<i>S. boydii</i> 2	S117	N	~2.6 F	N	N				
<i>S. boydii</i> 3	S118	N	~2.6 F	N	N				733 [SK#]
<i>S. boydii</i> 4	S119	N	N	N	N		~0.3	~1.5	
<i>S. boydii</i> 7	S120	N	N	N	N				

ssrA D# - T7# SGSP-PCR results									
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)							Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	<i>Eco</i> RV	<i>Hinc</i> II	
K12 MG1655									
<i>S. dysenteriae</i> 3	S101	N	N	N	N	N	N	N	
<i>S. dysenteriae</i> 9	S102	N	N	N	N		N	N	
<i>S. dysenteriae</i> 6	S103	N	N	N	N				
<i>S. sonnei</i>	S108	N	~1.8 F	N	N	N	~1.5 F	~3 F	483 [D#]
<i>S. flexneri</i> 6	S110	N	N	N	N	N	N	N	
<i>S. sonnei</i>	S113	N	~2.5 F	~2.5 F	N	N			178 [D#]
<i>S. sonnei</i> bio a	S114	N	~2 F	~2.5 F	N	N			138 [D#]
<i>S. sonnei</i> bio g	S115	N	~1.8 F	N	N	N			165 [D#]
<i>S. boydii</i> 1	S116	N	N	N	N	N	N	N	
<i>S. boydii</i> 2	S117	N	N	N	N				
<i>S. boydii</i> 3	S118	N	N	N	~3 F				
<i>S. boydii</i> 4	S119	N	N	~5 F	N		N	N	
<i>S. boydii</i> 7	S120	N	N	~1.6	N	N			

^a Indicates that no amplicon was generated

^b The addition of 'F' after the text indicates that the amplicon was faint

^c Text highlighted in bold indicates that the amplicon was sequenced

S. sonnei

Only one of the *S. sonnei* strains was characterised at the *ssrA* locus; S113 produced a D primer generated SGSP-PCR amplicon, the sequence obtained had the highest nucleotide identity to the sequence present in the D-arm of the K12 MG1655 *ssrA* GI (designated island family 2, see Figure A2. 21). This sequence is likely to be island DNA as it has a relatively low GC content and the region comprises a number of putative ORFs that are present in less than half of the *E. coli* strains available on the NCBI database (see Table A2. 6). The other three *S. sonnei* strains produced D# generated SGSP-PCR amplicons and representatives were sequenced, restriction profile and sequence analysis indicated that the sequence obtained was the same as the corresponding region in the Ss046 genome. However, as the part of the DF containing the D primer is inverted in this strain, the D primer generated amplicons did not walk into island DNA but into the Ss046 IS1 disrupted DF. Any U# generated SGSP-PCR amplicons from these strains were too faint to sequence; therefore the *ssrA* associated island DNA in S114, S115 and S108 was designated as uncharacterised (see Table 5.1).

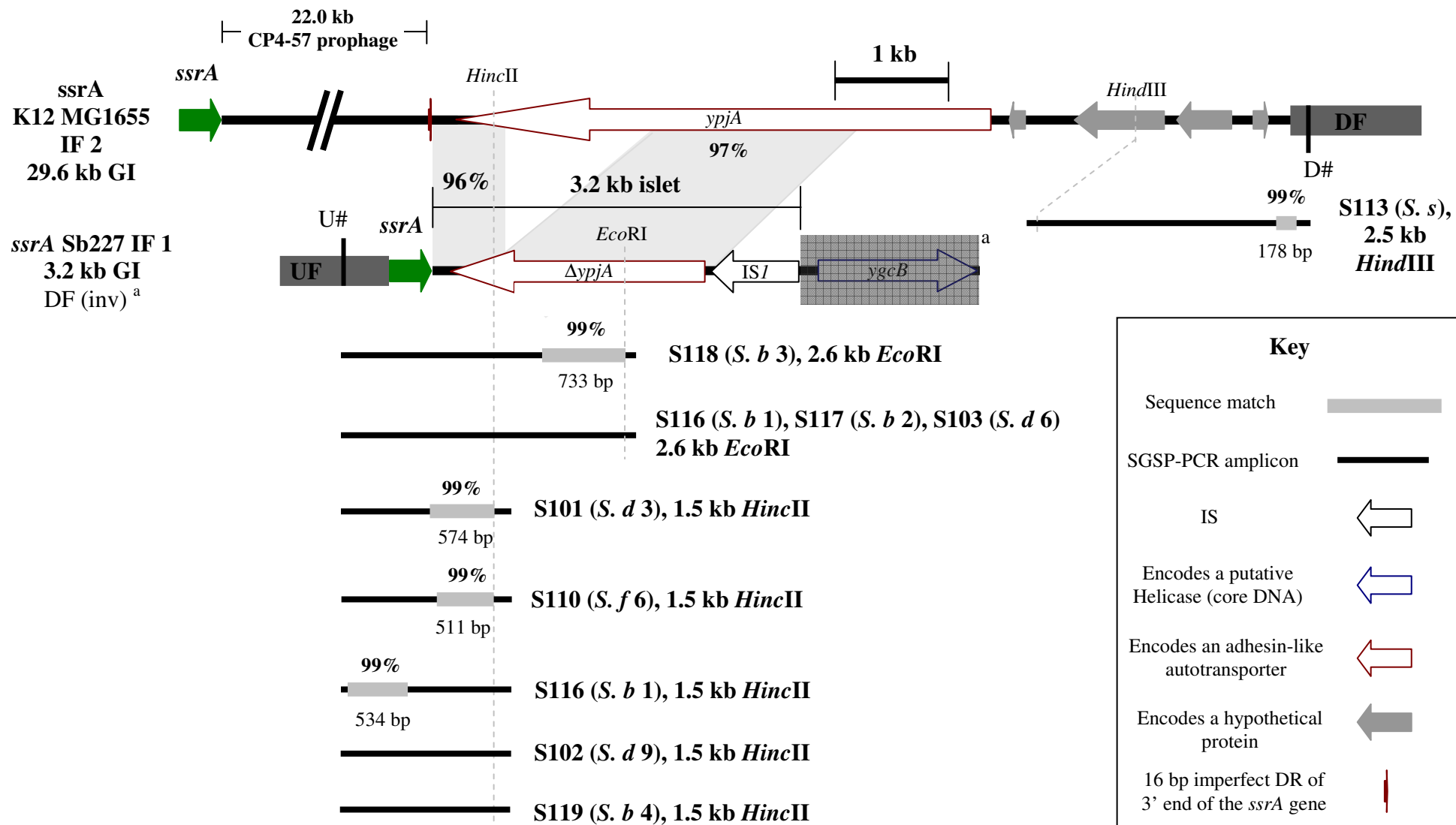


Figure A2. 21. *ssrA* SGSP-PCR results for the *S. sonnei*, *S. boydii*, *S. dysenteriae* and *S. flexneri* 6 strains.

^a The Sb227 *ssrA* DF is inverted and found 102 kb downstream of *ssrA*, however, blastn analysis indicated that the entire core region that comprises the DF present downstream of the truncated *ypjA* gene was inverted with respect to K12 MG1655. The inversion event was possibly mediated by two *IS1* elements that flank the corresponding sequence. The Sb227 *ssrA* GI was therefore determined to be 3.2 kb.

S. dysenteriae* and *S. boydii

All of the *S. boydii* and *S. dysenteriae* strains apart from S120 (*S. boydii* 7 strain) were characterised and found to harbour the same DNA as is found in the U-arm of the Sb227 *ssrA* GI (*ssrA*-IF1). The size of this GI could not be determined by *in silico* tRIP as the DF is inverted; however, it was found that the 99 kb core region that includes the *ssrA* DF is inverted with respect to K12 MG1655, the size of the Sb227 *ssrA* island was therefore determined to be 3.2 kb (see Figure A2. 21).

The island sequence data obtained from the above characterised strains indicates the presence of a disrupted gene designated *ypjA*, the intact gene is also found in the D-arm of the K12 MG1655 *ssrA* GI and encodes a putative adhesin-like autotransporter that when expressed leads to increased adhesion to solid surfaces and increased biofilm formation (Roux *et al.*, 2005). However, as this gene is disrupted in the above *Shigella* strains, it is likely to be non-functional.

S. flexneri

The only tRIP negative strain was S110 (*S. flexneri* 6 strain), the *ssrA* U-arm island sequence obtained had the highest nucleotide identity to the corresponding region in Sb227. This again,

as with other tRNA loci, indicates that this strains' *ssrA* island content is '*S. boydii* – like' rather than '*S. flexneri* – like'.

A2.6.8 *serX*

Table A2. 15. SGSP-PCR results of the *serX* tRIP negative strain-tRNA-loci.

<i>serX</i> U# - T7# SGSP-PCR results							
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)					Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	
K12 MG1655							
<i>S. dysenteriae</i> 3	S101			N ^a	~1.4 F		
<i>S. dysenteriae</i> 9	S102				~1.4		
<i>S. dysenteriae</i> 6	S103				~1.4		
<i>S. flexneri</i> 2b	S107			N	~2.8 F ^b		
<i>S. sonnei</i>	S108	N	N	N	N	~1.2 ^c	465 [SK#]
<i>S. flexneri</i> 6	S110			N	~2.8		392 [SK#]
<i>S. sonnei</i>	S113	N	N	N	N	~1.4	183 [SK#]
<i>S. sonnei</i> bio a	S114	N	N	N	N	~1.2	303 [U#]
<i>S. sonnei</i> bio g	S115	N	N	N	N	~1.2	
<i>S. boydii</i> 1	S116				~1.4		
<i>S. boydii</i> 2	S117				~1.4		
<i>S. boydii</i> 3	S118				~1.4		
<i>S. boydii</i> 4	S119				~1.4		350 [SK#]
<i>S. boydii</i> 7	S120	N	N	N	N	~0.5 F	

No D# SGSP-PCRs were performed at the *serX* site

^a Indicates that no amplicon was generated

^b The addition of ‘F’ after the text indicates that the amplicon was faint

^c Text highlighted in bold indicates that the amplicon was sequenced

S. flexneri

The U-arm sequencing results show that S107 (*S. flexneri* 2b strain) and S110 (*S. flexneri* 6 strain) harbour SRL PAI-like sequences (see Table 1.3) at the *serX* locus (see Table 5.1 also), this is further validated by their resistance to the antibiotics Ap, Cm, Str, and Tc (see Table 2.3) In all of the other *S. flexneri* strains, *serX* is unoccupied, perhaps due to the presence of the Sf301 prophage at the *serU* locus in these strains, which may be acting as a selfish entity, preventing other similar prophage sequences from integrating into the hosts chromosome (see section 7.5 for more details).

S. dysenteriae*, *S. boydii* and *S. sonnei

The U# results show that in all of the strains representative of the above three ‘species’ characterised by SGSP-PCR that they harbour the same DNA as is found associated with *serX* in Sb227 and Ss046 (see Figure A2. 22), a 1.5 kb islet that is comprised of a 263 bp repeat region directly downstream of the tRNA, followed by 1232 bp of DNA that is found only in Sb227, Ss046 and two other unfinished genomes (*S. boydii* BS512 and *E. coli* B7A [ETEC], GenBank accession no. AAJT000000000), (see Table A2. 6 also). This sequence has a GC content of 40.5% and contains a putative ORF that encodes a ‘hypothetical protein’. These features suggest that this sequence is very likely to be horizontally acquired DNA, possibly the remains of a larger element that has since excised from the chromosome.

The 263 bp repeat region upstream of this, adjacent to the *serX* gene, is also found adjacent to *serX* in K12 MG1655 and is found partially directly downstream of *serX* and *serW* in Sf301, the *serX* associated GI in CFT073 and the *serX* and *serW* associated GIs in EDL933. This region comprises a 9 bp DR of the 3’ terminus of a *ser* RNA locus, the length of the repeat region varies from strain to strain, and is also found partially downstream of *serW* in Sb227 and Ss046. This indicates that mobile generic elements inserted previously at both *serX* and *serW* in Sb227, Ss046, K12 MG1655 and Sf301 have excised in an imprecise manner, leaving

behind the 3' DR of a *ser* tRNA. This also suggests that one or more of these elements, at some time has occupied both *serX* and *serW* in the above strains. This seems plausible as the same prophage-like islands are already known to occupy both *serX* and *serW* in both *Shigella* and *E. coli* (the SRL PAI in *Shigella* (Turner *et al.*, 2003) also seen in this study and the O-islands that encode tellurite resistance in EHEC (Taylor *et al.*, 2002)).

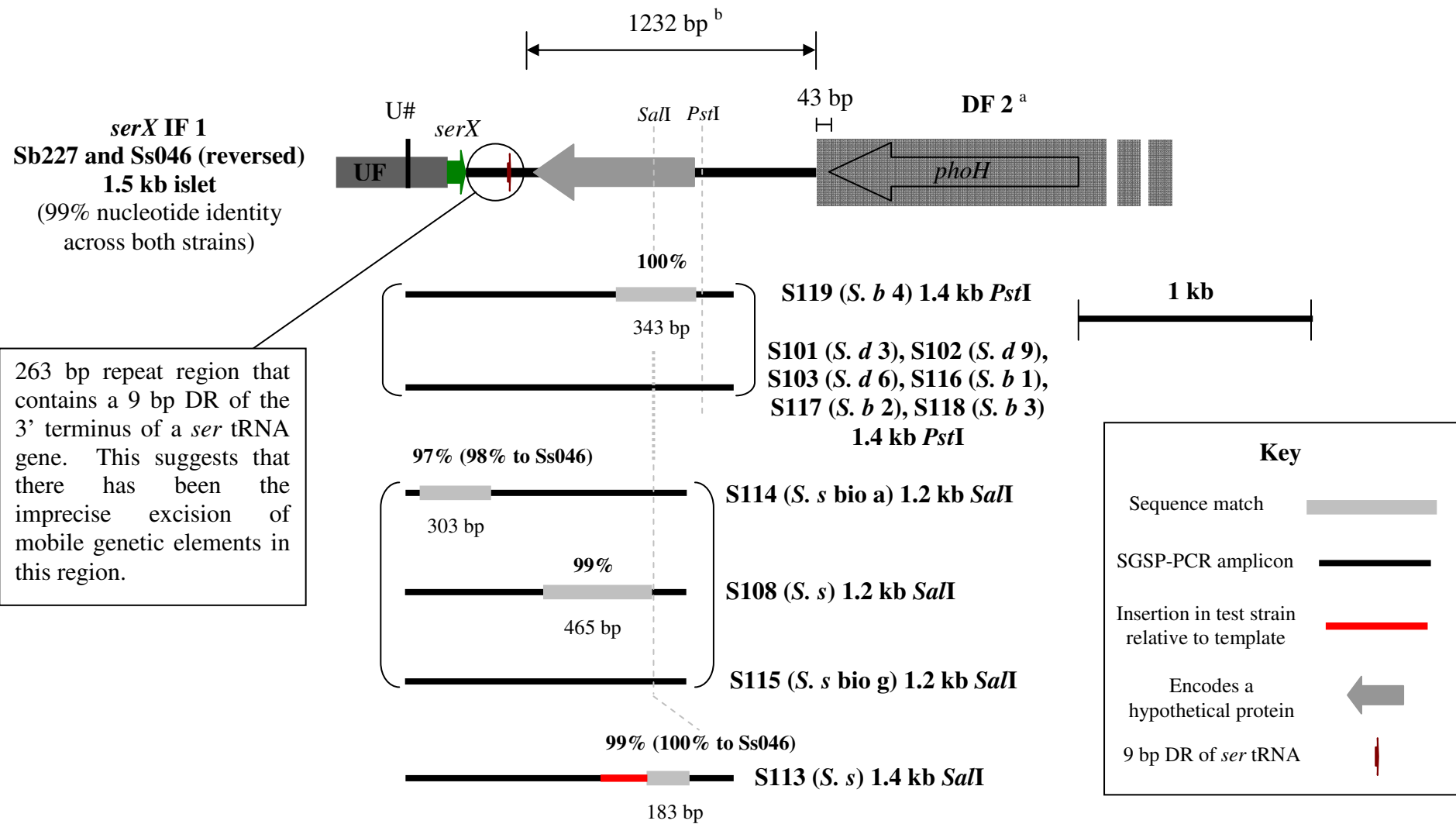


Figure A2. 22. *serX* SGSP-PCR results for the *S. sonnei*, *S. boydii*, and *S. dysenteriae* strains.

^a The original conserved DF is not present in Sb277 and Ss046, so a secondary conserved DF region (DF 2) was used to define the size of the GI.

^b The island DNA is only present in Sb227, Ss046 and the unfinished *S. boydii* BS512 and *E. coli* B7A (ETEC) chromosomes; it has a GC content of 40.5%

A2.6.9 *asnT*

Table A2. 16. SGSP-PCR results of the *asnT* tRIP negative strain-tRNA loci

<i>asnT</i> U# - T7# SGSP-PCR results									
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)							Length of sequence read (bp)
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	<i>Eco</i> RV	<i>Hinc</i> II	[primer used]
K12 MG1655									
<i>S. dysenteriae</i> 3	S101								
<i>S. dysenteriae</i> 9	S102	N ^a	N	~2.0 ^b	N	N	N	~0.3	712 [U# 3] ^c
<i>S. dysenteriae</i> 6	S104								
<i>S. flexneri</i> 1a	S104								
<i>S. flexneri</i> 1b	S105								
<i>S. flexneri</i> 2a	S106								
<i>S. flexneri</i> 2b	S107								
<i>S. sonnei</i>	S108	N	N	N	N	N	N	~0.3 ^e	194 [SK#]
<i>S. flexneri</i> 6	S110	N	N	N	~ 4.0 F ^d	N	N	~0.3 ^e	181 [SK#]
<i>S. flexneri</i> X	S111								
<i>S. flexneri</i> Y	S112								
<i>S. sonnei</i>	S113								
<i>S. sonnei</i> bio a	S114								
<i>S. sonnei</i> bio g	S115	N	N	N	N	N	N	~0.3	
<i>S. boydii</i> 1	S116	N	N	N	N	N	N	~0.3 ^e	581 [SK#]
<i>S. boydii</i> 2	S117								
<i>S. boydii</i> 3	S118								
<i>S. boydii</i> 4	S119	N	N	N	N	N	~2.5 F	~0.3	
<i>S. boydii</i> 7	S120								

<i>asnT</i> D# - T7# SGSP-PCR results									
Species & serotype	Strain code	Restriction library/SGSP-PCR amplicon (kb)							Length of sequence read (bp) [primer used]
		<i>Bam</i> HI	<i>Eco</i> RI	<i>Hind</i> III	<i>Pst</i> I	<i>Sal</i> I	<i>Eco</i> RV	<i>Hinc</i> II	
K12 MG1655									
<i>S. dysenteriae</i> 3	S101	N	~ 0.6	N	N	N			453 [SK#]
<i>S. dysenteriae</i> 9	S102	N	N	N	N	N	N	~0.6 F	
<i>S. dysenteriae</i> 6	S104	N	~ 0.6	N	~0.7 F	~ 1.2 kb			451 [SK#], 310 [SK#]
<i>S. flexneri</i> 1a	S104	N	N	~ 0.4	N	N			309 [D#]
<i>S. flexneri</i> 1b	S105			~0.4	N				
<i>S. flexneri</i> 2a	S106			~0.4	~0.7 F				
<i>S. flexneri</i> 2b	S107			~0.4	~0.7 F				
<i>S. sonnei</i>	S108	N	N	N	N	N	~2.7 F	~3.0 F	
<i>S. flexneri</i> 6	S110	N	N	N	~ 0.7	N	N	~0.6 F	658 [SK#] N/S ^f
<i>S. flexneri</i> X	S111	N	N	~0.4	~0.7 F, ~1.2 F	N			
<i>S. flexneri</i> Y	S112			~0.4	~0.7 F				
<i>S. sonnei</i>	S113	N	N	N	N	N			
<i>S. sonnei</i> bio a	S114	N	N	N	N	N			
<i>S. sonnei</i> bio g	S115	N	N	N	N	N	~2.7 F	~3.0 F	
<i>S. boydii</i> 1	S116	N	N	N	N	N	N	~0.6 F	
<i>S. boydii</i> 2	S117	N	N	N	~0.7 F	N			
<i>S. boydii</i> 3	S118	N	N	N	N	N			
<i>S. boydii</i> 4	S119	N	N	N	~0.7 F	N	N	~0.6 F	
<i>S. boydii</i> 7	S120	N	N	~ 1.6	~0.6 F	~1.2 F			750 [SK#]

^a Indicates that no amplicon was generated.

^b Text highlighted in bold indicates that the amplicon was sequenced.

^c SGSP-PCR with the original *asnT* U# used for tRIP and a second U# produced multiple bands, therefore a third *asnT* U# was used and this produced specific SGSP-PCR amplicons (see Table A2. 1).

^d The addition of 'F' after the text indicates that the amplicon was faint ^e The amplicons were specific but did not walk beyond the tRNA into the putative GI.

^f The SGSP-PCR amplicon was non-specific.

S. flexneri

Figure A2. 23 shows the *asnT* D# results for the characterised *S. flexneri* strains, indicating that they all harbour the same sequence as is present in the D-arm of the Sf301 *asnT* 4.5 kb islet D-arm (*asnT*-IF3). The only sequence data obtained matches numerous IS600 elements; however in this case the D-arms were not designated as 'unclassifiable' as the restriction pattern of the amplicons matched that of the Sf301 D-arm, so it was very likely that the amplicons obtained were specific. There has been some prophage activity at this locus in Sf301, as indicated by the two integrase gene fragments found downstream of *asnT*; however IS activity may have deleted any island DNA that was previously associated with *asnT* in this strain.

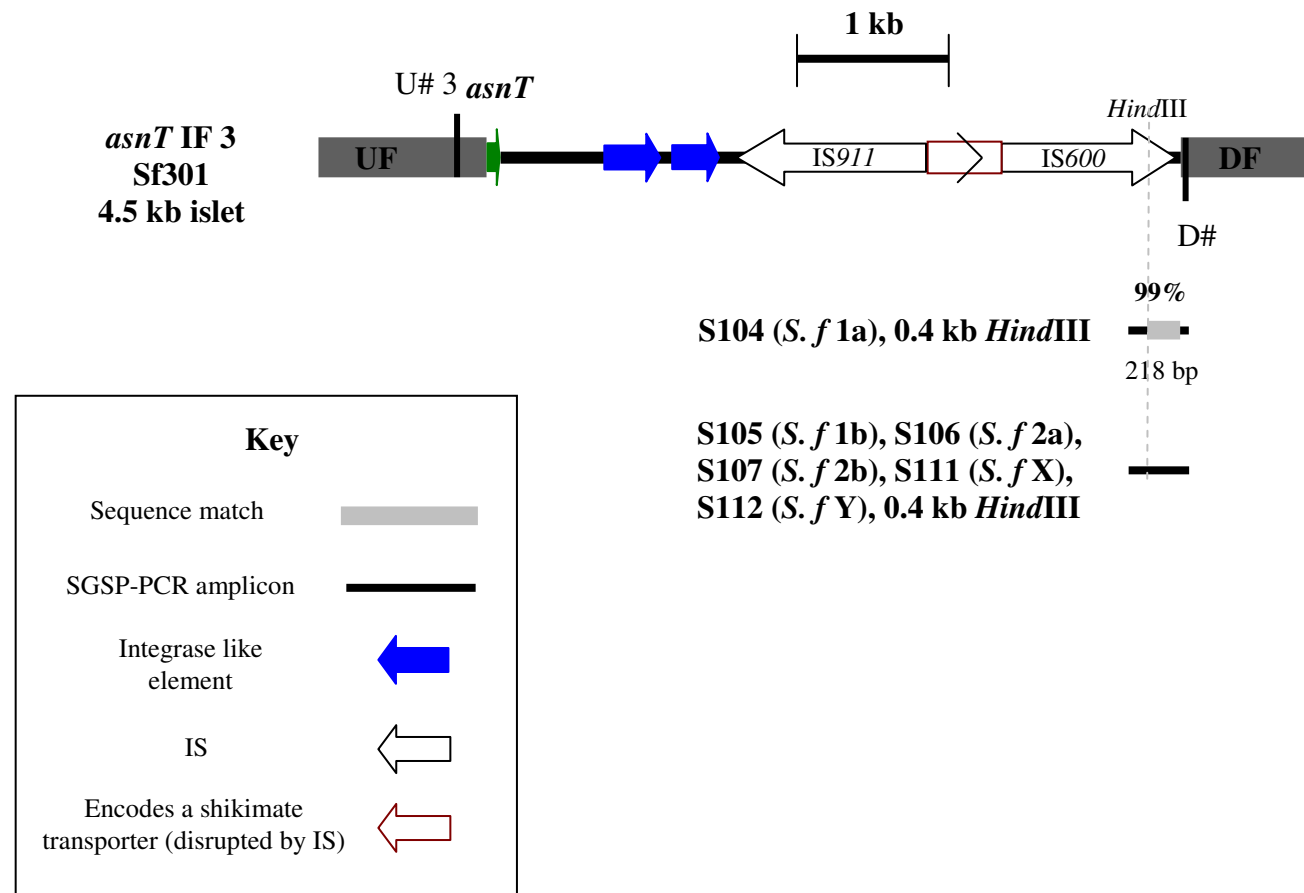


Figure A2. 23. *asnT* D# SGSP-PCR results for the *S. flexneri* strains belonging to island family 3.

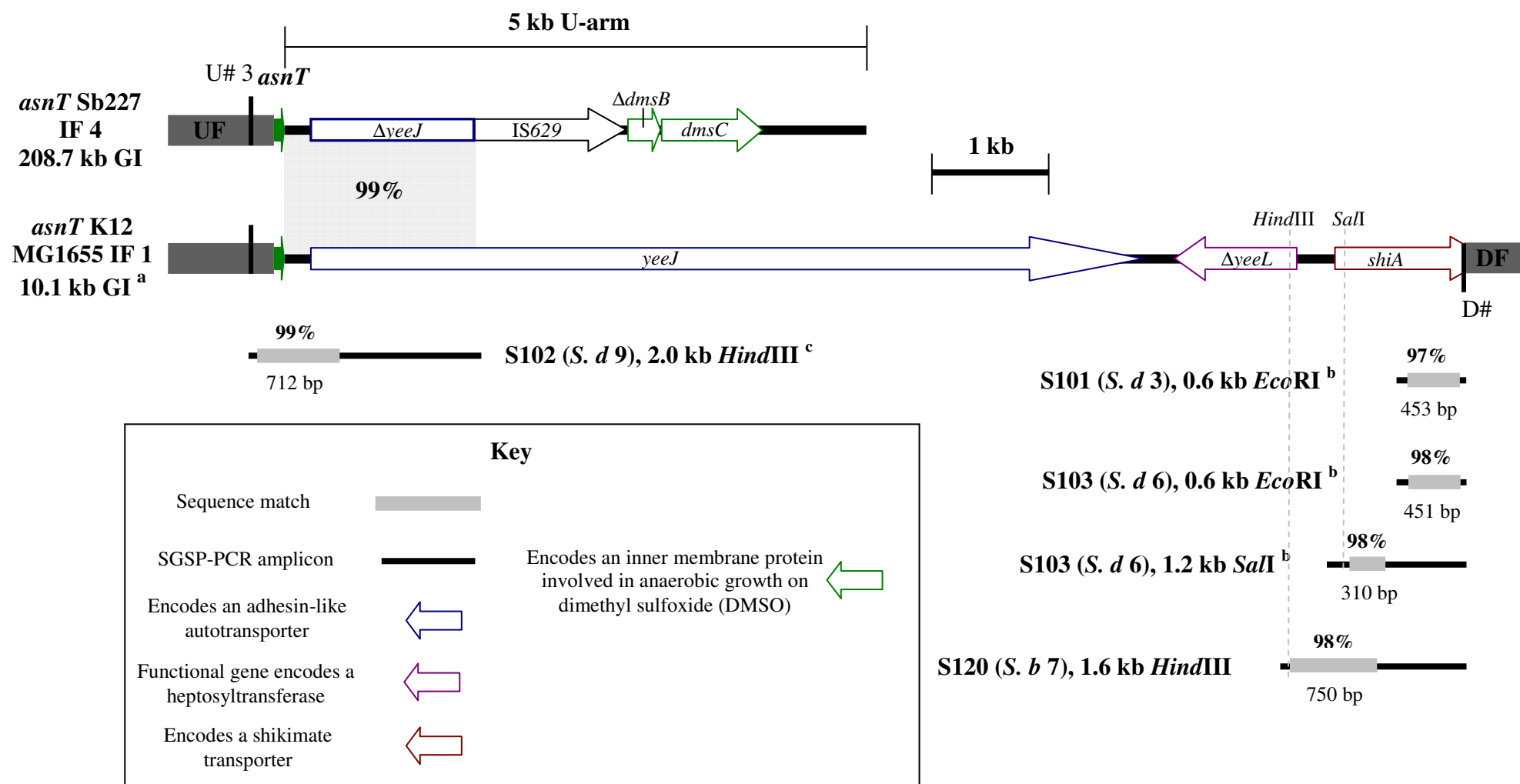


Figure A2. 24. SGSP-PCR results for the *S. dysenteriae* strains and S120 (*S. boydii* 7)

^a The EDL933 *asnT* GI is 11.0 kb in length and has 97% nucleotide identity to the K12 MG1655 *asnT* GI

^b S101 and S103 were designated into *asnT*-IF1, however, as the sequences obtained had the same nucleotide identity and Blastn scores to both the K12 MG1655 and EDL933 *asnT* GI sequences, the D-arms were given split assignments (see Table 5.1)

^c The S102 sequence had the same nucleotide identity and Blastn scores to both the Sb227 and K12 MG1655 *asnT* GI sequences, so the U-arm was given a split assignment (see Table 5.1). However as it was not characterised from the D-arm and the U-arm restriction pattern was different to both Sb227 and K12 MG1655, the GI in this strain was assigned to a unique island family –*asnT*-IF2.

Figure A2. 24 shows the SGSP-PCR results of the *S. dysenteriae* strains and one *S. boydii* strain.

A2.7 *asnV* Results

A2.7.1 Inversion of the *asnV* UF region in *E. coli* CFT073

The initial *in silico* tRIP screen followed by analysis using blastn and Artemis, indicated that compared to the other fully sequenced *E. coli* and *Shigella* strains studied, the CFT073 conserved *asnV* UF was inverted with respect to the DF and that this inversion was across a 4.3 kb region that comprised three *asn* tRNA loci (see Figure A2. 25). The inversion lies between the *asnV* and *asnW* tRNA loci and is exact to the 3' termini of each gene, so that in CFT073 there is no disruption of the tRNA loci. The 54.4 kb CFT073 genomic island is therefore associated with *asnW*. However, as the *asnV* DF is still intact, for the purpose of this study I have classified the GI as associated with the *asnV* locus, but with the UF (inv) (see Table 3.1). Therefore, when the tRIP screen was performed using CFT073 genomic DNA, the result was negative because the U and D primers were in tandem, however in this situation the negative tRIP results was not a true false-negative because of the presence of the large GI. This raised the question as to whether any other *E. coli*, or *Shigella* strains especially, had an inversion at the *asnV* UF, specifically if there were any that also did not harbour a GI at *asnV*, as in this situation, a negative tRIP result would be a false negative, due to the inversion of the U primer and not because of the presence of an island between the U and D flanks.

A2.7.2 Inversion of the *asnV* UF region in *S. sonnei* 046

At the time of this analysis the only other *Shigella* sequences available were the partially sequenced genomes of *Shigella sonnei* 53G and *Shigella dysenteriae* M131649 (M131), which were available on the coliBASE online utility (<http://colibase.bham.ac.uk/>) (Chaudhuri *et al.*, 2004). Blastn analysis showed that *S. sonnei* 53G did not harbour any island DNA at the *asnV* locus, but the *asnV* UF was inverted; an *asnV* U-D tRIP PCR with this strain would therefore be negative, however, this would be a false-negative result. The *S. sonnei* 53G

inversion covers a 6.4 kb region, was likely to have been mediated by two inverted *IS1* elements that flank the inversion and it is in a different location to the CFT073 inversion. However, the inversion still comprised the *asnV* U primer (see Figure A2. 26); the *S. sonnei* 53G sequence also represents the subsequently completely sequenced Ss046 chromosome, as the cognate regions have 99% nucleotide identity).

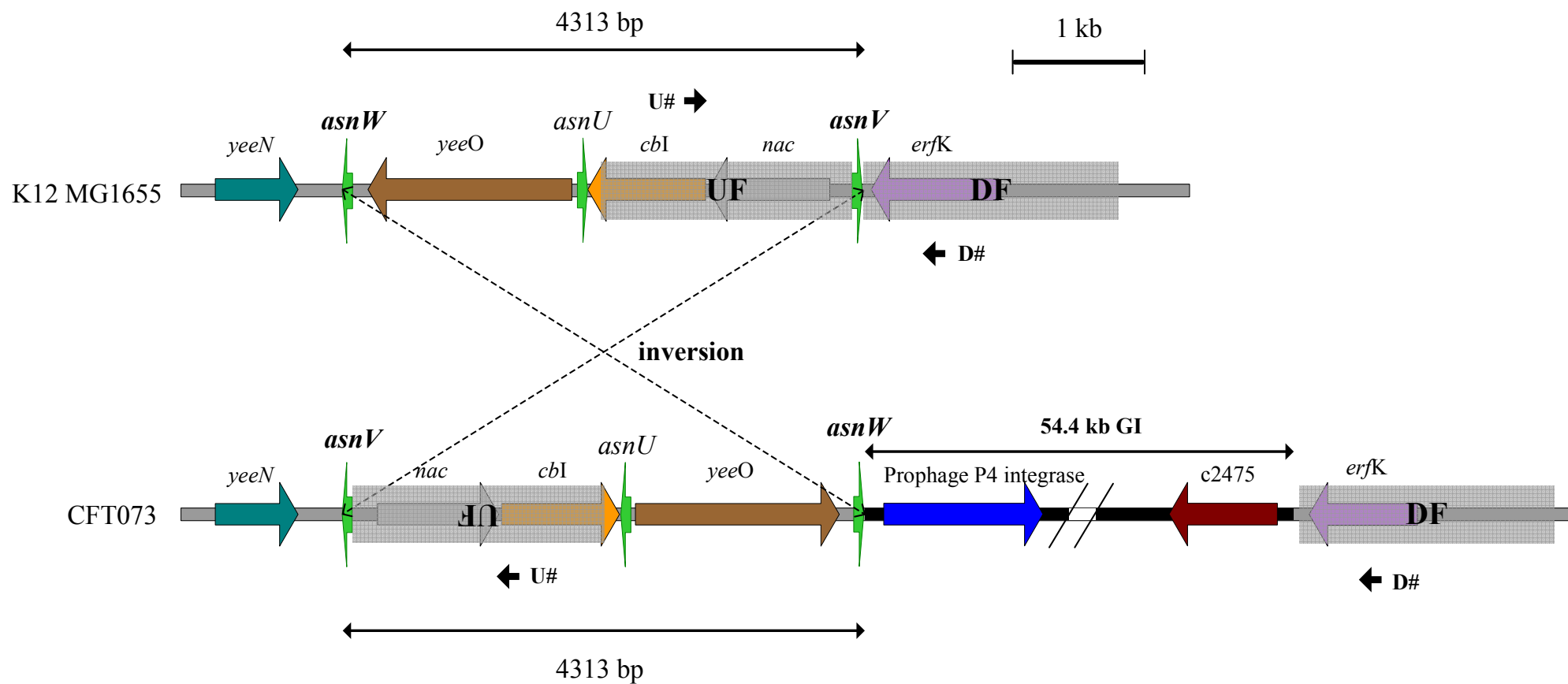


Figure A2. 25. Schematic showing the *asnV*-*asnW* inversion in *E. coli* CFT073 relative to *E. coli* K12 MG1655.

The locations of the *asnV* U and D primers are indicated by the small black arrows

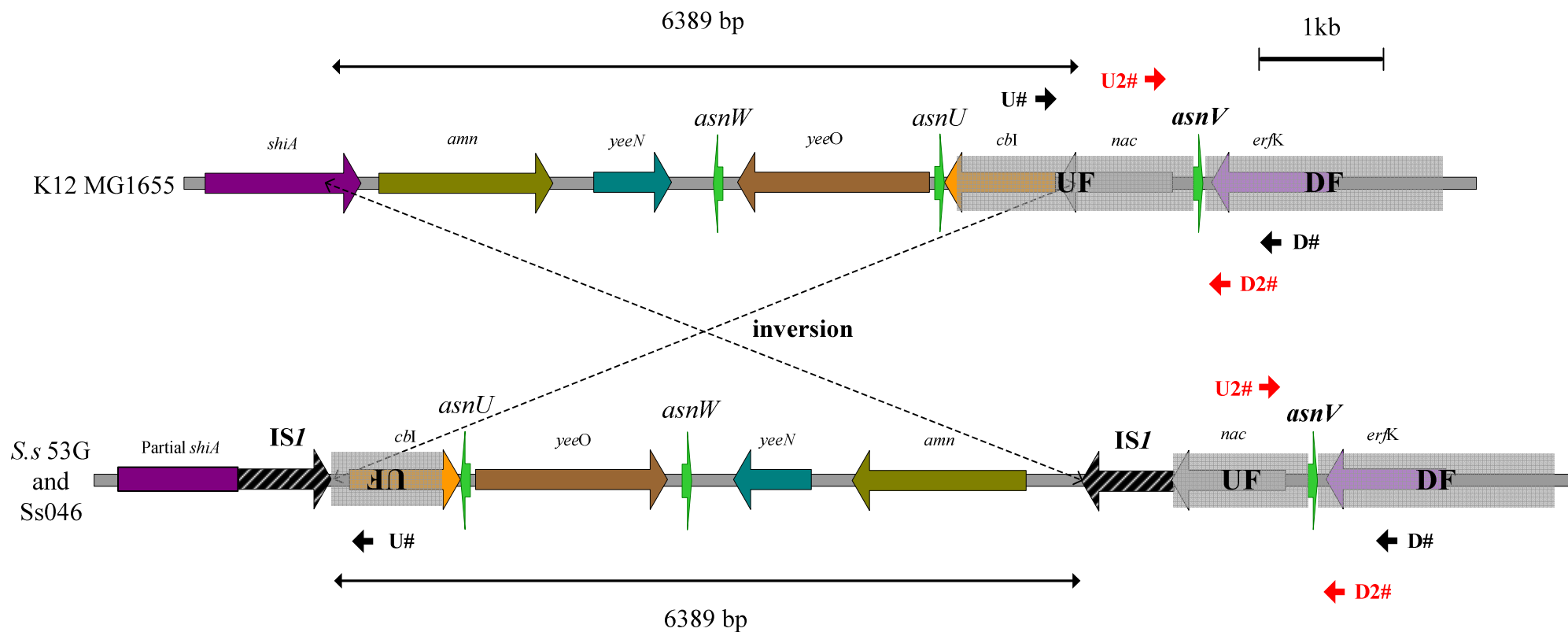


Figure A2. 26. Schematic showing the *asnV* UF inversion in *S. sonnei* 53G (unfinished) and *S. sonnei* Ss046 relative to *E. coli* K12 MG1655.

The locations of the original U/D primer pair and the second U/D primer pair are indicated by the small black arrows and small red arrows respectively.

At the time of this finding, I had performed the *asnV* tRIP screen across fourteen of the *Shigella* strains, and five were tRIP negative (see Table A2. 17).

Table A2. 17 The initial *asnV* tRIP results showing the fourteen *Shigella* strains screened with the original U and D primers.

Strain	Code	<i>asnV</i> U#-D#
<i>E. coli</i> K12 MG1655		+ ^a
<i>E. coli</i> CFT073		-
<i>S. dysenteriae</i> 3	S101	+
<i>S. dysenteriae</i> 9	S102	+
<i>S. dysenteriae</i> 6	S103	+
<i>S. flexneri</i> 1a	S104	+
<i>S. flexneri</i> 1b	S105	+
<i>S. flexneri</i> 2a	S106	+
<i>S. flexneri</i> 2b	S107	+
<i>S. flexneri</i> 3a ^d	S108	- ^b
<i>S. flexneri</i> 4a ^e	S109	+
<i>S. flexneri</i> 6	S110	N/A ^c
<i>S. flexneri</i> X	S111	N/A
<i>S. flexneri</i> Y	S112	N/A
<i>S. sonnei</i>	S113	-
<i>S. sonnei</i> bio a	S114	-
<i>S. sonnei</i> bio g	S115	N/A
<i>S. boydii</i> 1	S116	-
<i>S. boydii</i> 2	S117	-
<i>S. boydii</i> 3	S118	+
<i>S. boydii</i> 4	S119	N/A
<i>S. boydii</i> 7	S120	N/A

^a Positive tRIP result, with a tRIP amplicon no greater than 6.0 kb at this strain-tRNA locus, see Table 3.3 for full details of tRIP amplicon lengths

^b Negative tRIP result, indicating the presence of a putative GI at this strain-tRNA locus

^c Indicates that tRIP PCR was not performed at this strain-tRNA locus with this primer pair

^d After characterisation with SGSP-PCR and an API-20E test strip S108 was re-classified as a *S. sonnei* (see section 5.7 for more details)

^e After characterisation with SGSP-PCR and an API-20E test strip, S109 was re-classified as an atypical *E. coli* (see section 5.6 for more details)

The initial tRIP results indicated that three *S. sonnei* strains harboured a GI at the *asnV* locus, as did two *S. boydii* strains. It was therefore vital to check that these were not false negatives due to inversion or deletion of part of the *asnV* UF. I therefore designed a second U primer that was downstream of the original U primer, closer to the *asnV* tRNA gene and within the UF region in *S. sonnei* 53G that was not inverted (see Figure A2. 26), this was designated *asnV* U 2. I also designed a second D primer (*asnV* D 2) that was closer to the start of the DF and would therefore be more useful for SGSP-PCR amplicon sequence analysis. The new primers were then used in a tRIP screen across the original *asnV* tRIP-negative *Shigella* strains and the remaining strains that were to be screened (see Table A2. 18).

Table A2. 18 The final *asnV* tRIP screen results after screening with the U2 and D2 primers.

Strain	Code	<i>asnV</i> U#-D#	<i>asnV</i> U2#-D2#
<i>E. coli</i> K12 MG1655		+ ^a	+
<i>E. coli</i> CFT073		-	-
<i>S. dysenteriae</i> 3	S101	+	N/A
<i>S. dysenteriae</i> 9	S102	+	N/A
<i>S. dysenteriae</i> 6	S103	+	N/A
<i>S. flexneri</i> 1a	S104	+	N/A
<i>S. flexneri</i> 1b	S105	+	N/A
<i>S. flexneri</i> 2a	S106	+	N/A
<i>S. flexneri</i> 2b	S107	+	N/A
<i>S. flexneri</i> 3a ^d	S108	- ^b	+
<i>S. flexneri</i> 4a ^e	S109	+	N/A
<i>S. flexneri</i> 6	S110	N/A ^c	+
<i>S. flexneri</i> X	S111	N/A	+
<i>S. flexneri</i> Y	S112	N/A	+
<i>S. sonnei</i>	S113	-	+
<i>S. sonnei</i> bio a	S114	-	+
<i>S. sonnei</i> bio g	S115	N/A	+
<i>S. boydii</i> 1	S116	-	+
<i>S. boydii</i> 2	S117	-	+
<i>S. boydii</i> 3	S118	+	N/A
<i>S. boydii</i> 4	S119	N/A	+
<i>S. boydii</i> 7	S120	N/A	-

^{a,b,c,d,e} See Table A2. 17 footnotes for details

These results confirmed that the five tRIP-negative *Shigella* strains flagged by the original U-D *asnV* tRIP screen were false-negative results and they in fact harboured no island DNA between the *asnV* U and D flanking regions. The only *Shigella* strain that produced no tRIP amplicon with the second primer pair was the *S. boydii* 7 strain, however, SGSP-PCR from both the U and D primer did not produce any amplicons, so the putative GI at this site remains uncharacterised.

References

- ABAJY, M. Y., KOPEC, J., SCHIWON, K., BURZYNSKI, M., DORING, M., BOHN, C. & GROHMANN, E. (2007) A type IV-secretion-like system is required for conjugative DNA transport of broad-host-range plasmid pIP501 in gram-positive bacteria. *J Bacteriol*, 189, 2487-96.
- ADHIKARI, P., ALLISON, G., WHITTLE, B. & VERMA, N. K. (1999) Serotype 1a O-antigen modification: molecular characterization of the genes involved and their novel organization in the *Shigella flexneri* chromosome. *J Bacteriol*, 181, 4711-8.
- AHMED, Z. U., SARKER, M. R. & SACK, D. A. (1988) Nutritional requirements of shigellae for growth in a minimal medium. *Infect Immun*, 56, 1007-9.
- AL-HASANI, K., ADLER, B., RAJAKUMAR, K. & SAKELLARIS, H. (2001a) Distribution and structural variation of the she pathogenicity island in enteric bacterial pathogens. *J Med Microbiol*, 50, 780-6.
- AL-HASANI, K., HENDERSON, I. R., SAKELLARIS, H., RAJAKUMAR, K., GRANT, T., NATARO, J. P., ROBINS-BROWNE, R. & ADLER, B. (2000) The sigA gene which is borne on the she pathogenicity island of *Shigella flexneri* 2a encodes an exported cytopathic protease involved in intestinal fluid accumulation. *Infect Immun*, 68, 2457-63.
- AL-HASANI, K., RAJAKUMAR, K., BULACH, D., ROBINS-BROWNE, R., ADLER, B. & SAKELLARIS, H. (2001b) Genetic organization of the she pathogenicity island in *Shigella flexneri* 2a. *Microb Pathog*, 30, 1-8.
- ALLISON, G. E., ANGELES, D., TRAN-DINH, N. & VERMA, N. K. (2002) Complete genomic sequence of SfV, a serotype-converting temperate bacteriophage of *Shigella flexneri*. *J Bacteriol*, 184, 1974-87.
- ALLISON, G. E. & VERMA, N. K. (2000) Serotype-converting bacteriophages and O-antigen modification in *Shigella flexneri*. *Trends Microbiol*, 8, 17-23.

ALTSCHUL, S. F., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25, 3389-402.

ASHIDA, H., TOYOTOME, T., NAGAI, T. & SASAKAWA, C. (2007) Shigella chromosomal IpaH proteins are secreted via the type III secretion system and act as effectors. *Mol Microbiol*, 63, 680-93.

AUSUBEL, F.M., BRENT,R., KINGSTON,R.E., MOORE,D.D., SEIDMAN,J.G., SMITH,J.A. & STRUHL,K. (eds) (1987) *Current Protocols in Molecular Biology*, Greene Publishing Associates and Wiley-Interscience, New York, NY.

BEABER, J. W., HOCHHUT, B. & WALDOR, M. K. (2002) Genomic and functional analyses of SXT, an integrating antibiotic resistance gene transfer element derived from *Vibrio cholerae*. *J Bacteriol*, 184, 4259-69.

BERNIER-FEBREAU, C., DU MERLE, L., TURLIN, E., LABAS, V., ORDONEZ, J., GILLES, A. M. & LE BOUGUENEC, C. (2004) Use of deoxyribose by intestinal and extraintestinal pathogenic *Escherichia coli* strains: a metabolic adaptation involved in competitiveness. *Infect Immun*, 72, 6151-6.

BISHOP, A. L., BAKER, S., JENKS, S., FOOKES, M., GAORA, P. O., PICKARD, D., ANJUM, M., FARRAR, J., HIEN, T. T., IVENS, A. & DOUGAN, G. (2005) Analysis of the hypervariable region of the *Salmonella enterica* genome associated with tRNA(LeuX). *J Bacteriol*, 187, 2469-82.

BJOURSON, A. J. & COOPER, J. E. (1992) Band-stab PCR: a simple technique for the purification of individual PCR products. *Nucleic Acids Res*, 20, 4675.

BLATTNER, F. R., PLUNKETT, G., 3RD, BLOCH, C. A., PERNA, N. T., BURLAND, V., RILEY, M., COLLADO-VIDES, J., GLASNER, J. D., RODE, C. K., MAYHEW, G. F., GREGOR, J., DAVIS, N. W., KIRKPATRICK, H. A., GOEDEN, M. A., ROSE, D. J., MAU,

- B. & SHAO, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, 277, 1453-74.
- BLOCKER, A., JOUIHRI, N., LARQUET, E., GOUNON, P., EBEL, F., PARSOT, C., SANSONETTI, P. & ALLAOUI, A. (2001) Structure and composition of the *Shigella flexneri* "needle complex", a part of its type III secretion. *Mol Microbiol*, 39, 652-63.
- BLOMFIELD, I. C., VAUGHN, V., REST, R. F. & EISENSTEIN, B. I. (1991) Allelic exchange in *Escherichia coli* using the *Bacillus subtilis* *sacB* gene and a temperature-sensitive pSC101 replicon. *Mol Microbiol*, 5, 1447-57.
- BLUM, G., OTT, M., LISCHIEWSKI, A., RITTER, A., IMRICH, H., TSCHAPE, H. & HACKER, J. (1994) Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an *Escherichia coli* wild-type pathogen. *Infect Immun*, 62, 606-14.
- BOYD, E. F., LI, J., OCHMAN, H. & SELANDER, R. K. (1997) Comparative genetics of the *inv-spa* invasion gene complex of *Salmonella enterica*. *J Bacteriol*, 179, 1985-91.
- BUCHRIESER, C., BROSC, R., BACH, S., GUIYOULE, A. & CARNIEL, E. (1998) The high-pathogenicity island of *Yersinia pseudotuberculosis* can be inserted into any of the three chromosomal *asn* tRNA genes. *Mol Microbiol*, 30, 965-78.
- BURRUS, V., PAVLOVIC, G., DECARIS, B. & GUEDON, G. (2002) Conjugative transposons: the tip of the iceberg. *Mol Microbiol*, 46, 601-10.
- CAMPBELL, A. (2003) Prophage insertion sites. *Res Microbiol*, 154, 277-82.
- CASJENS, S., WINN-STAPLEY, D. A., GILCREASE, E. B., MORONA, R., KUHLEWEIN, C., CHUA, J. E., MANNING, P. A., INWOOD, W. & CLARK, A. J. (2004) The chromosome of *Shigella flexneri* bacteriophage Sf6: complete nucleotide sequence, genetic mosaicism, and DNA packaging. *J Mol Biol*, 339, 379-94.

CHAUDHURI, R. R., KHAN, A. M. & PALLAN, M. J. (2004) coliBASE: an online database for Escherichia coli, Shigella and Salmonella comparative genomics. *Nucleic Acids Res*, 32, D296-9.

CHEN, Z. & SCHNEIDER, T. D. (2006) Comparative analysis of tandem T7-like promoter containing regions in enterobacterial genomes reveals a novel group of genetic islands. *Nucleic Acids Res*, 34, 1133-47.

CHENNA, R., SUGAWARA, H., KOIKE, T., LOPEZ, R., GIBSON, T. J., HIGGINS, D. G. & THOMPSON, J. D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*, 31, 3497-500.

CLARK, C. A., BELTRAME, J. & MANNING, P. A. (1991) The oac gene encoding a lipopolysaccharide O-antigen acetylase maps adjacent to the integrase-encoding gene on the genome of Shigella flexneri bacteriophage Sf6. *Gene*, 107, 43-52.

CLERC, P. & SANSONETTI, P. J. (1987) Entry of Shigella flexneri into HeLa cells: evidence for directed phagocytosis involving actin polymerization and myosin accumulation. *Infect Immun*, 55, 2681-8.

COURT, D. L., OPPENHEIM, A. B. & ADHYA, S. L. (2007) A new look at bacteriophage lambda genetic networks. *J Bacteriol*, 189, 298-304.

DAUBIN, V. & OCHMAN, H. (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. *Genome Res*, 14, 1036-42.

DON, R. H., COX, P. T., WAINWRIGHT, B. J., BAKER, K. & MATTICK, J. S. (1991) 'Touchdown' PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res*, 19, 4008.

DUBNAU, D. (1999) DNA uptake in bacteria. *Annu Rev Microbiol*, 53, 217-44.

DUPONT, H. L., LEVINE, M. M., HORNICK, R. B. & FORMAL, S. B. (1989) Inoculum size in shigellosis and implications for expected mode of transmission. *J Infect Dis*, 159, 1126-8.

- EWING, B. & GREEN, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, 8, 186-94.
- EWING, W. H. (1949) Shigella Nomenclature. *J Bacteriol*, 57, 633-8.
- FENG, L., SENCHENKOVA, S. N., WANG, W., SHASHKOV, A. S., LIU, B., SHEVELEV, S. D., LIU, D., KNIREL, Y. A. & WANG, L. (2005) Structural and genetic characterization of the Shigella boydii type 18 O antigen. *Gene*, 355, 79-86.
- FOLKESSON, A., LOFDAHL, S. & NORMARK, S. (2002) The Salmonella enterica subspecies I specific centisome 7 genomic island encodes novel protein families present in bacteria living in close contact with eukaryotic cells. *Res Microbiol*, 153, 537-45.
- FOUTS, D. E. (2006) Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res*, 34, 5839-51.
- FRANCETIC, O., LORY, S. & PUGSLEY, A. P. (1998) A second prepilin peptidase gene in Escherichia coli K-12. *Mol Microbiol*, 27, 763-75.
- FUKIYA, S., MIZOGUCHI, H., TOBE, T. & MORI, H. (2004) Extensive genomic diversity in pathogenic Escherichia coli and Shigella Strains revealed by comparative genomic hybridization microarray. *J Bacteriol*, 186, 3911-21.
- GALAN, J. E. (1996) Molecular genetic bases of Salmonella entry into host cells. *Mol Microbiol*, 20, 263-71.
- GARMENDIA, J., FRANKEL, G. & CREPIN, V. F. (2005) Enteropathogenic and enterohemorrhagic Escherichia coli infections: translocation, translocation, translocation. *Infect Immun*, 73, 2573-85.
- GARRIDO, P., BLANCO, M., MORENO-PAZ, M., BRIONES, C., DAHBI, G., BLANCO, J., BLANCO, J. & PARRO, V. (2006) STEC-EPEC oligonucleotide microarray: a new tool for typing genetic variants of the LEE pathogenicity island of human and animal Shiga toxin-producing Escherichia coli (STEC) and enteropathogenic E. coli (EPEC) strains. *Clin Chem*, 52, 192-201.

GERDES, S. Y., SCHOLLE, M. D., CAMPBELL, J. W., BALAZSI, G., RAVASZ, E., DAUGHERTY, M. D., SOMERA, A. L., KYRPIDES, N. C., ANDERSON, I., GELFAND, M. S., BHATTACHARYA, A., KAPATRAL, V., D'SOUZA, M., BAEV, M. V., GRECHKIN, Y., MSEEH, F., FONSTEIN, M. Y., OVERBEEK, R., BARABASI, A. L., OLTVAI, Z. N. & OSTERMAN, A. L. (2003) Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *J Bacteriol*, 185, 5673-84.

GRIGORIEV, A. (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res*, 26, 2286-90.

GROISMAN, E. A. & OCHMAN, H. (1996) Pathogenicity islands: bacterial evolution in quantum leaps. *Cell*, 87, 791-4.

GUAN, S., BASTIN, D. A. & VERMA, N. K. (1999) Functional analysis of the O antigen glucosylation gene cluster of Shigella flexneri bacteriophage SfX. *Microbiology*, 145 (Pt 5), 1263-73.

HACKER, J., BENDER, L., OTT, M., WINGENDER, J., LUND, B., MARRE, R. & GOEBEL, W. (1990) Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal Escherichia coli isolates. *Microb Pathog*, 8, 213-25.

HACKER, J., BLUM-OEHLER, G., MUHLDOERFER, I. & TSCHAPE, H. (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol Microbiol*, 23, 1089-97.

HACKER, J. & KAPER, J. B. (2000) Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol*, 54, 641-79.

HALE, T. L. (1991) Genetic basis of virulence in Shigella species. *Microbiol Rev*, 55, 206-24.

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. & SHINAGAWA, H. (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res*, 8, 11-22.

HENDERSON, I. R., CZECZULIN, J., ESLAVA, C., NORIEGA, F. & NATARO, J. P. (1999) Characterization of pic, a secreted protease of *Shigella flexneri* and enteroaggregative *Escherichia coli*. *Infect Immun*, 67, 5587-96.

HOCHHUT, B., JAHREIS, K., LENGELER, J. W. & SCHMID, K. (1997) CTnscr94, a conjugative transposon found in enterobacteria. *J Bacteriol*, 179, 2097-102.

HORTON, R. M., HUNT, H. D., HO, S. N., PULLEN, J. K. & PEASE, L. R. (1989) Engineering hybrid genes without the use of restriction enzymes: gene splicing by overlap extension. *Gene*, 77, 61-8.

HOU, Y. M. (1999) Transfer RNAs and pathogenicity islands. *Trends Biochem Sci*, 24, 295-8.

HSIAO, W., WAN, I., JONES, S. J. & BRINKMAN, F. S. (2003) IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics*, 19, 418-20.

INGERSOLL, M., GROISMAN, E. A. & ZYCHLINSKY, A. (2002) Pathogenicity islands of *Shigella*. *Curr Top Microbiol Immunol*, 264, 49-65.

INGERSOLL, M. A., MOSS, J. E., WEINRAUCH, Y., FISHER, P. E., GROISMAN, E. A. & ZYCHLINSKY, A. (2003) The ShiA protein encoded by the *Shigella flexneri* SHI-2 pathogenicity island attenuates inflammation. *Cell Microbiol*, 5, 797-807.

ITO, H., KIDO, N., ARAKAWA, Y., OHTA, M., SUGIYAMA, T. & KATO, N. (1991) Possible mechanisms underlying the slow lactose fermentation phenotype in *Shigella* spp. *Appl Environ Microbiol*, 57, 2912-7.

JAHREIS, K., BENTLER, L., BOCKMANN, J., HANS, S., MEYER, A., SIEPELMEYER, J. & LENGELER, J. W. (2002) Adaptation of sucrose metabolism in the *Escherichia coli* wild-type strain EC3132. *J Bacteriol*, 184, 5307-16.

JIN, Q., YUAN, Z., XU, J., WANG, Y., SHEN, Y., LU, W., WANG, J., LIU, H., YANG, J., YANG, F., ZHANG, X., ZHANG, J., YANG, G., WU, H., QU, D., DONG, J., SUN, L., XUE, Y., ZHAO, A., GAO, Y., ZHU, J., KAN, B., DING, K., CHEN, S., CHENG, H., YAO, Z., HE, B., CHEN, R., MA, D., QIANG, B., WEN, Y., HOU, Y. & YU, J. (2002) Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res*, 30, 4432-41.

KARAOLIS, D. K., LAN, R. & REEVES, P. R. (1994) Sequence variation in *Shigella sonnei* (Sonnei), a pathogenic clone of *Escherichia coli*, over four continents and 41 years. *J Clin Microbiol*, 32, 796-802.

KARLIN, S. (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol*, 9, 335-43.

KARLIN, S., CAMPBELL, A. M. & MRAZEK, J. (1998a) Comparative DNA analysis across diverse genomes. *Annu Rev Genet*, 32, 185-225.

KARLIN, S., MRAZEK, J. & CAMPBELL, A. M. (1998b) Codon usages in different gene classes of the *Escherichia coli* genome. *Mol Microbiol*, 29, 1341-55.

KNAPP, S., HACKER, J., JARCHAU, T. & GOEBEL, W. (1986) Large, unstable inserts in the chromosome affect virulence properties of uropathogenic *Escherichia coli* O6 strain 536. *J Bacteriol*, 168, 22-30.

KOBAYASHI, I. (2001) Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res*, 29, 3742-56.

KOSKI, L. B., MORTON, R. A. & GOLDING, G. B. (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol*, 18, 404-12.

KOTLOFF, K. L., WINICKOFF, J. P., IVANOFF, B., CLEMENS, J. D., SWERDLOW, D. L., SANSONETTI, P. J., ADAK, G. K. & LEVINE, M. M. (1999) Global burden of Shigella infections: implications for vaccine development and implementation of control strategies. *Bull World Health Organ*, 77, 651-66.

LAI, V., WANG, L. & REEVES, P. R. (1998) Escherichia coli clone Sonnei (Shigella sonnei) had a chromosomal O-antigen gene cluster prior to gaining its current plasmid-borne O-antigen genes. *J Bacteriol*, 180, 2983-6.

LAN, R., ALLES, M. C., DONOHOE, K., MARTINEZ, M. B. & REEVES, P. R. (2004) Molecular evolutionary relationships of enteroinvasive Escherichia coli and Shigella spp. *Infect Immun*, 72, 5080-8.

LAN, R. & REEVES, P. R. (2002) Escherichia coli in disguise: molecular origins of Shigella. *Microbes Infect*, 4, 1125-32.

LAWRENCE, J. G. & OCHMAN, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol*, 44, 383-97.

LAWRENCE, J. G. & OCHMAN, H. (1998) Molecular archaeology of the Escherichia coli genome. *Proc Natl Acad Sci U S A*, 95, 9413-7.

LEDERBERG, J., LEDERBERG, E. M., ZINDER, N. D. & LIVELY, E. R. (1951) Recombination analysis of bacterial heredity. *Cold Spring Harb Symp Quant Biol*, 16, 413-43.

LIN, R. J., CAPAGE, M. & HILL, C. W. (1984) A repetitive DNA sequence, rhs, responsible for duplications within the Escherichia coli K-12 chromosome. *J Mol Biol*, 177, 1-18.

LINDBERG, A. A., KARNELL, A. & WEINTRAUB, A. (1991) The lipopolysaccharide of Shigella bacteria as a virulence factor. *Rev Infect Dis*, 13 Suppl 4, S279-84.

LIO, P. & VANNUCCI, M. (2000) Finding pathogenicity islands and gene transfer events in genome data. *Bioinformatics*, 16, 932-40.

LIVNY, J. & FRIEDMAN, D. I. (2004) Characterizing spontaneous induction of Stx encoding phages using a selectable reporter system. *Mol Microbiol*, 51, 1691-704.

LUCCHINI, S., THOMPSON, A. & HINTON, J. C. (2001) Microarrays for microbiologists. *Microbiology*, 147, 1403-14.

LUCK, S. N., TURNER, S. A., RAJAKUMAR, K., ADLER, B. & SAKELLARIS, H. (2004) Excision of the Shigella resistance locus pathogenicity island in Shigella flexneri is stimulated by a member of a new subgroup of recombination directionality factors. *J Bacteriol*, 186, 5551-4.

LUCK, S. N., TURNER, S. A., RAJAKUMAR, K., SAKELLARIS, H. & ADLER, B. (2001) Ferric dicitrate transport system (Fec) of Shigella flexneri 2a YSH6000 is encoded on a novel pathogenicity island carrying multiple antibiotic resistance genes. *Infect Immun*, 69, 6012-21.

LURIA, S. E. & DELBRÜCK, M. (1943) Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics*, 28, 491-511.

MALLOFF, C., DULLAGHAN, E., LI, A., STOKES, R., FERNANDEZ, R. & LAM, W. (2003) Two-dimensional DNA displays for comparisons of bacterial genomes. *Biol Proced Online*, 5, 143-152.

MALLOFF, C. A., FERNANDEZ, R. C., DULLAGHAN, E. M., STOKES, R. W. & LAM, W. L. (2002) Two-dimensional display and whole genome comparison of bacterial pathogen genomes of high G+C DNA content. *Gene*, 293, 205-11.

MALLOFF, C. A., FERNANDEZ, R. C. & LAM, W. L. (2001) Bacterial comparative genomic hybridization: a method for directly identifying lateral gene transfer. *J Mol Biol*, 312, 1-5.

MANOIL, C. & BECKWITH, J. (1985) Tnp_{phoA}: a transposon probe for protein export signals. *Proc Natl Acad Sci U S A*, 82, 8129-33.

MANTRI, Y. & WILLIAMS, K. P. (2004) Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res*, 32, D55-8.

MAURELLI, A. T. (1989) Regulation of virulence genes in *Shigella*. *Mol Biol Med*, 6, 425-32.

MAURELLI, A. T. (2007) Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens. *FEMS Microbiol Lett*, 267, 1-8.

MAURELLI, A. T. & SANSONETTI, P. J. (1988) Genetic determinants of *Shigella* pathogenicity. *Annu Rev Microbiol*, 42, 127-50.

MAVRIS, M., MANNING, P. A. & MORONA, R. (1997) Mechanism of bacteriophage SfII-mediated serotype conversion in *Shigella flexneri*. *Mol Microbiol*, 26, 939-50.

MEDINI, D., DONATI, C., TETTELIN, H., MASIGNANI, V. & RAPPUOLI, R. (2005) The microbial pan-genome. *Curr Opin Genet Dev*, 15, 589-94.

MEHTA, P., CASJENS, S. & KRISHNASWAMY, S. (2004) Analysis of the lambdoid prophage element ϕ 14 in the *E. coli* K-12 genome. *BMC Microbiol*, 4, 4.

MEHRABIAN, S. & TOHIDPOUR, M. (2005) Colicin type, biochemical type and drug resistance pattern of 154 strains of *Shigella sonnei* isolated in Iran. *Pak J Med Sci*, 21, 340-44.

MILLER, J. F. (2003) Bacteriophage and the evolution of epidemic cholera. *Infect Immun*, 71, 2981-2.

MILLER, V. L. & 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*. *J Bacteriol*, 170, 2575-83.

MORABITO, S., TOZZOLI, R., OSWALD, E. & CAPRIOLI, A. (2003) A mosaic pathogenicity island made up of the locus of enterocyte effacement and a pathogenicity island of *Escherichia coli* O157:H7 is frequently present in attaching and effacing *E. coli*. *Infect Immun*, 71, 3343-8.

- MORELLE, G. (1989) A plasmid extraction procedure on a miniprep scale. *Focus*, 11, 7-8.
- MORITZ, R. L. & WELCH, R. A. (2006) The *Escherichia coli* *argW*-*dsdCXA* genetic island is highly variable, and *E. coli* K1 strains commonly possess two copies of *dsdCXA*. *J Clin Microbiol*, 44, 4038-48.
- MOSS, J. E., CARDOZO, T. J., ZYCHLINSKY, A. & GROISMAN, E. A. (1999) The *selC*-associated SHI-2 pathogenicity island of *Shigella flexneri*. *Mol Microbiol*, 33, 74-83.
- MUNIESA, M., SCHEMBRI, M. A., HAUF, N. & CHAKRABORTY, T. (2006) Active Genetic Elements Present in the Locus of Enterocyte Effacement in *Escherichia coli* O26 and Their Role in Mobility. *Infect Immun*, 74, 4190-9.
- MURPHY, K. C. & CAMPELLONE, K. G. (2003) Lambda Red-mediated recombinogenic engineering of enterohemorrhagic and enteropathogenic *E. coli*. *BMC Mol Biol*, 4, 11.
- NATARO, J. P., SERIWATANA, J., FASANO, A., MANEVAL, D. R., GUERS, L. D., NORIEGA, F., DUBOVSKY, F., LEVINE, M. M. & MORRIS, J. G., JR. (1995) Identification and cloning of a novel plasmid-encoded enterotoxin of enteroinvasive *Escherichia coli* and *Shigella* strains. *Infect Immun*, 63, 4721-8.
- NIYOGI, S. K. (2005) Shigellosis. *J Microbiol*, 43, 133-43.
- NORIEGA, F. R., LIAO, F. M., FORMAL, S. B., FASANO, A. & LEVINE, M. M. (1995) Prevalence of *Shigella* enterotoxin 1 among *Shigella* clinical isolates of diverse serotypes. *J Infect Dis*, 172, 1408-10.
- NUNES-DUBY, S. E., KWON, H. J., TIRUMALAI, R. S., ELLENBERGER, T. & LANDY, A. (1998) Similarities and differences among 105 members of the *Int* family of site-specific recombinases. *Nucleic Acids Res*, 26, 391-406.
- OU, H. Y., CHEN, L. L., LONNEN, J., CHAUDHURI, R. R., THANI, A. B., SMITH, R., GARTON, N. J., HINTON, J., PALLAN, M., BARER, M. R. & RAJAKUMAR, K. (2006) A

novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria. *Nucleic Acids Res*, 34, e3.

OU, H. Y., HE, X., HARRISON, E. M., KULASEKARA, B. R., THANI, A. B., KADIOGLU, A., LORY, S., HINTON, J. C., BARER, M. R., DENG, Z. & RAJAKUMAR, K. (2007) MobilomeFINDER: web-based tools for in silico and experimental discovery of bacterial genomic islands. *Nucleic Acids Res*, 35, W97-W104.

OU, H. Y., SMITH, R., LUCCHINI, S., HINTON, J., CHAUDHURI, R. R., PALLAN, M., BARER, M. R. & RAJAKUMAR, K. (2005) ArrayOme: a program for estimating the sizes of microarray-visualized bacterial genomes. *Nucleic Acids Res*, 33, e3.

PARSOT, C. & SANSONETTI, P. J. (1996) Invasion and the pathogenesis of *Shigella* infections. *Curr Top Microbiol Immunol*, 209, 25-42.

PERNA, N. T., PLUNKETT, G., 3RD, BURLAND, V., MAU, B., GLASNER, J. D., ROSE, D. J., MAYHEW, G. F., EVANS, P. S., GREGOR, J., KIRKPATRICK, H. A., POSFAI, G., HACKETT, J., KLINK, S., BOUTIN, A., SHAO, Y., MILLER, L., GROTEBECK, E. J., DAVIS, N. W., LIM, A., DIMALANTA, E. T., POTAMOUSIS, K. D., APODACA, J., ANANTHARAMAN, T. S., LIN, J., YEN, G., SCHWARTZ, D. C., WELCH, R. A. & BLATTNER, F. R. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, 409, 529-33.

PETROVSKAYA, V. G. & BONDARENKO, V. M. (1977) Recommended corrections to the classification of *Shigella flexneri* on a genetic basis. *Int. J. Syst Bacteriol*, 27, 171–175.

PETROVSKAYA, V. G. & KHOMENKO, N. A. (1979) Proposals for improving the classification of members of the genus *Shigella*. *Int. J. Syst. Bacteriol*, 29, 400–402.

PHILIPPE, N., ALCARAZ, J. P., COURSANGE, E., GEISELMANN, J. & SCHNEIDER, D. (2004) Improvement of pCVD442, a suicide plasmid for gene allele exchange in bacteria. *Plasmid*, 51, 246-55.

- POSFAI, G., KOOB, M. D., KIRKPATRICK, H. A. & BLATTNER, F. R. (1997) Versatile insertion plasmids for targeted genome manipulations in bacteria: isolation, deletion, and rescue of the pathogenicity island LEE of the *Escherichia coli* O157:H7 genome. *J Bacteriol*, 179, 4426-8.
- PRESTON, G. M., HAUBOLD, B. & RAINEY, P. B. (1998) Bacterial genomics and adaptation to life on plants: implications for the evolution of pathogenicity and symbiosis. *Curr Opin Microbiol*, 1, 589-97.
- PUGSLEY, A. P. (1993) The complete general secretory pathway in gram-negative bacteria. *Microbiol Rev*, 57, 50-108.
- PUPO, G. M., LAN, R. & REEVES, P. R. (2000) Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A*, 97, 10567-72.
- PURDY, G. E. & PAYNE, S. M. (2001) The SHI-3 iron transport island of *Shigella boydii* 0-1392 carries the genes for aerobactin synthesis and transport. *J Bacteriol*, 183, 4176-82.
- RAJAKUMAR, K., SASAKAWA, C. & 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. *Infect Immun*, 65, 4606-14.
- RAVATN, R., STUDER, S., SPRINGAEL, D., ZEHNDER, A. J. & VAN DER MEER, J. R. (1998) Chromosomal integration, tandem amplification, and deamplification in *Pseudomonas putida* F1 of a 105-kilobase genetic element containing the chlorocatechol degradative genes from *Pseudomonas* sp. Strain B13. *J Bacteriol*, 180, 4360-9.
- REITER, W. D., PALM, P. & YEATS, S. (1989) Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. *Nucleic Acids Res*, 17, 1907-14.
- REYRAT, J. M., PELICIC, V., GICQUEL, B. & RAPPUOLI, R. (1998) Counterselectable markers: untapped tools for bacterial genetics and pathogenesis. *Infect Immun*, 66, 4011-7.

RITTER, A., GALLY, D. L., OLSEN, P. B., DOBRINDT, U., FRIEDRICH, A., KLEMM, P. & HACKER, J. (1997) The Pai-associated leuX specific tRNA⁵(Leu) affects type 1 fimbriation in pathogenic *Escherichia coli* by control of FimB recombinase expression. *Mol Microbiol*, 25, 871-82.

ROUX, A., BELOIN, C. & GHIGO, J. M. (2005) Combined inactivation and expression strategy to study gene function under physiological conditions: application to identification of new *Escherichia coli* adhesins. *J Bacteriol*, 187, 1001-13.

SANSONETTI, P. J., HALE, T. L., DAMMIN, G. J., KAPFER, C., COLLINS, H. H., JR. & FORMAL, S. B. (1983) Alterations in the pathogenicity of *Escherichia coli* K-12 after transfer of plasmid and chromosomal genes from *Shigella flexneri*. *Infect Immun*, 39, 1392-402.

SASAKAWA, C., BUYSSE, J. M. & WATANABE, H. (1992) The large virulence plasmid of *Shigella*. *Curr Top Microbiol Immunol*, 180, 21-44.

SCHAAD, U. B. & WEDGWOOD, J. (1992) Lack of quinolone-induced arthropathy in children. *J Antimicrob Chemother*, 30, 414-6.

SCHICKLMAIER, P., MOSER, E., WIELAND, T., RABSCH, W. & SCHMIEGER, H. (1998) A comparative study on the frequency of prophages among natural isolates of *Salmonella* and *Escherichia coli* with emphasis on generalized transducers. *Antonie Van Leeuwenhoek*, 73, 49-54.

SCHMIDT, H. & HENSEL, M. (2004) Pathogenicity islands in bacterial pathogenesis. *Clin Microbiol Rev*, 17, 14-56.

SCHUBBE, S., KUBE, M., SCHEFFEL, A., WAWER, C., HEYEN, U., MEYERDIERKS, A., MADKOUR, M. H., MAYER, F., REINHARDT, R. & SCHULER, D. (2003) Characterization of a spontaneous nonmagnetic mutant of *Magnetospirillum gryphiswaldense* reveals a large deletion comprising a putative magnetosome island. *J Bacteriol*, 185, 5779-90.

SHIGA, K. (1898) Ueber den Dysenterie bacillus (*Bacillus dysenteriae*), *Zentralbl Bakteriol Parasitenkd Abt I Org.* 24, 817-824.

SHINAGAWA, H. & ITO, T. (1973) Inactivation of DNA-binding activity of repressor in extracts of lambda-lysogen treated with mitomycin C. *Mol Gen Genet*, 126, 103-10.

SPARLING, P. F. (1966) Genetic transformation of *Neisseria gonorrhoeae* to streptomycin resistance. *J Bacteriol*, 92, 1364-71.

SULLIVAN, J. T. & RONSON, C. W. (1998) Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. *Proc Natl Acad Sci U S A*, 95, 5145-9.

SUR, D., NIYOGI, S. K., SUR, S., DATTA, K. K., TAKEDA, Y., NAIR, G. B. & BHATTACHARYA, S. K. (2003) Multidrug-resistant *Shigella dysenteriae* type 1: forerunners of a new epidemic strain in eastern India? *Emerg Infect Dis*, 9, 404-5.

TALUKDER, K. A., DUTTA, D. K., SAFA, A., ANSARUZZAMAN, M., HASSAN, F., ALAM, K., ISLAM, K. M., CARLIN, N. I., NAIR, G. B. & SACK, D. A. (2001) Altering trends in the dominance of *Shigella flexneri* serotypes and emergence of serologically atypical *S. flexneri* strains in Dhaka, Bangladesh. *J Clin Microbiol*, 39, 3757-9.

TALUKDER, K. A., ISLAM, M. A., KHAJANCHI, B. K., DUTTA, D. K., ISLAM, Z., SAFA, A., ALAM, K., HOSSAIN, A., NAIR, G. B. & SACK, D. A. (2003) Temporal shifts in the dominance of serotypes of *Shigella dysenteriae* from 1999 to 2002 in Dhaka, Bangladesh. *J Clin Microbiol*, 41, 5053-8.

TAUSCHEK, M., GORRELL, R. J., STRUGNELL, R. A. & ROBINS-BROWNE, R. M. (2002) Identification of a protein secretory pathway for the secretion of heat-labile enterotoxin by an enterotoxigenic strain of *Escherichia coli*. *Proc Natl Acad Sci U S A*, 99, 7066-71.

- TAUXE, R. V., MCDONALD, R. C., HARGRETT-BEAN, N. & BLAKE, P. A. (1988) The persistence of *Shigella flexneri* in the United States: increasing role of adult males. *Am J Public Health*, 78, 1432-5.
- TAYLOR, D. E., ROOKER, M., KEELAN, M., NG, L. K., MARTIN, I., PERNA, N. T., BURLAND, N. T. & BLATTNER, F. R. (2002) Genomic variability of O islands encoding tellurite resistance in enterohemorrhagic *Escherichia coli* O157:H7 isolates. *J Bacteriol*, 184, 4690-8.
- TAYLOR, R. K., MANOIL, C. & MEKALANOS, J. J. (1989) Broad-host-range vectors for delivery of TnphoA: use in genetic analysis of secreted virulence determinants of *Vibrio cholerae*. *J Bacteriol*, 171, 1870-8.
- TOYOTOME, T., SUZUKI, T., KUWAE, A., NONAKA, T., FUKUDA, H., IMAJOH-OHMI, S., TOYOFUKU, T., HORI, M. & SASAKAWA, C. (2001) *Shigella* protein IpaH(9.8) is secreted from bacteria within mammalian cells and transported to the nucleus. *J Biol Chem*, 276, 32071-9.
- TURNER, S. A., LUCK, S. N., SAKELLARIS, H., RAJAKUMAR, K. & ADLER, B. (2001) Nested deletions of the SRL pathogenicity island of *Shigella flexneri* 2a. *J Bacteriol*, 183, 5535-43.
- TURNER, S. A., LUCK, S. N., SAKELLARIS, H., RAJAKUMAR, K. & ADLER, B. (2003) Molecular epidemiology of the SRL pathogenicity island. *Antimicrob Agents Chemother*, 47, 727-34.
- TURNER, S. A., LUCK, S. N., SAKELLARIS, H., RAJAKUMAR, K. & ADLER, B. (2004) Role of attP in integrase-mediated integration of the *Shigella* resistance locus pathogenicity island of *Shigella flexneri*. *Antimicrob Agents Chemother*, 48, 1028-31.
- ULLRICH, S., KUBE, M., SCHUBBE, S., REINHARDT, R. & SCHULER, D. (2005) A hypervariable 130-kilobase genomic region of *Magnetospirillum gryphiswaldense* comprises

a magnetosome island which undergoes frequent rearrangements during stationary growth. *J Bacteriol*, 187, 7176-84.

VAN DER PLOEG, J. R., WEISS, M. A., SALLER, E., NASHIMOTO, H., SAITO, N., KERTESZ, M. A. & LEISINGER, T. (1996) Identification of sulfate starvation-regulated genes in *Escherichia coli*: a gene cluster involved in the utilization of taurine as a sulfur source. *J Bacteriol*, 178, 5438-46.

VARGAS, M., GASCON, J., JIMENEZ DE ANTA, M. T. & VILA, J. (1999) Prevalence of *Shigella* enterotoxins 1 and 2 among *Shigella* strains isolated from patients with traveler's diarrhea. *J Clin Microbiol*, 37, 3608-11.

VERMA, N. K., BRANDT, J. M., VERMA, D. J. & LINDBERG, A. A. (1991) Molecular characterization of the O-acetyl transferase gene of converting bacteriophage SF6 that adds group antigen 6 to *Shigella flexneri*. *Mol Microbiol*, 5, 71-5.

VERMA, N. K., VERMA, D. J., HUAN, P. T. & LINDBERG, A. A. (1993) Cloning and sequencing of the glucosyl transferase-encoding gene from converting bacteriophage X (SFX) of *Shigella flexneri*. *Gene*, 129, 99-101.

VOKES, S. A., REEVES, S. A., TORRES, A. G. & PAYNE, S. M. (1999) The aerobactin iron transport system genes in *Shigella flexneri* are present within a pathogenicity island. *Mol Microbiol*, 33, 63-73.

WALDRON, D. E. & LINDSAY, J. A. (2006) Sau1: a novel lineage-specific type I restriction-modification system that blocks horizontal gene transfer into *Staphylococcus aureus* and between *S. aureus* isolates of different lineages. *J Bacteriol*, 188, 5578-85.

WANG, R. F. & KUSHNER, S. R. (1991) Construction of versatile low-copy-number vectors for cloning, sequencing and gene expression in *Escherichia coli*. *Gene*, 100, 195-9.

WANG, Y. D., ZHAO, S. & HILL, C. W. (1998) Rhs elements comprise three subfamilies which diverged prior to acquisition by *Escherichia coli*. *J Bacteriol*, 180, 4102-10.

WEI, J., GOLDBERG, M. B., BURLAND, V., VENKATESAN, M. M., DENG, W., FOURNIER, G., MAYHEW, G. F., PLUNKETT, G., 3RD, ROSE, D. J., DARLING, A., MAU, B., PERNA, N. T., PAYNE, S. M., RUNYEN-JANECKY, L. J., ZHOU, S., SCHWARTZ, D. C. & BLATTNER, F. R. (2003) Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. *Infect Immun*, 71, 2775-86.

WELCH, R. A., BURLAND, V., PLUNKETT, G., 3RD, REDFORD, P., ROESCH, P., RASKO, D., BUCKLES, E. L., LIOU, S. R., BOUTIN, A., HACKETT, J., STROUD, D., MAYHEW, G. F., ROSE, D. J., ZHOU, S., SCHWARTZ, D. C., PERNA, N. T., MOBLEY, H. L., DONNENBERG, M. S. & BLATTNER, F. R. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci U S A*, 99, 17020-4.

WILLIAMS, K. P. (2002) Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res*, 30, 866-75.

WINSTANLEY, C. (2002) Spot the difference: applications of subtractive hybridisation to the study of bacterial pathogens. *J Med Microbiol*, 51, 459-67.

YANG, F., YANG, J., ZHANG, X., CHEN, L., JIANG, Y., YAN, Y., TANG, X., WANG, J., XIONG, Z., DONG, J., XUE, Y., ZHU, Y., XU, X., SUN, L., CHEN, S., NIE, H., PENG, J., XU, J., WANG, Y., YUAN, Z., WEN, Y., YAO, Z., SHEN, Y., QIANG, B., HOU, Y., YU, J. & JIN, Q. (2005) Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res*, 33, 6445-58.

ZHANG, R. & ZHANG, C. T. (2004) A systematic method to identify genomic islands and its applications in analyzing the genomes of *Corynebacterium glutamicum* and *Vibrio vulnificus* CMCP6 chromosome I. *Bioinformatics*, 20, 612-22.

ZHAO, S., SANDT, C. H., FEULNER, G., VLAZNY, D. A., GRAY, J. A. & HILL, C. W.
(1993) Rhs elements of *Escherichia coli* K-12: complex composites of shared and unique components that have different evolutionary histories. *J Bacteriol*, 175, 2799-808.