

**ROLE OF EPIGENETIC MODIFICATION IN REGULATING
UROPATHOGENIC *ESCHERICHIA COLI* VIRULENCE**

KUROSH MEHERSHAHI

(M.Sc., University of Pune)

**A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY**

**DEPARTMENT OF MEDICINE
YONG LOO LIN SCHOOL OF MEDICINE
NATIONAL UNIVERSITY OF SINGAPORE**

2016

Supervisor:

Assistant Professor Dr. Swaine Lin Chen

Examiners:

Associate Professor Dr. Thomas Dick

Assistant Professor Dr. Chng Shu Sin

Associate Professor Dr. Scott Rice, Nanyang Technological University

DECLARATION

I hereby declare that this thesis is my original work
and it has been written by me in its entirety.

I have duly acknowledged all the sources of information
which have been used in the thesis.

This thesis has also not been submitted for any degree in any
university previously.

KUROSH MEHERSHAHI

29 September 2016

For my Mimi, Granny and Memas,

*The strongest and most loving women I have known, who always believed in
me more than I ever did and whom I miss dearly.*

ACKNOWLEDGEMENTS

I would like to first and foremost convey my deepest gratitude to my supervisor Dr. Swaine Chen. Looking back at how far I have come and how much I've learnt it feels incredible and this would not have been possible without his guidance. He has always been there patiently supporting all his students including me. His ability to prioritise science above all else and strive to excel at everything you apply yourself to are really inspiring. His emphasis on developing skills and reasoning rather than merely results gave me the chance to make mistakes and learn from them, gradually getting better. Although intimidated at first, I consider myself really lucky to have been his student. Thank you Swaine!

I would also like to thank the entire Chen lab, past and present. Here goes: Rashmi, Shazmina, Huibin, Jacquelin, Kristin, Lu Ting, Bala, Jade, Siyi, Pauline, En Jun, Stacey, Shanon, Milena, Rachel, Sathya, Sarah, Maurice, Marie, Penny, Adriano, Stacey... I hope I didn't miss anyone. Each one of these people have in their own special way added to the exceptionally fun and supportive environment in the lab.

Special thanks to my lab BFF Dr. Majid Eshaghi, whose constant support and encouragement has been very important in getting me here. Majid has been like a father to all of us, ever ready to help even at his busiest. Majid always enjoys a challenge and this came in very handy when experiments refused to cooperate. Nevertheless, if you needed advice, a sympathetic ear or a good laugh Majid always delivered. Merci Majid! Varnica is another person who has seen me at my best and my worst and I have had the pleasure of being there for her's. As a fellow grad student, we have seen it all together and her support has pulled me through some rough times, so thanks Ms. V!

I would also like to express my heartfelt gratitude to Dr. William Burkholder and Dr. Chng Shu Sin. Thank you for your patience and guidance as my TAC members. I really

appreciate the advice and encouragement I got from the two of them. I consider myself lucky for the friends I have made here in Singapore. We have been each other's support structure and this PhD would have been much more difficult without them. Thank you for making Singapore feel like home away from home.

Special thanks to Grishma Rane, who has always been there providing unwavering support throughout these four years and more. I could never have done this without her and consider myself immeasurably lucky to have her in my life. I would like to express my heartfelt gratitude to Dr. Shahin Irani, whose warmth, guidance and constant encouragement helped keep me on the right track and got me to where I am. Another person who has been instrumental in my scientific journey and whom I am forever indebted to is Dr. Atanu Basu. I might have given up on research had it not been for the inspiring nature of Basu Sir. His energy and enthusiasm for science, music, food, literature and everything else are incredible. When things got tough I would push myself knowing that I needed to make him proud. I and all his students miss him dearly and fondly remember sitting around the desk and discussing experiments every morning.

Finally, I want to thank my parents who are the most important and influential people in the world to me. My dad who taught me that being happy is a conscious decision we make every single day. His meticulous hard working nature and ridiculous sense of humour are my inspiration. My mom, hands down the most loving, kind and huggable person in the world. For always knowing what I was feeling without me having to say it and calming me when I needed it. I consider myself lucky to have two of the most loving, supporting and self-sacrificing people ever as my parents. My sister Kainaz, a perpetual irritant and yet loving inspiration, whom I miss all the time. And Rohan, the latest addition to the family and a spirited dose of fun. Thanks guys! I would also like to thank my grandparents and my uncle, for being there always and for their loving support. I hope to make you all proud.

TABLE OF CONTENTS

SUMMARY	5
LIST OF TABLES	7
LIST OF FIGURES	8
LIST OF ABBREVIATIONS	9
1. INTRODUCTION	12
1.1 <u>U</u> rinary <u>T</u> ract <u>I</u> nfections (UTIs)	12
1.2 <u>U</u> ropathogenic <u>E</u> <i>sch</i> erichia <u>c</u> oli (UPEC)	13
1.2.1 Pathogenesis of UPEC	13
1.2.2 UPEC virulence factors and genomics	15
1.2.3 Treatment of UPEC induced UTI	17
1.3 Epigenetic DNA Methylation	19
1.3.1 Types of DNA Methyltransferases	20
1.3.1.1 Restriction Modification Systems	21
1.3.1.2 Orphan Methyltransferases	25
1.3.2 Detection of DNA Methylation	26
1.4 Role of DNA Methylation in altering virulence	28
1.4.1 Orphan Methyltransferases	28
1.4.1.1 <u>D</u> NA <u>A</u> denine <u>M</u> ethyltransferase (<i>dam</i>)	29
1.4.1.2 <u>D</u> NA <u>C</u> ytosine <u>M</u> ethyltransferase (<i>dcm</i>)	32
1.4.1.3 <u>C</u> ell <u>C</u> ycle <u>R</u> egulated DNA <u>M</u> ethyltransferase (<i>ccrM</i>)	33
1.4.1.4 Other Orphan Methyltransferases	33
1.4.2 Restriction Modification System Methyltransferases	35

1.4.2.1 Phasevarions	36
1.4.2.2 Complex Phasevariable Type I Restriction Modification Systems	38
1.4.2.3 Classical Restriction Modification Systems	40
1.5 Role of DNA Methylation in Bacterial Evolution	41
1.6 Hypothesis and Aims	44
<hr/>	
2. MATERIALS AND METHODS	45
2.1 Media and Culture conditions	45
2.2 Strain Generation	45
2.3 Plasmid Generation	48
2.4 Restriction Modification Assay	48
2.5 Mouse infections	49
2.6 RNAseq	50
2.7 Whole Genome Sequencing (WGS)	51
2.8 Single Molecule Real Time (SMRT) sequencing	51
2.9 Quantitative RT-PCR	52
2.10 Motility Assay	53
2.11 Biofilm Assay	53
2.12 Growth curves	54
2.13 Phenotype Microarray	54
<hr/>	
3. RESULTS:	56
Elucidating the role of Type I restriction modification system mediated methylation in regulating Uropathogenic <i>E. coli</i> virulence and physiology	
3.1 Single Molecule Real Time (SMRT) sequencing identifies a novel methylated motif in UTI89	57

3.2 Characterization of Type I Restriction Modification Systems	58
3.2.1 Novel UTI89 Type I methylation is mediated by a functional <u>Restriction Modification System</u> (RMS)	58
3.2.2 Generating UTI89 strains bearing diverse <i>E. coli</i> Type I methylation motifs	63
3.3 Effect of Type I methylation on UTI89 virulence <i>in vitro</i>	65
3.3.1 Type I methylation does not affect UTI89 motility	65
3.3.2 Type I methylation does not affect biofilm formation in UTI89	66
3.3.3 Type I RMS mediated methylation does not affect UTI89 growth	68
3.4 Effect of Type I methylation on UTI89 virulence <i>in vivo</i>	71
3.4.1 Loss of native UTI89 methylation does not affect virulence <i>in vivo</i>	71
3.4.2 Switching Type I methylation in UTI89 does not affect <i>in vivo</i> virulence	73
3.5 Effect of Type I methylation on <i>Escherichia coli</i> gene expression	75
3.5.1 RNA-seq reveals no consequence of altered Type I DNA methylation on UTI89 transcriptome	75
3.5.2. RNA-seq reveals minimal Type I methylation mediated gene expression changes in K12 and CFT073	77
3.6 Phenotypic consequences of altered Type I DNA methylation in <i>E. coli</i>	80
3.6.1 High throughput phenotypic screen does not identify any differences owing to Type I DNA methylation	80
3.7 Summary	84
4. RESULTS:	85
Identification and Characterization of a novel Ribonucleotide reductase gene mutant	
4.1 UTI89 SWAP K12 strain KSM3-61-6 has a functional K12 Type I RMS	86
4.2 UTI89 SWAP K12 strain KSM3-61-6 is defective in kidney colonisation	87
4.3 RNA-seq reveals differentially expressed genes in strain KSM3-61-6	89

4.4 Phenotype Microarray identifies additional phenotypic differences for KSM3-61-6	93
4.5 Secondary mutation independent of methylation is responsible for KSM3-61-6 phenotypes	96
4.6 Summary	99
5. DISCUSSION	101
5.1 UTI89 encodes a novel Type I methylation motif mediated by a functional RMS	103
5.2 Altered UTI89 Type I methylation does not affect virulence	105
5.3 Type I methylation does not affect global gene expression patterns or phenotypes in multiple <i>E. coli</i> strains	106
5.4 CONCLUSIONS	109
REFERENCES	117
PUBLICATIONS	126

SUMMARY

Urinary tract infections (UTIs) represent one of the most common infections in humans. Prevalence is especially high in women, with 50% having at least one symptomatic episode in their lifetime and 25% of primary infections followed by a recurrent episode. Although UTIs are generally self-limiting and perceived to be therapeutically manageable with hospitalisations being rare, they still exert a significant burden on public healthcare systems. Moreover, due to frequent sometimes ineffectual antibiotic usage, the emergence of drug resistance is a serious concern. Uropathogenic *Escherichia coli* (UPEC) are the primary causative agents of UTIs accounting for about 80% of community acquired as well as nosocomial infections. Investigation of UPEC has revealed details about this pathogen's complex life cycle, which includes both intracellular and extracellular stages within diverse host niches. To grow, survive and efficiently colonise the harsh nutrient starved environment of the human urinary tract, these pathogens have acquired an arsenal of diverse virulence factors. Since UPEC do not encode a universal set of virulence factors, regulation of virulence factor expression mediated by DNA methylation could play an important role in pathogenesis.

DNA methylation refers to the covalent attachment of methyl groups to adenine or cytosine nucleotides in the context of specific sequences. This sequence specific methylation of DNA is mediated by methyltransferases, which are either solitary enzymes or part of restriction modification systems (RMSs). Solitary or orphan methyltransferases have been studied extensively in several bacterial species and are known to play a role in regulating DNA replication, mismatch repair, chromatin stability and gene expression. On the other hand, the primary function of RMS mediated methylation is host defence, by differentially tagging self and non-self DNA, so as to allow the selective degradation of non-self DNA. The precise molecular mechanisms via which DNA methylation regulates gene expression are known only for a few epigenetically regulated genomic loci. Also, any additional roles possessed by RMS associated methyltransferases by virtue of their ability to methylate DNA is poorly understood.

Our analysis of the methylome of cystitis causing UPEC isolate UTI89 identified methylation motifs corresponding with two previously known orphan methyltransferases and a novel bipartite methylation sequence 5'-CCA(N₇)CTTC-3' characteristic of a Type I RMS. We assigned this novel methylation to the Type I gene cluster *hsdSMR* and proceeded to investigate the role of this methylation in UTI89 biology besides host defence. The UTI89 Type I gene cluster was functionally validated, confirming that both the restriction and modification functions of this putative RMS are intact. To mimic the diversification of RMS repertoires, we replaced the native Type I methylation of UTI89 with that of lab strain K12 and pyelonephritis causing UPEC isolate CFT073. Swapping of either a single *hsdS* gene or the entire *hsdSMR* locus successfully resulted in completely changing all the Type I methylation marks, while retaining functional RMSs. To check if replacing or abolishing about 700 genomic methylation marks affects virulence, we utilised UPEC specific *in vitro* assays as well as the transurethral murine model of ascending UTI. Observing no differences at all *in vitro* or *in vivo*, we hypothesised that Type I RMS mediated methylation may have evolved not to interfere with bacterial physiology. To systematically test this hypothesis, we used RNA-seq to analyse gene expression patterns in different growth stages and phenotype microarray to screen a range of about 2000 diverse phenotypes, to shed light on any Type I methylation mediated effects. Confirming our hypothesis, UTI89 mutants with altered Type I methylation display no changes to gene expression patterns or phenotypic traits. To further strengthen our argument, we extended our broad high throughput transcriptomic and phenotypic screen to K12 and CFT073 Type I RMS deletion mutants. Similar to UTI89, in both these cases no global alterations were observed, with only exponential phase cultures showing a few genes (2-4) belonging to a single prophage being affected in K12. Overall, our study suggests that Type I RMS mediated methylation does not alter global *E. coli* gene expression. And in the case of UPEC isolate UTI89, this epigenetic mark convincingly does not affect virulence or other cellular processes.

LIST OF TABLES

Table 1: UTI89 methylated motifs	57
Table 2: HA titres for UTI89 strains	71
Table 3: Differentially expressed genes in UTI89 Type I methylation mutants	75-76
Table 4: Differentially expressed genes in CFT073 <i>ΔhdsMR</i>	78
Table 5: Differentially expressed genes in K12 <i>ΔhdsMR</i>	79-80
Table 6: Phenotype Microarray results for UTI89 and K12 mutants	82
Table 7: Stationary phase differentially expressed genes in KSM3-61-6	90
Table 8: Log phase differentially expressed genes in KSM3-61-6	90-91
Table 9: Phenotype Microarray results for KSM3-61-6	94-95
Table 10: Mutations present in KSM3-61-6	96-97
Table 11: List of Strains	112
Table 12: List of Plasmids	112
Table 13: List of Primers	113-116

LIST OF FIGURES

Figure 1: Characteristics of Type I-III Restriction Modification Systems	22-23
Figure 2: Generating seamless allelic replacements	47-48
Figure 3: Principle of the Restriction Modification Assay	59
Figure 4: Restriction Modification assay for Type I Restriction Modification Systems	61-62
Figure 5: Restriction Modification Assay for UTI89 SWAP strains	64
Figure 6: Motility of UTI89 mutants	66
Figure 7: Biofilm formation by UTI89 mutants	67-68
Figure 8: Growth curves for UTI89 methylation mutants	69
Figure 9: Growth curves for K12 and CFT073 methylation mutants	70
Figure 10: Deletion of the UTI89 Type I RMS has no effect on <i>in vivo</i> urinary tract infection	72-73
Figure 11: Co-infection with UTI89 SWAP K12 strain	74
Figure 12: Phenotype Microarray panels for UTI89 and K12 mutants	83-84
Figure 13: Restriction modification assay for strain KSM3-61-6	86-87
Figure 14: Growth curves for KSM3-61-6	87-88
Figure 15: Co-infection with KSM3-61-6	88-89
Figure 16: qRT-PCR for differentially expressed genes in KSM3-61-6	92
Figure 17: Motility of strain KSM3-61-6.	93
Figure 18: Phenotype Microarray panel for KSM3-61-6	94
Figure 19: Azathioprine sensitivity of UTI89 SWAP K12 mutants.	95-96
Figure 20: Motility of UTI89 <i>ybaL</i> and <i>nrdA</i> mutants	97-98
Figure 21: Azathioprine sensitivity of UTI89 <i>ybaL</i> and <i>nrdA</i> mutants	98

LIST OF ABBREVIATIONS

5mC	5-methylcytosine
4mC	N4-methylcytosine
6mA	N6-methyladenine
5hmC	5-hydroxymethylcytosine
Agn43	Antigen 43
APEC	Avian Pathogenic <i>Escherichia coli</i>
ATP	Adenosine triphosphate
bp	Base pairs
cAMP	Cyclic adenosine monophosphate
<i>ccrM</i>	Cell cycle regulated methyltransferase
cDNA	Complementary deoxyribonucleic acid
cfu	Colony forming units
CGI	CpG islands
CI	Competitive index
<i>cnfl</i>	Cytotoxic necrotising factor 1
<i>dam</i>	DNA adenine methyltransferase
<i>dcm</i>	DNA cytosine methyltransferase
DMSO	Dimethylsulfoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
dpi	Days post infection
EDTA	Ethylenediaminetetraacetic acid
EHEC	Enterohaemorrhagic <i>Escherichia coli</i>

EOT	Efficiency of transformation
EtBr	Ethidium Bromide
ExPEC	Extraintestinal pathogenic <i>Escherichia coli</i>
FDR	False discovery rate
GTP	Guanosine triphosphate
HA	Hemagglutination assay
h	Hour
<i>hsd</i>	Host specificity determinant
HUS	Hemolytic uremic syndrome
IBC	Intracellular bacterial communities
IPTG	Isopropyl β -D-1-thiogalactopyranoside
Kbp	Kilo base pairs
LB	Lysogeny broth
LPS	Lipopolysaccharides
Mbp	Mega base pairs
mg	Milligram
min	Minute
ml	Milliliter
mm	Millimetre
mM	Millimolar
mRNA	Messenger RNA
MRSA	Methicillin resistant <i>Staphylococcus aureus</i>
ng	Nanogram
OD	Optical density
ORF	Open reading frame
PBS	Phosphate buffered saline

PCR	Polymerase chain reaction
QIR	Quiescent intracellular reservoir
qRT-PCR	Quantitative real time PCR
QS	Quality score
RBC	Red blood cell
RMS	Restriction Modification System
rUTI	Recurrent urinary tract infection
rRNA	Ribosomal ribonucleic acid
RNA	Ribonucleic acid
RNR	Ribonucleotide reductase
<i>sat</i>	Secreted autotransporter protein
SAM	S-Adenosyl methionine
SMRT	Single Molecule Real Time
ST	Sequence type
TRD	Target recognition domain
tRNA	Transfer ribonucleic acid
μg	Microgram
μl	Microliter
μM	Micromolar
UPEC	Uropathogenic <i>Escherichia coli</i>
UTI	Urinary Tract Infection
<i>vat</i>	Vacuolating autotransporter protein
WGS	Whole genome sequencing
YESCA	Yeast extract Casamino acids

1. INTRODUCTION

1.1 Urinary Tract Infections (UTIs)

Urinary Tract Infections (UTIs) refer to infections of the urinary tract and encompass infections with diverse aetiologies, clinical manifestations, sites of infection and patient outcomes. UTIs are one of the most common infections in humans afflicting more than 150 million people annually worldwide [1]. UTIs can be clinically classified as either complicated or uncomplicated. Uncomplicated UTIs refer to infections in otherwise healthy individuals with no anatomical or immunological abnormalities. Factors which increase the risk of uncomplicated UTIs are female gender, a prior UTI, vaginal infection, sexual activity, diabetes, obesity and host genetics. Complicated UTIs on the other hand refer to infections associated with a compromised urinary tract; due to factors such as urinary obstruction, urinary retention, immune suppression, surgery of the urinary tract and presence of a foreign medical device such as an indwelling catheter [2]. UTIs can be further classified as lower and upper UTIs. Lower UTIs or cystitis refers to an infection of the bladder commonly associated with symptoms such as dysuria (painful urination), increased urinary frequency and urgency, suprapubic pain and in severe cases haematuria (blood in urine). Upper UTIs or pyelonephritis is an infection of the kidneys, which includes symptoms of cystitis as well as fever, flank pain, vomiting and nausea. In 30% of cases, bacteria might eventually spread from the kidneys to the bloodstream causing a life-threatening systemic infection [3].

About 50% of women and 12% of men will experience a symptomatic UTI in the course of their life time [4]. This strikingly high prevalence of UTIs especially in women is believed to be due to the shorter length of the female urethra and the close proximity of the vaginal as well as gastrointestinal flora in the peri urethral area. Both these anatomical features come together to facilitate colonization of the urinary tract in women. Furthermore, 25% of these women will subsequently experience a recurrent UTI (rUTI). A rUTI is defined as 2 episodes of culture

positive UTI in the last 6 months or 3 episodes in the last 12 months [5]. rUTI can either be due to recurrent ascending infections from an external source such as the gastrointestinal flora or chronic persistent infection of the bladder. There is evidence that strains of Uropathogenic *Escherichia coli* (UPEC) causing rUTI bear close resemblance to the colonic flora of the same individual [6] and in 77% of cases, UPEC causing rUTI are identical to the primary infection strain [7]. Although episodes of symptomatic UTI are usually self-limiting and amenable to antibiotic treatment, it does result in 6.1 symptomatic days and 2.4 days of restricted activity [4]. Taken together with the high rate of prevalence and recurrence, UTIs constitute a significant burden on global public healthcare systems and a serious cause of reduction of quality of life for women.

1.2 Uropathogenic *Escherichia coli* (UPEC)

Uropathogenic *Escherichia coli* (UPEC) is the primary causative agent for both uncomplicated (75%) and complicated (65%) UTIs. Although UTIs can be caused by a plethora of other bacterial and fungal pathogens, UPEC are by far the primary aetiological agents. Uncomplicated UTIs are also caused by other pathogens such as *Klebsiella pneumoniae* (6%), *Staphylococcus saprophyticus* (6%), *Enterococcus faecalis* (5%), group B *Streptococcus* (3%) and others. For complicated UTIs, common causes after UPEC include *Enterococcus faecalis* (11%), *Klebsiella pneumoniae* (8%), *Candida spp.* (7%), *Staphylococcus aureus* (3%) and others [1]. Owing to its pre-eminence as an uropathogen; the pathogenesis, genomics and potential therapeutics against UPEC have been subjects of immense interest.

1.2.1 Pathogenesis of UPEC

Escherichia coli is a bacterium capable of successful colonization of a diverse range of environmental as well as host niches, in the form of both commensals and pathogens. The pathogenic cycle of UPEC in uncomplicated UTI involves several distinct intracellular and extracellular environments and an arsenal of virulence factors to aid in colonization of the urinary

tract. After migration to the bladder via the urethra, the first step involves adhesion and invasion of terminally differentiated superficial bladder epithelial cells lining the bladder lumen. This attachment and invasion is mediated by filamentous adhesive appendages present on the bacterial surface called pili. One of the main pili involved in attachment to the urothelium is the UPEC Type I pilus [8]. These Type I pili mediate intimate contact with the urothelium by binding to mannosylated glycoprotein receptors such as uroplakin 1a and $\alpha 3\beta 1$ integrins present on the host cells. UPEC invasion is Type I pilus dependent and involves localized actin and microtubule rearrangements, which aids in a zipper-like internalization of bacteria [9, 10]. UPEC enters urothelial cells in a clathrin dependent manner and initially reside in Ca^{2+} and cAMP regulated exocytic vesicles [11, 12]. At these initial stages of infection, the host response consists of antimicrobial peptides, neutrophil infiltration, shear force exerted by urine, exfoliation of superficial urothelial cells and expulsion of vesicles containing bacteria [12, 13].

Upon invasion, UPEC are capable of giving rise to complex biofilm like Intracellular Bacterial Communities (IBCs) within the host cell cytoplasm. Fast replicating, rod shaped motile bacterial cells initially form loose colonies in the cytoplasm, gradually giving way to compact slow growing, non-motile coccoid cells which fill up the entire cytoplasm [14]. These form pod like bulges with bacteria enmeshed in polysaccharide rich matrix within the cell cytoplasm, which protrude from the urothelial surface with approximately 10^4 bacteria each [15, 16]. These bacteria containing IBCs are an acute transient state in the infection, whose main purpose is subversion of host immune response in the form of neutrophil attack, secreted pro-inflammatory cytokines and removal by urine flow. Moreover bacteria present in IBCs are phenotypically resistant to antibiotics [15]. Finally, bacteria present in mature IBCs acquire a filamentous form and begin fluxing from the host cell, with the potential to initiate a second round of IBC formation. Exfoliation of the terminally differentiated superficial layer of cells, exposes the underlying layer of undifferentiated cells which are also susceptible to UPEC. However UPEC in these transitional bladder epithelial cells exist in a barely replicating quiescent state enmeshed in actin [17]. These Quiescent Intracellular Reservoirs (QIR) upon differentiation of their host

cells due to epithelial turnover are directed to acidic compartments, where cytoskeletal alterations trigger growth and expulsion [18]. QIR formation depends on the integrity of the urothelium during the initial infection and offers an excellent means of subverting host response to UTI, while maintaining a potential source for a rUTI. Supporting this model, QIRs within transitional cells are capable of seeding reinfection if epithelial exfoliation is chemically induced [19].

Acute stages of UPEC mediated UTI trigger an inflammatory response in the host, which sets the stage for the long term outcome of the infection [13]. A severe initial inflammatory response sensitizes the bladder and remodels the tissue to a chronic cystitis state, wherein the hyperplastic state of the urothelium restricts UPEC to an extracellular state incapable of IBC formation. A mild inflammatory response can result in resolution of the infection, with a small possibility of QIR formation and rUTI. A deficient response usually leads to persistent asymptomatic bacteriuria [20]. Biofilm formation in the extracellular niche, either urothelium or catheter associated is yet another important step in bladder colonization. From the bladder, UPEC can ascend to and proliferate in the kidneys and ultimately result in bacteraemia if the host barriers are sufficiently compromised [21]. IBC formation, which represents an unique and important stage in the UPEC infection cycle is not an artefact of the murine model, as clinical UPEC isolates can form IBCs in 4 different mice strains [22]. Also, urine samples from patients with UPEC contained IBCs (22% of subjects) and filamentous bacteria (45% of subjects), hallmarks of the murine *in vivo* infection model [23].

1.2.2 UPEC virulence factors and genomics

UPEC virulence factors can be broadly categorised as adhesins, iron acquisition systems, toxins and others. Adhesins consist of filamentous surface structures called pili or fimbriae, which mediate attachment to specific host surface receptors. These include Type I pili, P pili, S pili, F1C pili and Dr adhesins [1, 24]. Type I and P pili represent two of the most important and consequently well studied pili recognising mannosylated glycoprotein receptors and globotriacylceramide containing glycolipid receptors respectively. Since the distribution within the urinary tract of their

corresponding receptors varies, accordingly Type I pili primarily mediate attachment in the bladder and P pili in the kidneys. However, Type I pilus is the most exciting UPEC virulence factor because it is required for attachment, invasion as well as biofilm and IBC formation, with Type I pilus deficient strains being significantly attenuated *in vivo* both in mice [24] and patients [25]. Iron is an essential nutrient for bacterial growth and owing to the incredibly iron deficient nature of the urinary tract, UPEC encode several seemingly redundant iron acquisition systems. These iron acquisition systems primarily consist of siderophores such as Aerobactin, Enterobactin, Salmochelin and Yersiniabactin, which bind to iron with high affinity and scavenge it from the environment [26, 27]. Secreted UPEC toxins such as α -hemolysin, cytotoxic necrotising factor 1 (cnf1), vacuolating autotransporter protein (vat) and secreted autotransporter protein (sat) play diverse roles in UTI progression. Pore forming toxin α -hemolysin can lead to host cell lysis and the release of nutrients, however physiologically relevant levels of the toxin are sub-lytic. α -hemolysin is also capable of altering host serine/threonine kinase (Akt) signalling, disrupting cellular processes such as cell cycle progression, metabolism, vesicular trafficking, inflammation and survival [28]. Cnf1 similarly via constitutive activation of Rho family of GTPases can alter host cytoskeletal arrangement and inflammatory response. Cnf1 can stimulate urothelial apoptosis and exfoliation, exposing underlying transitional cells to UPEC [1, 21]. Other virulence factors include Antigen 43 (Agn43) an adhesive autotransporter protein involved in host cell adhesion and biofilm formation, and flagella, which are supramolecular mediators of bacterial motility, which play important roles in host cell adhesion, biofilm formation and ascension through the urinary tract [29-31].

Pathogenic *E. coli* are simplistically classified as either diarrheagenic *E. coli* or Extraintestinal Pathogenic E. coli (ExPEC). Uropathogenic *E. coli* come under the broad group of ExPEC. The purpose of such general pathotypes is to cluster bacteria on the basis of similarities in clinical outcomes, pathogenesis and virulence factors [27]. Comparative genomic analysis of multiple pathovars including enterotoxigenic, enteropathogenic, enteroaggregative and commensal *E. coli* revealed about 2200 genes conserved in all the pathovars representing the core genome. However,

few pathovar specific genes could be identified and no unique genomic signatures emerged. Taken together, the 17 genomes analysed did reveal an astonishingly diverse pangenome of about 13,000 genes [32]. Comparing the genomes of enterohemorrhagic, uropathogenic and commensal *E. coli*; a few trends begin to emerge. Genomic content differs not only between pathogenic and commensal *E. coli* (8-22% more open reading frames in pathogens) and between pathotypes as expected, but also between sequenced UPEC isolates [33, 34]. UPEC isolates are enriched in virulence factors such as iron acquisition systems, adhesins and autotransporters compared to other pathogenic *E. coli*. However, the sequenced UPEC genomes differ in the types and numbers of virulence factors encoded, primarily associated with pathogenicity islands [27]. Based on genomic analysis there exists a substantial amount of similarity between UPEC and Avian Pathogenic E. coli (APEC), with regards to serogroups, phylogenetic groups and virulence genes. These genomic overlaps include pathogenicity islands and plasmid encoded virulence genes [27, 35]. Thus, APEC might represent a direct zoonotic risk or at least a reservoir of potential virulence phenotypes. It is evident that the classification uropathogenic *E. coli* (UPEC) encompasses a diverse set of pathogenic bacteria having evolved to retain a high degree of genome plasticity and with access to an impressive arsenal of virulence factors, enabling successful adaptation to and colonization of the human urinary tract.

1.2.3 Treatment of UPEC Induced UTI

Treatment strategies for both UTI and rUTI involves antibiotic usage, which can alleviate symptoms and reduce bacterial colonization. Commonly prescribed antibiotics include trimethoprim-sulfamethoxazole, Ciprofloxacin, Nitrofurantoin and Ampicillin [36]. Analysis of a panel of 17 different antibiotics revealed that though planktonic bacteria grown in liquid media are susceptible to killing, when tested against biofilms and IBCs *in vitro* efficacy drops drastically, with IBCs susceptible only to Nitrofurantoin, Ciprofloxacin and Sparfloxacin. This *in vitro* efficacy was entirely absent when extended and tested *in vivo* [37]. Further studies in mice have revealed that 36% will have a recurrent bacteriuric episode similar to humans. Although 10 days of

trimethoprim-sulfamethoxazole treatment reduces recurrences and eliminates faecal colonization, it is unable to eliminate the bladder reservoir [38].

Our understanding of the complex UPEC life cycle and virulence mechanisms sheds light on the persistence of this pathogen even in the face of long term antibiotic therapy and permits the rational design of novel prophylactic small molecule inhibitors. Pilicides which are inhibitors of crucial steps in the assembly of adhesive pili are one such inhibitor. Pilicides with broad spectrum activity, such as ec240 which is active against Type I, P and S pili exist. ec240 functions by inhibiting the phase variable expression of type I pili and effectively shutting down transcription [39]. Mannosides are orally bioavailable, fast acting, high affinity Type I pili receptor antagonists which prevent binding of Type I pili to their mannosylated glycoprotein receptors. Effective against both uncomplicated UTI and catheter associated UTIs [40], mannosides have even shown efficacy against acute as well as chronic bladder colonization by multidrug resistant lineage of UPEC [41]. Combination therapies of new classes of antibiotics with β -lactamase inhibitors, chemical inducers of urothelium exfoliation [19, 42], pilicides and mannosides holds great promise. Vaccines targeting UPEC are primarily directed against the type I pilus adhesion subunit protein FimH, iron acquisition system components and secreted toxins. Vaccination against type I adhesin has been shown to mount an effective urinary tract specific antibody response capable of blocking host pathogen interaction and thus bacterial colonization, in both mice [43, 44] and monkeys [44]. Vaccines against different siderophore components are effective in specific host niches only and vaccination against toxin like α -hemolysin does not prevent colonization, only mitigates severity of clinical outcome [1]. Two commercially available vaccines exist, which either contain membrane components from 18 *E. coli* or 10 heat killed UPEC. However, these are not universally approved or effective [45].

Owing to high prevalence and morbidity combined with a significant rate of recurrence and the immense financial and human burden of UPEC induced UTIs, the shortcomings of current antibiotic therapies become apparent. Although these antibiotics are effective in the short term,

repeated prolonged use in UTI treatment runs the risk of altering host microbiota and more importantly drug resistance in UPEC. Indeed, multidrug resistant UPEC are now beginning to emerge which encode an impressive array of resistance mechanisms such as extended spectrum β -lactamases (ESBLs) such as cefotaximases (CTX-Ms) and oxacillinases (OXAs). Most disturbingly, carbapenemases active against a broad spectrum of β -lactam antibiotics including the last line carbapenems, such as the New Delhi Metallo- β -lactamase (NDM1) have begun to rapidly emerge. Drug resistance amongst clinical UPEC isolates can vary from 8.6% to 48.7% depending on the antibiotic [2] and globally disseminated lineages of drug resistant UPEC such as Sequence Type or ST131, ST69, ST73 and ST95 now exist [46].

Knowledge of Uropathogenic *E. coli* infectious life cycle, virulence factors and genomics has helped clinical management of this pervasive infectious disease and further research into these fundamental bacterial processes will definitely aid in designing better therapeutic strategies.

1.3 Epigenetic DNA Methylation

Epigenetics refers to functionally relevant, heritable information encoded by the genome of living organisms, which does not include the underlying nucleotide sequence. This definition can cover several distinct mechanisms such as regulatory RNAs, chromatin remodelling and covalent modification of DNA. Arguably one of the most common and well-studied epigenetic mechanisms in all kingdoms of life is DNA methylation. Methylation of DNA is mediated by enzymes called DNA methyltransferases and involves covalent attachment of a methyl group either to the carbon ring of cytosine (producing 5-methylcytosine, 5mC) or the exocyclic amino groups of cytosine or adenine (producing N4-methylcytosine, 4mC or N6-methyladenine, 6mA respectively) [47]. The functional role of DNA methylation in eukaryotes has been studied primarily in the context of 5mC methylation of CpG nucleotides. More than half of vertebrate genomes possess CpG nucleotide rich stretches called CpG islands or (CGI) in their genomes. 5mC methylation of CGIs present at the transcription start sites are repressive marks and result in gene silencing. Although 5mC does

occur at CGIs present at other sites and also other genomic contexts such as gene bodies, regulatory elements and repeat sequences, we are still learning about their mechanisms and functions [48]. 6mA which was believed to be an exclusively prokaryotic epigenetic mark is now being discovered in regulated sequence and tissue specific contexts in higher eukaryotes such as *Chlamydomonas reinhardtii*, *Caenorhabditis elegans*, and *Drosophila melanogaster* [49]. DNA methylation in humans for example, plays an important role in differentiation and development such as X-chromosome inactivation, genetic imprinting, controlling transposition, development and tissue specific gene expression [50]. DNA methylation also plays an important role in cancer by altering gene expression and mutation rates of genes [48]. This brief overview does not do justice to the fascinating role of DNA methylation in eukaryotes, however the focus of this study is to elucidate the role of this important epigenetic mechanism in bacteria, specifically UPEC.

1.3.1 Types of DNA Methyltransferases

The first evidence for DNA methylation was found in *Escherichia coli* K12 by Joe Bertani and Jean Weigle in 1953, when they reported a barrier to λ phage transduction. Briefly, when λ phage from a heterologous *E. coli* strain was used to infect *E. coli* K12, it failed. But, λ phage propagated in *E. coli* K12 could effortlessly reinfect the homologous strain. This led to 2 pivotal and elegant conclusions: 1. Bacteria are capable of imparting upon their DNA an identification “tag” and 2. Bacteria are capable of recognising and restricting a foreign identification tag. Years of research would show that this barrier was mediated by a Restriction Modification System (RMS) and the identification tag imparted to DNA was DNA methylation. The study of RMSs discovered over half a century ago has led to several major steps in our understanding of DNA methylation, cleavage, translocation, bacterial defence mechanisms and phage evasion strategies, but most importantly allowed us to develop molecular biology tools essential in genome manipulation [51]. DNA methyltransferases in bacteria are of two types: Components of Restriction Modification Systems and Solitary or Orphan Methyltransferases

1.3.1.1 Restriction Modification Systems

Restriction Modification Systems (RMSs) as the name suggests perform two contrasting enzymatic functions, DNA methylation or modification and DNA cleavage or restriction. RMSs perform these contrasting activities in the context of a specific DNA recognition sequence. RMSs ensure methylation of host genomic or self DNA at the specific recognition sequence, however, upon encountering the unmethylated recognition sequence in exogenous or non-self DNA, trigger cleavage. Thus allowing differentiation between self and non-self DNA and selectively degrading potentially harmful foreign DNA such as phages, plasmids, etc. Based on this primary function, RMSs could be considered a prokaryotic innate immune system [52]. RMSs display exceptional diversity and prevalence in both bacterial and archaeal genomes. About 90% of all sequenced prokaryotic genome possess at least one RMS, with 80% containing multiple systems. The distribution of the number of RMSs positively correlates with the size of the host genome. Obligate intracellular pathogens like *Buchnera*, *Borrelia* and *Rickettsia* are devoid of RMSs and naturally competent bacterial species such as *Helicobacter*, *Neisseria*, *Haemophilus* and *Streptococcus* encode a disproportionately large number of RMSs, more than 20 in *Helicobacter pylori* [53]. Arguably, this correlates with the exposure of these different bacterial species to exogenous DNA such as phages. Although primarily involved in host defence, DNA methyltransferases associated with RMSs by virtue of their ability to methylate DNA have the potential to perform additional epigenetic functions. Based on recognition sequence, co-factor requirement, sub-unit composition and cleavage position RMSs can be classified into 4 types, I to IV, summarised in Figure 1.

Type I Restriction Modification Systems consist of three genes, *hsdS*, *hsdM* and *hsdR*. HsdS is the sequence recognition protein, HsdM the methyltransferase and HsdR the endonuclease of these RMSs. Type I RMSs are complex multi subunit systems, with the functional restriction holoenzyme, $R_2M_2S_1$ consisting of two endonuclease (R), two methyltransferase (M) and one specificity (S) subunit. The functional methyltransferase consists of a M_2S_1 heterotrimeric complex. The recognition sequence of Type I RMSs is bipartite, consisting of two specific parts separated by

a degenerate central sequence of fixed length, for example, AAC(N₆)GTGC. Methylation occurs at specific nucleotides on both strands of the recognition sequence and requires S-Adenosyl methionine (SAM) as a methyl donor and cofactor [54]. Restriction of DNA occurs by recognition of two unmethylated sites and ATP/Mg²⁺ dependent bi-directional translocation of DNA until two Type I heteropentameric complexes collide, triggering cleavage. Cleavage thus occurs at a random site each time which is located hundreds to thousands of bp away from the recognition site [55]. RMS recognition sites have been classically identified based on their plasmid cleavage patterns, since both methylation and restriction functions are mediated by the same recognition sequence. The variable position of the Type I RMS cut site explains the difficulty in identifying and hence relative paucity of characterized Type I recognition sequences [51].

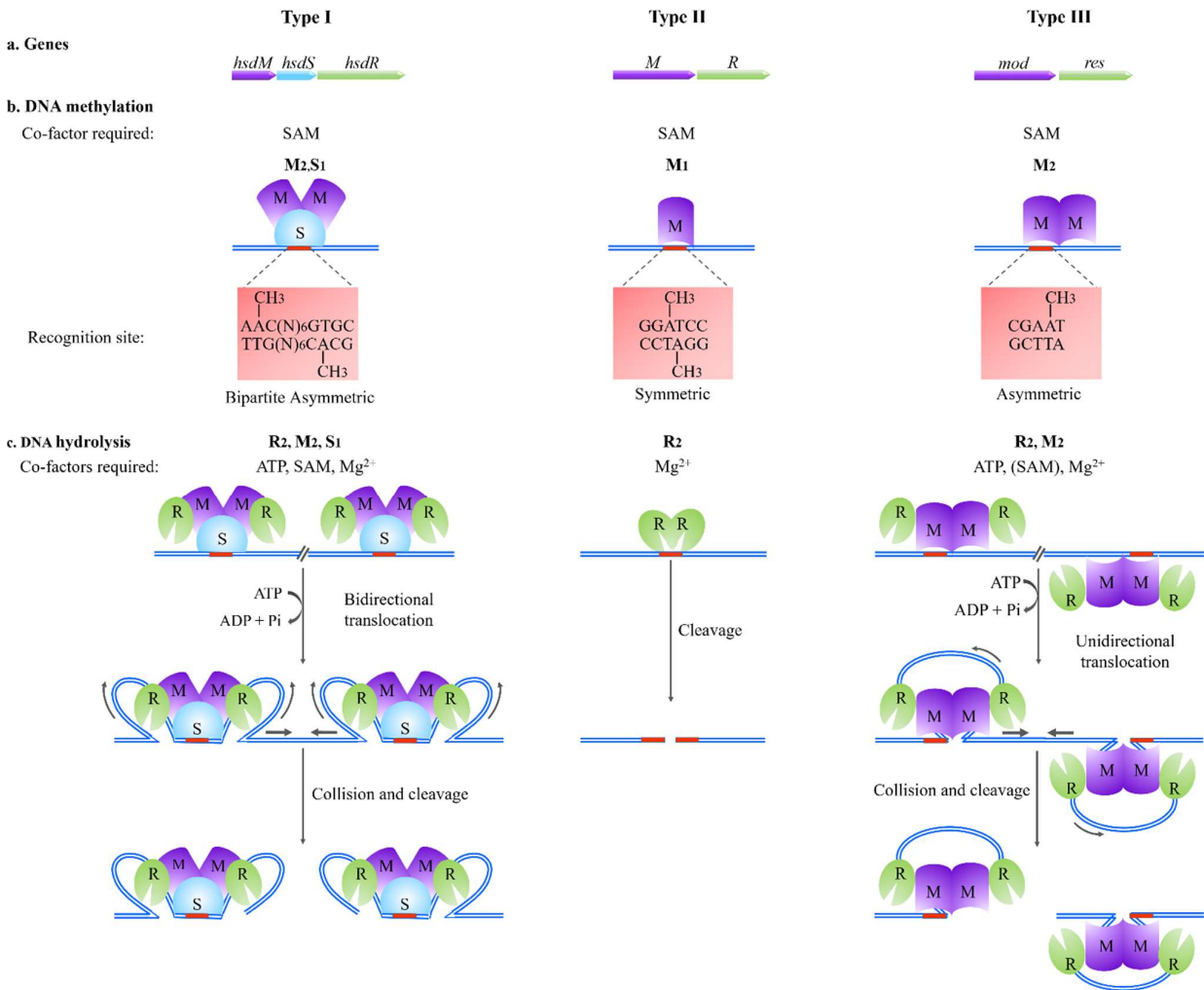


Figure 1: Characteristics of Type I-III Restriction Modification Systems. **A)** Genetic organisation of RMS genes. *hsdM*, *M* and *mod* – DNA methyltransferases. *hsdR*, *R* and *res* – endonucleases. *hsdS* – Type I RMS specificity determining DNA binding sub-unit. **B)** DNA methylation – holoenzyme components, cofactor requirements and recognition site characteristics. **C)** DNA cleavage – holoenzyme components, cofactor requirements, cleavage site and mechanism. M = Methyltransferase, R = Endonuclease and S = Specificity subunit.

The Type I HsdS sub-unit consists of two highly variable target recognition domains (TRDs) separated by conserved regions between and flanking the TRDs. It is these TRDs, each binding to one half of the bipartite recognition sequence, which determines the Type I recognition sequence. The central conserved region present between the TRDs can vary in length and this correlates with the length of the degenerate sequence present in the recognition site [56, 57]. In addition to the flanking conserved regions, *hsdS* TRDs sometimes contain flanking invertible repeats [58]. The presence of these elements permits Type I RMSs to exchange not just the *hsdS* gene, but also individual TRDs and even vary the central degenerate sequence length to generate a phenomenally diverse repertoire of recognition sequences. Evidence for such Type I RMS diversity comes from the existence of systems with combinations of TRDs which match precisely the combinations of bipartite recognition sequences [51, 54, 56]. Type I RMS genes *hsdSMR* usually occur at the same locus, however the number of sub-units and genomic location can vary. Multiple HsdS each with a different recognition sequence, located at different loci can be recruited by the same HsdMR to form a functional RMS. This genetic complementation occurs naturally in diverse bacteria such as *Mycoplasma pulmonis* [59, 60], *Bacteroides fragilis* [61] and *Streptococcus pneumoniae* [62, 63] as well as experimentally in the lab, provided the genes belong to the same family of Type I RMSs. Overall Type I RMSs represent extremely flexible molecular machines with the potential to mediate a diverse repertoire of epigenetic DNA methylation.

Type II Restriction Modification Systems consist of two genes, encoding a methyltransferase (M) and an endonuclease (R). Both genes contain DNA recognition domains and are capable of enzymatic activity independent of each other. The recognition site for these systems consists of short (4-8bp) usually palindromic but occasionally asymmetric sequences, for example GAATTC. Cofactors required by these systems include SAM for methylation and Mg^{2+} for restriction. The site of endonuclease digestion is at a fixed position at or near the recognition site, with no need for DNA translocation. The precise nature of DNA digestion mediated by these systems has allowed for their easy biochemical characterization. Also, the independent recognition and enzymatic activities has allowed cloning and commercial production of these endonucleases as precise tools in molecular biology. The classification of Type II systems is based more on enzymatic behaviour than phylogeny, resulting in a rather heterogeneous group of systems [64]. Since the two activities of these systems can act independently of each other, loss or inefficient methyltransferase activity risks endonucleolytic attack on the host genome. Thus, Type II RMSs can be considered as selfish genetic elements [65]. This has also resulted in the evolution of temporal regulatory mechanisms for the methyltransferase and endonuclease, so as to ensure adequate genomic methylation and mitigate toxicity to the host [66].

Type III Restriction Modification Systems are also encoded by two genes, a methyltransferase (mod) and an endonuclease (res) gene. Only the methyltransferase has the ability to recognise the Type III recognition site. Thus the functional methyltransferase consists of a homodimeric (mod_2) holoenzyme, while the functional endonuclease consists of either heterotrimeric ($mod_2 res_1$) or heterotetrameric ($mod_2 res_2$) holoenzyme. The recognition site for Type III RMSs consists of short (4-6bp) non-palindromic sequences, for example CGAAT, with methylation characteristically occurring only on one of the two strands. DNA cleavage requires binding of Type III enzyme complex to two inversely oriented (head to head oriented) unmethylated recognition sites, followed by DNA translocation, stalling due to collision between the two systems and cleavage at a fixed position approximately 25 to 27 bp from one of the recognition sites. Type III RMSs require SAM as the methyl donor for methylation and Mg^{2+} as well as ATP for

translocation and cleavage of DNA [67]. Similar to other RMSs methylation can occur in the absence of the restriction component, as the methyltransferase encodes the DNA recognition domain [68]. However, unlike Type II RMSs where the recognition domain is present on both M and R, alteration in DNA recognition domain does not risk endonucleolytic cleavage of the host genome. Thus, similar to Type I RMSs we observe a mod gene organization where numerous alleles differing only in a central DNA target recognition domain exist with conserved flanking sequences and the potential to easily acquire novel specificities via horizontal gene transfer [69].

Type IV Restriction Modifications Systems are not considered as classical RMSs, since they do not possess a methyltransferase component. Unlike other RMSs, they recognise and cleave modified DNA [68]. The recognition site for these systems has very low specificity, such as RC (where R = A or G) separated by about 40 to 3000 bp. Also, the type of methylation detected by these systems are highly variable, such as 5-methylcytosine (5-mC), 5-hydroxymethylcytosine (5-hmC), N4-methylcytosine (4-mC), N6-methyladenine (6-mA) or even 5-glucosylhydroxymethylcytosine. These systems utilise GTP for translocation and cleavage of DNA [70]. Since the recognition sequence is relatively promiscuous and the type of methylation detected is rare but often found in some phage genomes, Type IV RMSs still behave as host defence mechanisms, capable of killing an infected host as well [71]. However, since these systems do not encode a methyltransferase component, they are relatively less interesting from the epigenetic perspective.

1.3.1.2 Orphan Methyltransferases

Orphan or solitary methyltransferases as the name implies are stand-alone enzymes lacking any partner endonuclease. Like Type II RMS methyltransferases, orphan methyltransferases predominantly utilize SAM to methylate either cytosine or adenine in short (4–6 bp) palindromic sequences, for example GATC, CCWGG (W = A or T) and RCCGGY (R = A or G and Y = C or T) [72-74]. Orphan methyltransferases are believed to be evolutionary degradation products of RMSs, where the endonuclease has gradually been lost from the genome. Indeed, orphan

methyltransferase *dam* and Type II RMS DpnI both recognise the same palindrome, GATC. A study of over 1000 genomes revealed that orphan methyltransferases far outnumber methyltransferases encoded by RMSs, based on sequence analysis [75]. A far more comprehensive and broad analysis of actual DNA methylation patterns in 230 (217 bacterial and 13 archaeal species) prokaryotic genomes revealed that over 48% of methylated Type II motifs are due to orphan systems. This number is probably higher, since the presence of the endonuclease gene does not necessarily mean that its enzymatic activity is functional, which is not experimentally verified in these studies. Type I and III orphan methyltransferases in contrast constitute only 2.8% and 6.7% respectively of the identified methyltransferases [76]. Also, while RMSs show a high degree of variability, orphan methyltransferases are phylogenetically well-conserved across a particular genus/family and even essential in some bacteria [75, 76]. Thus, although orphan methyltransferases may have lost their ability to function as classical RMSs, the conserved nature of these enzymes and their ability to methylate DNA at specific sequences is indicative of alternative epigenetic functions.

1.3.2 Detection of DNA Methylation

Direct detection of methylated bases in DNA has been classically achieved by using either radiolabelled substrates or mass spectroscopic analysis. But depending on the method, they have severe drawbacks such as lack of genome wide analysis and inability to determine the specific sequence context of methylation [49, 51]. The first truly genome wide, strand and context specific detection of DNA methylation was achieved with the advent of bisulfite sequencing. This method involves the conversion of cytosine to uracil via bisulfite treatment followed by conventional next generation sequencing, resulting in C to T conversion of unmethylated cytosines only. This method has been the gold standard for detection of methylated DNA and a major aid in our understanding of the function of DNA methylation. However, Bisulfite sequencing can specifically detect 5-mC and 5-hmC only, and is also sensitive to differences in efficiency of bisulfite conversion of DNA [77]. The field of DNA methylation has been revolutionized with the advent of third generation

sequencing platforms, especially Single Molecule Real Time (SMRT) Sequencing by Pacific Biosciences [77].

SMRT sequencing is unique because the incorporation of fluorescently labelled nucleotides in the sequencing reaction is detected in real-time, allowing an in-depth analysis of sequencing polymerase kinetics. This real-time detection is possible since each polymerase enzyme molecule is immobilized in individual reaction wells called zero mode waveguides [78]. Incorporation of a nucleotide complementary to a modified nucleotide such as N6-methyladenine, N4-methylcytosine and 5-methylcytosine amongst others, causes a characteristic and significant change in polymerase kinetics [79]. When compared to an *in silico* control for polymerase kinetics of unmodified DNA it is possible to reliably detect, genome wide, strand and context specific methylation patterns in bacterial as well as eukaryotic DNA [80]. Combined with the long read length (>10kbp), lack of sequencing bias, high consensus accuracy and use of native DNA, SMRT sequencing allows *de novo* assembly of bacterial genomes and the reliable analysis of the epigenome [81, 82].

Since SMRT sequencing allows the detection of methylation independent of the presence or type of RMS restriction activity, researchers can now easily identify Type I and III RMSs, vastly increasing the numbers of such systems [76, 83, 84]. SMRT allows the detection of Type II and III RMSs with degenerate specificity, such as those observed in *Bacillus cereus* [85]. Detection of methylation systems in diverse bacterial and archaeal species, which differ in important genomic characteristics such as size, G+C and repeat content, is possible [76, 84]. SMRT has also allowed the identification of unique sub-types of RMSs such as Type IIG, IIS and IIT whose identification by conventional methods such as studying DNA cleavage patterns would have been difficult. This is because of several reasons such as presence of a single fused M/R protein for Type IIG, presence of cleavage sites outside the recognition site either on one or both sides, diversity of recognition sites (symmetric, asymmetric or even bipartite) and in some cases, a tendency towards incomplete cleavage of DNA [85, 86]. Finally, SMRT sequencing has allowed the analysis and characterisation of the complex methylomes of bacteria such as *Helicobacter pylori*, which can simultaneously

express 17 to 20 functional methyltransferases [87]. Thus, it is now possible to temporally detect specific methylated, hemimethylated or unmethylated sites at a genome wide level [76, 88-90]. With tools such as SMRT sequencing our ability to characterize DNA methyltransferases has greatly improved, revealing a vast diversity of such systems and possible roles in epigenetic regulation.

1.4 Role of DNA Methylation in altering virulence

The virulence phenotype of a pathogenic bacterium is a direct consequence of specific spatial and temporal gene expression patterns. Methylation of DNA irrespective of the enzyme responsible, has the potential to alter expression of genes required not only for essential physiological processes, but also virulence. The aim of this section is to summarise the effect of DNA methylation mediated by orphan as well as Restriction Modification System methyltransferases on a diverse set of bacterial pathogens.

1.4.1 Orphan Methyltransferases

Arguably the majority of methyltransferases studied to date capable of altering virulence are orphans. The most well characterised being DNA Adenine Methyltransferase (*dam*), DNA Cytosine Methyltransferase (*dcm*), and Cell Cycle Regulated DNA Methyltransferase (*ccrM*). For *dam* and *ccrM* the molecular mechanisms used in order to execute specific functions is also known [91]. It is important to note that methylation of DNA can alter the thermodynamics and curvature of DNA, and also the ability of important regulatory proteins such as RNA polymerase and transcription factors to bind to their target sites [92]. Besides the locus specific effects of these two orphan methyltransferases the mechanistic cause for global gene expression changes for orphan as well as RMS methyltransferases remains unclear.

1.4.1.1 **DNA Adenine Methyltransferase (*dam*)**

Dam is a highly conserved methyltransferase found in *Gammaproteobacteria* which includes many important Gram negative pathogens such as *Escherichia coli*, *Salmonella spp*, *Vibrio cholera*, *Yersinia spp*, *Haemophilus influenzae* and *Neisseria meningitidis*. The target sequence for Dam is 5'-GATC-3' with the adenine on both strands undergoing N6-methylation [91]. Unlike RMS associated methyltransferases with the same sequence specificity, Dam is a highly processive enzyme. Although Dam shows equal binding affinity for both hemi and unmethylated sites irrespective of the local sequence context, the kinetics of methylation are affected up to 12 fold by the local sequence context of individual GATC sites, especially poly-adenine tracts [93]. These features of Dam mediated methylation play an important role and known Dam regulatory mechanisms involve arrays of GATC sites present in the promoters of regulated genes whose methylation status can affect protein binding and gene expression. Dam has been primarily studied in *Escherichia coli* and *Salmonella spp*.

E. coli Dam plays an important role in regulating DNA replication and mismatch repair. Methylation of three GATC sites in the promoter of the *dnaA* gene, located at the origin of replication causes expression of DnaA which is responsible for initiation of replication. Immediately thereafter, protein SeqA binds to these specific hemimethylated GATC sites to prevent further DnaA synthesis till the completion of one round of replication. Thus, Dam methylation is required for co-ordination of replication and cell division [94]. As replication proceeds GATC sites immediately after the replication fork are transiently hemimethylated. The mismatch repair system *mutHLS* scans DNA looking for mismatches (replication errors) and utilises the methylation status of GATC to differentiate the template strand from the newly synthesized strand to correct errors [92]. Antibiotic stress stimulates error prone DNA polymerase IV which, in the absence of *dam* directed mismatch repair, reduces the survival of pathogenic *E. coli* like UPEC and even drug resistant strains [95]. Although *dam* is not an essential gene in *E. coli*, deletion mutants are severely attenuated, partially due to increased mutation rates and have also been tested as potential live

vaccines *in vivo* [96]. A Dam methylation site present at the -10 RNA polymerase binding site of insertion element IS10 transposase gene *tnp* regulates its transcription. Only during replication, the expression of *tnp* is possible due to transient hemimethylation and thus Dam methylation is capable of regulating transposition [94]. Deletion of *dam* in *E. coli* resulted in the differential expression of several hundred genes, confirmed by transcriptomics and proteomics in multiple conditions as well as growth phases. Genes upregulated in the *dam* deficient strain include those involved in respiration, DNA and amino acid metabolism, and SOS response, while downregulated genes were found to be involved in anaerobic respiration, chemotaxis and motility. Computational analysis revealed a biased distribution of GATC sites in the regulatory regions of the *E. coli* chromosome, with an enrichment in regulatory regions flanking genes identified as differentially expressed [97, 98].

Dam methylation has also been studied in the context of specific *E. coli* pathotypes such as Enterohaemorrhagic *E. coli* (EHEC). EHEC adhesion and actin pedestal formation upon contact with intestinal epithelial cells is dramatically increased in *dam* deletion strains. Virulence factors such as intimin and Type 3 secretion system effectors show increased expression via an unknown mechanism independent of transcription. This Dam mediated effect is unique both in its potential mechanism and in the fact that it increases virulence [73]. Pyelonephritis associated adhesive P-pili and autotransporter adhesin Agn43 are important Uropathogenic *E. coli* (UPEC) virulence factors as discussed earlier and are subject to Dam methylation regulated phase variation. Phase variation refers to the rapid reversible on/off switching of gene expression, which allows the host to quickly respond and adapt to environmental cues. Upstream of P- pili genes *papAB*, there are six binding sites for the regulatory protein leucine-responsive protein (Lrp) of which two have overlapping Dam GATC sites. Lrp cannot bind to its binding site when it is methylated and depending on this methylation status, Lrp can either bind and repress P-pilus or vice versa [99]. Similarly, the upstream region of *agn43* has three binding sites for LysR-like factor OxyR and these OxyR binding sites each have an overlapping GATC site. Methylated GATC sites do not allow binding of OxyR and permit transcription, while hemi or unmethylated sites result in OxyR binding

and *agn43* repression [100]. Both these *dam* methylation regulated epigenetic switches of UPEC are complex multi-step epigenetic mechanisms which aid in virulence.

Salmonella spp. also encode a *dam* homologue which has been the subject of great interest. *Salmonella enterica* serovar *typhimurium* lacking *dam* is attenuated, with significantly reduced murine LD₅₀ values and even considered as a potential live vaccine [101, 102]. *Salmonella* lacking *dam* methylation was incapable of colonising deep tissue sites such as the liver and spleen [101], and showed reduced cytotoxicity, invasion and intracellular survival in phagocytic cells *in vitro* [102]. *Salmonella typhimurium* lacking *dam* methylation shows up to 139 genes upregulated and 37 genes downregulated when compared to wild type. These gene expression changes resulted in altered levels of secretory proteins, reduced motility and chemotaxis, reduced bile resistance and attenuated SOS response [103, 104]. Deletion of *dam* resulted in gene expression changes only in pathogenic and not lab adapted *Salmonella typhimurium* strains [104]. A locus specific mechanism for dysregulation has been elucidated for the *Salmonella* Std fimbrium. Std fimbrial expression is activated by the LysR-like protein HdfR and repressed by hemimethylated GATC binding protein SeqA. In the absence of *dam* methylation wherein SeqA no longer binds to GATC sites or a deletion of SeqA itself, HdfR mediated upregulation of Std fimbrial transcription occurs; which contributes to *in vivo* attenuation [105]. Another phase variable locus reminiscent of the *agn43* locus of *E. coli* is the *opvAB* locus of *Salmonella enterica*. In the OpvAB^{ON} state a shorter O-chain length LPS is formed, while OpvAB^{OFF} state results in normal longer LPS O-chain length. The *opvAB* operon contains four *dam* GATC methylation sites and four LysR-like protein OxyR binding sites. Depending on the combinations of GATC sites methylated, OxyR binds at the remaining unmethylated sites and this rapid switching of methylation and OxyR binding can switch expression of OpvAB on or off [106]. The normal LPS O-chain serves as a receptor for several *Salmonella* phages and hence the OpvAB^{ON} sub-population although less virulent can survive phage attack. After the phage pressure is relieved this epigenetic phasevariable element can switch again and restore virulence [107].

Other Gammaproteobacteria also possess *dam* homologues. In *Yersinia pseudotuberculosis* and *Vibrio cholerae*, *dam* is essential and it is believed that *dam* methylation could be involved in co-ordinating replication and maintenance of the two chromosomes harboured by these bacterial species [91]. *Y. pseudotuberculosis* and *V. cholera* strains overexpressing *dam* are attenuated and completely protective in the murine gastric infection model [108]. Moreover, *Y. pseudotuberculosis dam* methylation is involved in mismatch repair, resistance to detergents and overexpression mutants ectopically secrete virulence factors [108, 109]. Similarly, *dam* is also essential in *Yersinia enterocolitica* with overexpression mutants showing increased rate of spontaneous mutations, increased motility, increased proportion of rough O-side chain containing LPS and reduced resistance to detergents [110, 111]. Similarly, lack or overexpression of *dam* methylation results in altered virulence factor expression which results in attenuation of *Aeromonas hydrophila* [112], *Actinobacillus actinomycetemcomitans* [113] and *Klebsiella pneumoniae* [114].

1.4.1.2 **DNA Cytosine Methyltransferase (*dcm*)**

Dcm is an orphan methyltransferase which performs 5-methylcytosine methylation of the second cytosine in the sequence 5'-CCWGG-3' (W = A or T). Although conserved in all sequenced *E. coli*, the exact biological role of this methyltransferase has remained elusive [115]. Cells lacking *dcm* show numerous gene expression changes, but only during stationary phase. Most notably, lack of *dcm* results in reduced expression of ribosomal genes and increased expression of stress induced sigma factor RpoS as well as its downstream targets [74]. Other *dcm* regulated genes such as cold shock protein b (*cspb*) could be required for optimal stationary phase growth at 20°C [116]. Lack of *dcm* also results in increased expression of antibiotic efflux transporter SugE, resulting in increased resistance to EtBr [117]. Although not involved in virulence, there is evidence for *E. coli dcm* methylation mediated regulation of stationary phase gene expression via an unknown mechanism.

1.4.1.3 Cell Cycle Regulated DNA Methyltransferase (*ccrM*)

CcrM is a N6-methyladenine methyltransferase targeting adenine in the sequence 5'-GANTC-3' (where N = A, T, G or C). This methyltransferase is conserved amongst most *Alphaproteobacteria* including *Agrobacterium tumefaciens*, *Rhizobium meliloti* and *Brucella abortus* [91]. However, CcrM has been studied mainly in the context of the Gram negative bacterium *Caulobacter crescentus*, known for its dimorphic life cycle consisting of a swarmer and stalked cell type. *ccrM* plays an important role in regulating the *C. crescentus* cell replication cycle. Unlike *dam*, *ccrM* is not constitutively expressed and does not perform mismatch repair [118]. Instead it is one of four hierarchical cell cycle master regulators, namely DnaA, GcrA, CtrA and CcrM in that order. DNA replication is initiated by DnaA and results in hemimethylated GANTC sites starting from the origin, progressing through the chromosome as replication proceeds. Each regulator sequentially activates the next until finally CcrM is expressed, which then rapidly methylates all the hemimethylated GANTC sites in the genome followed by cell division [90, 118]. This gradual transition of the *C. crescentus* epigenome from methylated to hemimethylated and back to fully methylated has been tracked in exquisite single base resolution at every step in the replication cycle by SMRT sequencing [90]. Computational analysis indicates that GANTC sites are overrepresented in the intergenic regions. Comparing the time of activation of genes, the position of GANTC sites in their promoters and the methylation state of these GANTC sites at the time of their regulation, we can list 59 genes regulated by *ccrM*. Not surprisingly these genes are mainly involved in cell cycle progression, cell division and DNA metabolism [90, 118]. Epigenetic DNA methylation thus plays an important role in the complex cell cycle progression of *C. crescentus* and potentially other *Alphaproteobacteria*.

1.4.1.4 Other Orphan Methyltransferases

Additional orphan methyltransferases from diverse bacteria have been recently discovered and their role in epigenetically altering gene regulation analysed. *Vibrio cholerae* Methyltransferase (*vchM*) is a 5-methylcytosine methyltransferase targeting the first cytosine in the sequence 5'-

RCCGGY-3' (where R = A or G and Y = C or T). *V. cholerae* mutant lacking *vchM* shows both *in vitro* growth defect as well as attenuation in the murine gastric infection model. The methylation mediated by *vchM* is not dynamic, rather it is stably maintained on both chromosomes throughout the cell cycle. Deletion of *vchM* does result in global gene expression changes, with 134 genes upregulated and 63 downregulated. Analysis of its genetic interactions revealed that envelope stress responsive sigma factor RpoE (σ^E) is not essential in a *vchM* deletion mutant. Genes regulated by *vchM* include several cell envelope maintenance proteins including those involved in LPS biosynthesis and those localised at the cell envelope. This implies that *vchM* mediated epigenetics results in altered expression which induces envelope stress, making the RpoE gene essential in wild type. No satisfactory molecular mechanism exists, except for one gene whose expression is regulated by a single RCCGGY site within its coding sequence [72].

Mycobacterium tuberculosis encodes the Mycobacterial adenine methyltransferase *mamA* gene, which methylates adenine in the sequence 5'-CTCCAG-3'. Deletion of this methyltransferase resulted in only modest gene expression changes (11 genes). However, in several of the dysregulated genes, the methylation site CTCCAG occurs at the same position relative to the transcription start site and overlaps with a putative σ factor -10 binding site. The lack of *mamA* methylation does not result in *in vivo* or *in vitro* defects. However, under hypoxic growth conditions mimicking *M. tuberculosis* granulomas, the *mamA* mutant is attenuated [119]. Further analysis of 12 different *M. tuberculosis* strains revealed a functional orphan Type I RMS methyltransferase (*hsdSM*, *hsdR* is missing) in addition to *mamA*. For both *mamA* and *hsdSM* the number of constitutively unmethylated sites is higher than expected, possibly due to masking by DNA binding regulatory proteins. Intriguingly, the relatively rare (600 to 700 sites / genome) Type I recognition sites of *hsdSM* are significantly enriched in intergenic regions [89]. This provides interesting hints towards additional regulatory roles for these orphan methyltransferases.

Helicobacter pylori encodes an orphan 5-methylcytosine methyltransferase called *hpyAVIBM* and this particular allele is present in 83% of symptomatic strains and only in 25% of asymptomatic

strains tested. Deletion of this methyltransferase results in altered gene expression, specifically outer membrane proteins, LPS biosynthesis genes, motility associated genes, virulence factors including adhesins and other RMSs. Phenotypically *H. pylori* strains lacking *hpyAVIBM* exhibit altered motility, LPS profile and elicit a heightened IL8 response. This orphan methyltransferase also possesses polymeric AG repeat tracts, which due to slipped-strand mispairing can result in phasevariable (reversible ON/OFF) expression of *hpyAVIBM* [120].

Mycoplasma hyorhinitis encodes two 5-methylcytosine methyltransferases (*mhy1* and *mhy2*), both methylating cytosine in the dinucleotide 5'-CG-3'. *M. hyorhinitis* in planktonic growth displays two subpopulations, displaying high and low levels of CG methylation. However, upon infection only the low CG methylation population survives, escapes endosomal degradation, undergoes exocytosis and re-infects new cells. Thus low CG methylation pattern in the *M. hyorhinitis* genome predisposes it to intracellular survival and replication. Although Mhy1 and Mhy2 methyltransferases escape and enter the mammalian host nucleus, their effect on host CpG methylation and its functional consequences are unknown. Thus, epigenetic regulation by this orphan methyltransferase potentially alters gene expression to aid in intracellular survival [121].

1.4.2 Restriction Modification System Methyltransferases

There is a paucity of information regarding the role of methyltransferases associated with RMSs in regulation of bacterial virulence. This probably stems from the fact that RMSs are believed to be first and foremost bacterial defence systems against exogenous DNA. There are however a few exceptions, where the functions of this diverse and prevalent set of epigenetic information have been investigated. It should be noted that not all studies test for the activity of the restriction endonuclease component of a RMS. This is important because any inactivating mutation such as a truncation, single nucleotide polymorphism or frame shift mutation in the endonuclease of a RMS would functionally result in an orphan methyltransferase.

1.4.2.1 Phasevarions

Phasevarions are the most well characterised set of RMS associated methyltransferases. Phasevariation refers to the high-frequency, rapid and reversible ON/OFF switching of gene expression patterns or regulons. Phasevariable Type III RMSs have been discovered in several important host adapted pathogens such as *Helicobacter pylori*, *Neisseria meningitidis*, *Neisseria gonorrhoeae* and *Haemophilus influenzae*. They can regulate the expression of a diverse set of genes, depending on the presence or absence of DNA methylation [122, 123]. These phasevariable regulons or Phasevarions possess characteristic homopolymeric dinucleotide, tetranucleotide or pentanucleotide repeat tracts of varying lengths either in the gene body or promoter of the methyltransferase (*mod*) gene [69, 123]. Phasevariation occurs by a mechanism called slipped strand mispairing, wherein owing to DNA polymerase errors at these repeat tracts, the number of repeats changes resulting in reversible ON/OFF expression of the *mod* gene [123]. This phasevariable methylation by Type III systems resulting in altered gene expression is a unique mechanism for increasing phenotypic diversity and fitness within the host.

The *mod* gene encoding the methyltransferase of Type III RMSs contains a DNA target recognition domain (TRD) which determines the recognition site specificity [67]. Computational analysis of different *mod* alleles methylating different sequences demonstrated significant identity between these genes from distinct bacterial species, hinting at horizontal gene transfer. Furthermore, the TRD of *mod* alleles shows mosaicism with regions of inter- and intra-allelic similarity indicative of recombination driven diversification of *mod*. In certain cases *mod* alleles such as the *modA12* allele of *N. meningitidis* demonstrates not only an *en bloc* transfer of the entire TRD to other *mod* genes, but also an enrichment in clinical isolates, indicative of selection for this particular allele [69]. Importantly, in 70% of sequenced *N. meningitidis* and 20% of sequenced *H. influenzae* phasevarions, the restriction endonuclease (*res*) has an inactivating mutation. This inactivation of *res* genes is also observed in *N. gonorrhoeae* phasevarions associated with two alleles (*modA12* and *modA13*). There are also examples of inactivated *res* in association with

phasevarions in *Mycoplasma spp.* and *Helicobacter pylori* [69, 123, 124]. The frequent identification of Phasevariable Type III RMSs with an inactive restriction (*res*) component demonstrates a selective advantage conferred by the phasevariable gene regulation mediated by the methyltransferase and not the restriction phenotype.

Phasevariable Type III RMSs capable of coordinated switching of gene expression patterns have also been found in *H. influenzae* [124]. Screening of clinical isolates of non-typeable *H. influenzae* responsible for otitis media demonstrated an overrepresentation of five *modA* methyltransferase alleles in more than two thirds of all cases. The effect of phasevariable methylation was studied on both the transcriptome and the proteome, in the context of different *mod* alleles. Differentially expressed genes included outer membrane proteins which include vaccine candidates. Phenotypically phasevariation of *mod* resulted in altered antibiotic sensitivity, evasion of phagocytic killing, biofilm formation and in a chinchilla model of middle ear infection, the predominant *modA2* allele was preferentially isolated in the ON state [125]. *Neisseria gonorrhoeae* contains multiple phasevarions even within the same strain, FA1090. Phasevariation of *modA13* (recognition site: 5'-AGAAA-3') resulted in global gene expression changes which included antibiotic efflux systems, ABC transporters, genes involved in DNA metabolism and others. Moreover, phase variation of *modA13* methylation resulted in altered antibiotic resistance, biofilm formation and adhesion-invasion of *in vitro* cell line model [126]. Similarly phasevariation of the other system *ngoAXmod* (recognition site: 5'-CCACC-3') resulted in differential expression of over 121 genes. Once again there were phenotypic consequences which included altered growth rates, attachment-invasion of cell lines and biofilm formation [127]. *Neisseria meningitidis* strains such as MC58 also encode multiple phasevariable methyltransferases such as *modA11* (recognition site: 5'-CGYAG-3') and *modA12* (recognition site: 5'-ACACC-3'), both of which regulate the expression of multiple genes including surface expressed vaccine candidates Lactoferrin-binding proteins A and B, amongst others [126]. Testing the antibiotic sensitivity of different *N. meningitidis* phasevariable systems against a panel of 13 antibiotics, revealed altered sensitivity to at least two [128]. Finally, analysis of the target site distribution revealed the presence of

methylation sites of phasevarions in the intergenic regions of genes regulated by them [129]. *Helicobacter pylori* has one of the largest repertoires of phasevarions, up to 17 different *mod* alleles. One of the most common phasevarions, containing allele *modH*, is responsible for altered expression of only 6 genes but this includes important virulence factors like *flaA* (motility associated) and *hopG* (outer membrane protein associated with colonisation and gastric cancer) [130]. *Moraxella catarrhalis* also has a phasevariable Type III RMS with three different alleles (*modM1-3*). Screening of clinical isolates revealed that *modM2* is the most common allele, while symptomatic stratification revealed that the allele *modM3* shows a significant association with otitis media. Possibly due to the gene expression switching mediated by this phasevarion conferring a fitness advantage in this host niche. *modM2* is also involved in the regulation of up to 34 genes, many of them virulence factors involved in evasion of host immune responses and colonisation [131].

Although phasevarions have been primarily studied in the context of Type III RMSs, the homopolymeric repeat tracts responsible for phasevariation are also found within Type I and II RMSs [126, 132]. *Campylobacter jejuni* encodes a phasevariable Type IIG RMS, wherein both methyltransferase and endonuclease are fused into a single polypeptide. The stochastic ON/OFF switching of methylation mediated by this system results in altered biofilm formation as well as attachment and invasion of colorectal epithelial cell line. Phasevariation of this system caused differential expression of over 200 genes, with over 23% showing the presence of more than four methylation sites in their coding sequence. This Type IIG RMS is fully functional, with both methylation and restriction functions experimentally confirmed [86]. Potentially phasevariable RMSs have also been identified in a more diverse set of bacteria, including *Mycoplasma spp*, *Pasteurella haemolytica* and *Streptococcus thermophilus* [123].

1.4.2.2 Complex Phasevariable Type I Restriction Modification Systems

Streptococcus pneumoniae clonal populations have been previously shown to have two distinct phenotypic sub-populations based on colony opacity, an opaque variant preferentially

associated with invasive disease and a transparent variant associated with nasopharyngeal carriage. *S. pneumoniae* possesses a unique Type I RMS, which encodes for two distinct *hsdS* genes, one partial *hsdS*, one methyltransferase component *hsdM*, one endonuclease component *hsdR* and one tyrosine recombinase (*psrA* or *creX*) at the same genomic locus [62, 63]. The Type I RMS specificity determining HsdS subunit contains two target recognition domains (TRDs), each recognising one half of a bipartite recognition sequence. The two and a half *hsdS* genes of these Streptococcal Type I RMSs essentially contain five TRDs, each with flanking inverted repeats. It is observed that the five *hsdS* TRDs can undergo random, high frequency recombination to give six different viable *hsdS* combinations with two TRDs each. Thus, a single Type I RMS can recombine to yield six distinct specificities, with both methylation and restriction functions verified experimentally [62, 63]. One of the recombinations responsible for TRD rearrangements is performed by the tyrosine recombinase present at the same locus [62]. The numbers for each of the six recognition sites for strain D39 Type I RMS in its genome ranged drastically from 424 up to 1029. The distribution of these sites did not show any bias for promoters, small RNA, genes or operons; but were enriched on the lagging strand of the genome, a unique feature amongst Type I RMSs [63].

Phenotypically, the phasevariable opaque and transparent colony types are due to changes in methylation patterns mediated by this Type I RMS [62, 63]. Of the six possible phasevariable methylation patterns one correlated with a predominantly opaque colony type (SpnIIIA in strain D39 and *hsdS*_{A1} in strain 556). A non-phasevariable mutant with the SpnIIIA methylation pattern for D39 also strongly correlated with bacteraemia in a murine model of invasive disease, while the other variants did not [63]. Similarly, the other variants resulted in a transparent colony type to different extents (for example SpnIIIB in strain D39 and *hsdS*_{A2} or *hsdS*_{A3} in strain 556). A non-phasevariable mutant with the transparent colony type and methylation pattern correlated with excellent nasopharyngeal colonisation in a murine carriage model [62, 63]. Differentially methylated variants of *S. pneumoniae* not only accounted for variable colony type, but also variable internalization by phagocytic and non-phagocytic cells as well as differential gene expression

patterns most importantly for capsular genes. Also, when the invasive murine model is inoculated with a wild type phase variable D39 strain, it contains a heterogeneous population of all six *hsdS* variants (SpnIII A to F). However, within 4 hours this population predominantly shifts to the SpnIII A opaque invasive variant, indicative of selection [63].

These unique, complex phasevariable Type I RMSs enable *S. pneumoniae* to rapidly and reversibly switch between methylation specificities. Each methylation specificity differentially regulates gene expression resulting in distinct phenotypic patterns, each with a distinct advantage in different host niches. These unique RMSs have also been found in other bacteria such as *Mycoplasma pulmonis* [59], *Bacteroides fragilis* [61], *Treponema medium*, *Campylobacter upsaliensis*, *Enterococcus faecalis* and *Streptococcus agalactiae* [62] but not functionally analysed. It is reasonable to assume that these Type I RMSs can rapidly switch methylation specificities and in the process epigenetically alter phenotypic traits of their host organisms.

1.4.2.3 Classical Restriction Modification Systems

As noted earlier, cases of classical RMS mediated bacterial gene regulation are rare. One such system was discovered in Hemolytic uremic syndrome (HUS)-linked *E. coli* O104:H4. This diarrhoeal *E. coli* strain was responsible for the serious 2011 HUS outbreak in Germany, primarily due to the acquisition of genes *stxAB* encoding the Shiga toxin from a new prophage (ϕ Stx104). Further analysis revealed that the phage ϕ Stx104 also encoded a fully functional Type II RMS targeting the sequence 5'-CTGCAG-3', with over 2486 sites in the genome. Methylation mediated by this Type II RMS resulted in differential expression of 1951 genes representing 38% of *E. coli* coding sequences. Upregulated genes include those involved in ion transport, while motility associated genes were downregulated. Interestingly, an increased abundance of methylation sites was observed within 500bp of motility associated genes, possibly hinting at a direct mechanism of methylation mediated regulation of gene expression. Comparison of strain C227-11 bearing the phage encode Type II RMS with another closely related *E. coli* O104 strain without the prophage, namely strain 55989 showed significant overlap of differentially expressed genes when compared

to C227-11 mutant lacking the Type II RMS [88]. This classical RMS although recently acquired via transduction had significant effects on the gene expression of this pathogenic *E. coli*.

Helicobacter pylori is a host adapted pathogen with an immensely complex methylome with even closely related strains showing highly variable sets of methyltransferases. In fact, 10% of *H. pylori* strain specific genes are RMSs [132]. An analysis of five different *H. pylori* strains revealed evidence of Inter and Intragenic target recognition domain rearrangements for Type I RMSs. This results in the evolution of diverse RMS specificities, with minimal risk of endonucleolytic attack of host DNA due to the modular nature, low numbers of recognition sites and genetic complementation of these systems. Computational analysis revealed regions of hyper and hypomethylation in the *H. pylori* genomes. Hypermethylated regions repeatedly consisted of RNA polymerase beta subunit gene (*rpoB*) and chaperonin (*groEL*) genes. Deletion of one classical Type I RMS of *H. pylori* strain P12 resulted in downregulation of only four genes encoding ATP/GTP-binding proteins and belonging to the same operon. These genes also had three copies of the Type I RMS methylation site within their coding sequence, with one site overlapping a long 22 bp palindromic sequence possibly a DNA binding protein recognition site [58]. Thus a single classical Type I RMS of *H. pylori* did not result in massive global gene expression changes, but did affect expression at a single locus via an unknown mechanism.

1.5 Role of DNA Methylation in Bacterial Evolution

Bacterial evolution is driven not only by spontaneous mutations, but also by the acquisition of horizontally transferred genes. Genetic exchange between closely related bacterial species can result in successful integration due to higher sequence homology. Such frequently transferred genetic material often encodes advantageous alleles whose frequency in a population increases due to selection. However, genetic material can also be transferred between completely unrelated clades due to illegitimate recombination. Although this occurs at a lower frequency, successful gene transfer between distant lineages often confers upon the recipient unique and complex abilities

[133]. However, maintenance of species identity requires controlled uptake of exogenous genetic material. The genetic uniqueness of bacterial species often enables the efficient colonisation of a specific environmental or host niche, which can be lost if a certain degree of genetic isolation is not maintained [134]. A systematic study of bacterial evolutionary mechanisms, using over 600 *Streptococcus pneumoniae* genomes revealed the presence of multiple co-circulating lineages. Stable genomic islands are infrequently transferred between lineages, mainly via transformation, and by integrative and conjugative elements. On the other hand, the prophage content is highly diverse even within lineages, indicative of horizontal transfer which represents a more frequent short term means of genetic exchange [135].

Restriction Modification Systems (RMSs) exert their primary function of host defence by blocking potentially harmful foreign DNA, such as phages and plasmids. However, RMSs never provide full protection against incoming DNA, which is also inherently transient [54, 68]. Moreover, other complementary systems for host defence against phages exist such as CRISPR [136] and BREX [137]. Thus, the selection pressure exerted by phage infection does not fully explain the exceptional diversity and wide distribution of bacterial RMSs [138]. Recognition site avoidance is a mechanism by which phages can subvert RMSs by underrepresenting the recognition sites of RMSs in their genomes; for example palindrome avoidance to subvert Type II RMSs. A systematic study of all available genomes and RMS recognition sites revealed that palindrome avoidance i.e. Type II RMS recognition site avoidance in the corresponding host genome is not universal and occurs only in about 50% of cases [139, 140]. However, recognition sites for other RMSs such as Type I, IIG, IIM, III and IV are not avoided in the vast majority of cases [140]. This implies that for the vast majority of functional RMSs site avoidance also does not explain need for RMS diversity and prevalence. Other slightly indirect functions besides host defence might exist for bacterial RMSs. Their ability to provide a barrier against the majority of foreign DNA, enables them to maintain species identity and drive bacterial evolution [52].

The role of DNA methylation and RMSs in bacterial evolution can be illustrated with the bacterium *Staphylococcus aureus*. The majority of human *S. aureus* infections are caused by one of ten distinct lineages, characterised by their surface architecture. Methicillin-resistant *Staphylococcus aureus* represents six of these lineages and depending on the situation is responsible for 50% of infections [141]. *S. aureus* possesses a functional Type I RMS locus (Sau1) which encodes lineage specific *hsdS* alleles. The presence of this RMS is not only a hindrance to cloning, but also inhibits the transfer of resistance genes and mobile genetic elements from other bacterial species as well as different *S. aureus* lineages to one another [142, 143]. Type I RMS sites are distributed evenly in the genome and each lineage appears to be evolving independently [143]. A less conserved but fully functional Type III RMS is also found in some clinical *S. aureus* strains [144]. The restriction barrier presented by these RMSs is capable of blocking horizontal gene transfer between different MRSA lineages. In fact, the *hsdS* allele of *S. aureus* is used for rapid molecular typing of hospital MRSA isolates to track outbreaks, since its distribution is lineage-specific [141]. This trend of bacterial clades with distinct core and accessory genomes superimposed with clade specific repertoires of functional RMSs is also observed in *Burkholderia pseudomallei*, indicative of a barrier to uncontrolled horizontal gene transfer between *B. pseudomallei* clades [145]. The globally disseminated drug resistant ST131 lineage of Uropathogenic *E. coli* (UPEC) also shows a distinct lineage specific RMS repertoire [146].

Thus, in several diverse bacterial species, there exist lineage specific repertoires of RMSs, which control DNA uptake. By virtue of their defensive function in regulating the uptake of exogenous DNA, RMSs act as regulators of bacterial evolution. Thus, methylation mediated by RMSs has the ability over time to drive bacterial evolution and possibly select for traits or gene expression patterns advantageous to the host. One of the most relevant and obvious examples is the transfer and spread of antibiotic resistance genes within specific bacterial lineages. This represents another secondary, non-defensive albeit more long term role of methylation associated with RMSs.

1.6 Hypothesis and Aims

Bacterial DNA methylation is an important epigenetic mechanism with the ability to regulate gene expression. The majority of information regarding diverse methylation mediated functions is restricted to orphan methyltransferases. There is a paucity of information regarding the full extent and importance of RMS mediated methylation in bacterial virulence as well as basic physiological processes. Also, for most methylation mediated gene expression changes both local and global, we still don't know how this epigenetic modification exerts its effect. Uropathogenic *E. coli* (UPEC) is the primary causative agent for urinary tract infections (UTIs) and has been studied in great detail with respect to its life cycle, niche specific gene expression patterns, immune evasion strategies and diverse repertoire of virulence factors. But the role of epigenetic DNA modification in this important human pathogen is relatively unclear. Based on the pervasiveness of DNA methyltransferases and their ability to regulate bacterial gene expression, we hypothesise that DNA methylation could potentially play an important role in regulating UPEC virulence or other cellular processes. The following are the major aims of my PhD thesis:

Aim 1: Characterise the methylome of UPEC strain UTI89 and assign the identified methylated motifs to their corresponding methyltransferases.

Aim 2: Analyse whether a loss or modification of UTI89 methylation has any effect on virulence and gene expression patterns, using both *in vivo* models and high throughput strategies.

Aim 3: Extend our analysis to other *E. coli* strains with similar methyltransferase systems.

2. MATERIALS AND METHODS

2.1 Media and Culture conditions

All strains were routinely propagated in Lysogeny Broth (LB) [10g/L Tryptone, 5g/L Yeast extract, 10g/L Sodium chloride] at 37°C unless specifically noted otherwise. Other media used include M9 minimal medium [1X M9 salts, 2mM Magnesium sulphate, 0.1mM Calcium chloride, 0.2% glucose] and Yeast extract-Casamino acids (YESCA) medium [10g/L Casamino acids, 1g/L Yeast extract], for specific purposes. Liquid media and agar plates were supplemented where appropriate with ampicillin (100 µg/ml), kanamycin (50 µg/ml) and chloramphenicol (20 µg/ml), purchased from SIGMA, Singapore.

2.2 Strain Generation

All strains and plasmids used in this study are listed in Tables 11 and 12. Primers used for this study are listed in Table 13 and were ordered from SIGMA, Singapore. Deletion mutants were generated in *E. coli* using the lambda Red recombinase based homologous recombination protocol described previously with minor modifications [147]. Red recombinase was expressed from helper plasmid pKM208 in the target strain. Cells were made electrocompetent by diluting an overnight culture 1:100 and growing to $OD_{600} = 0.2 - 0.25$, followed by induction with 1mM IPTG for 30-45mins in LB-ampicillin at 30°C. The induced cells were heat shocked at 42°C for 15 mins and placed on ice for 15 mins, with gentle swirling every 5 mins. Cells were then washed with sterile water twice and sterile 10% glycerol once at 5000 rpm for 10 mins at 4°C, careful to maintain cold conditions throughout. Cells were resuspended in 1/100 of initial culture volume of 10% glycerol and stored at -80°C. Primers were designed to amplify a positive selection cassette (from pKD3 for *cat* or pKD4 for *neo*) using universal priming sites and 50 bp flanking homology to the targeted genomic locus. 1 µg of this PCR product was then electroporated into thawed recombination ready

competent cells using 1mm electroporation cuvettes in a GenePulser XCELL system set at 400 Ω resistance, 25 μ F capacitance and 1500V output voltage (Bio-Rad, Singapore). Cells were recovered in LB at 37°C for 2 hours with shaking, 2 hours statically and finally plated on appropriate selection plates. Double crossover homologous recombination allowed replacement of the genomic target with the amplified selection cassette and correct clones were isolated by purifying to single colonies by restreaking on new plates, validation by PCR using test primers flanking the site of recombination and finally sequencing the targeted genomic locus by Sanger sequencing (1st Base, Singapore).

For generating strains with seamless start to stop codon replacement of a target gene with a different allele, a negative selection strategy was used [148]. As described in Figure 2.A, UTI89 strains bearing different Type I RMS alleles were generated via two successive rounds of homologous recombination. A dual positive-negative selection cassette (from pSLC-217) was used to knockout the native allele, rendering the strain kanamycin resistant and sensitive to growth on rhamnose due to inducible toxin RelE, both present on the same selection cassette. A second round of homologous recombination was used to replace the selection cassette with the desired allele. Desired clones were identified after each round of recombination by purifying to singles by streaking on selective plates (LB with kanamycin and M9 with 0.2% Rhamnose), PCR validation using primers flanking the site of recombination and finally Sanger sequencing the genomic target (1st Base, Singapore).

For replacing an essential gene such as *nrdA* with a mutant allele, a slightly altered negative selection strategy was utilised, Figure 2.B. Rather than replacing the target gene which is not possible, the dual selection cassette was inserted at an upstream locus which was not lethal to the host, in this case gene *yfaL*. Next, a PCR product bearing the desired mutant allele was used to replace a large piece of DNA including the essential gene, using homologous recombination with screening performed as described above. The resulting mutants do not carry over any selection markers or DNA scars and represent seamless allelic replacement mutants.

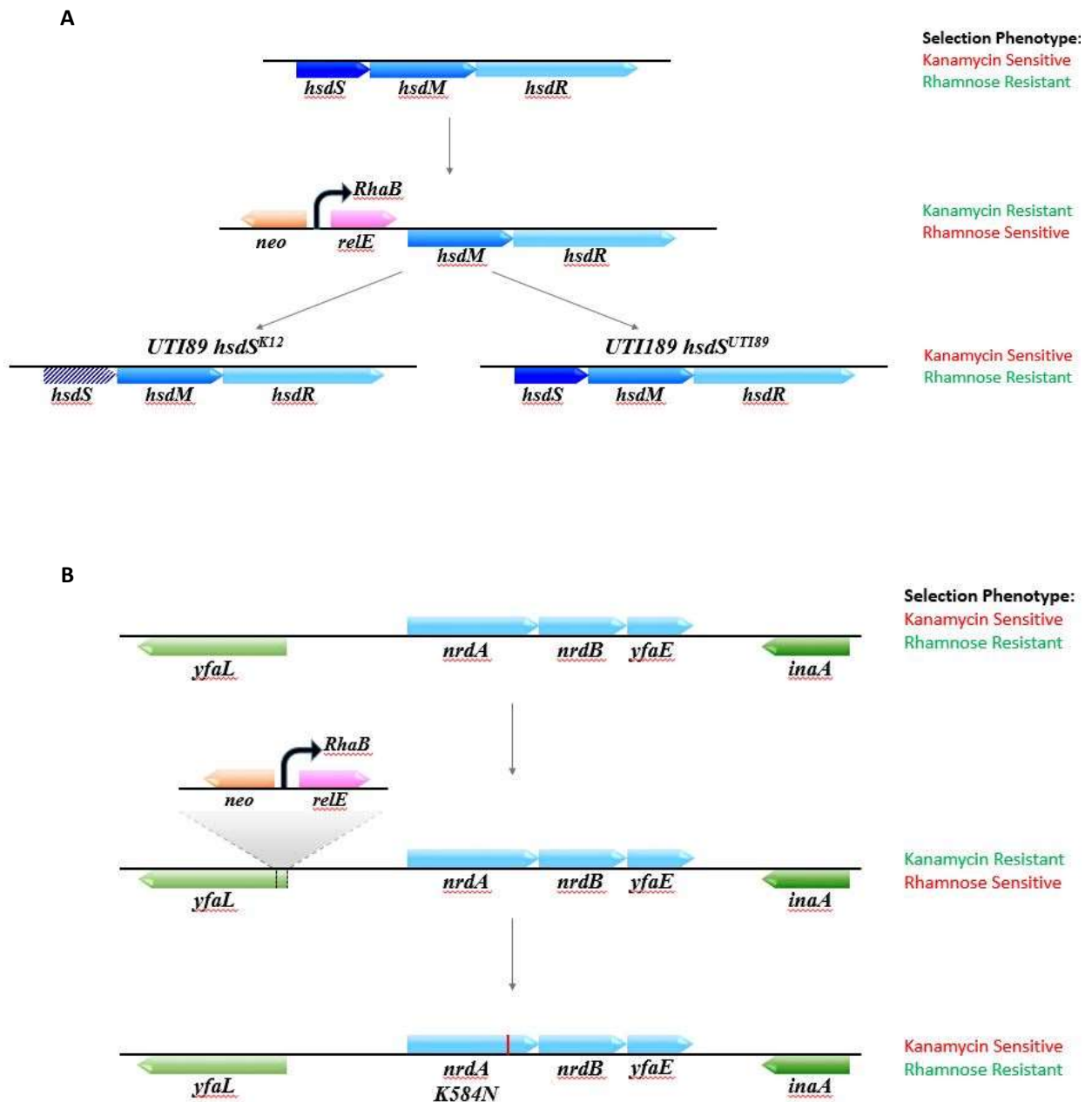


Figure 2: Generating seamless allelic replacements. A) Replacing a non-essential gene such as *hsdS*. Wild type *hsdS* gene is first replaced with a dual positive-negative selection cassette via homologous recombination. Next, using PCR products with flanking homology to the target locus and bearing the desired mutant *hsdS* allele (K12 or wild type UTI89), homologous recombination

replaces the selection cassette to generate the desired strain. **B)** Replacing an essential gene such as *nrdA* with a mutant allele *nrdA* K584N. First, the dual positive-negative selection cassette is inserted at a permissive locus further upstream, depicted by a vertical black line. Followed by a second recombination with a PCR product not only flanking the selection cassette but also the wild type *nrdA* gene so as to replace both and generate the desired mutant. Selection is performed at each step in both **A)** and **B)** using positive selection marker; for example, kanamycin and the rhamnose inducible negative selection marker; for example, toxin RelE. Blue arrows represent homologous recombination steps.

2.3 Plasmid Generation

Plasmids bearing one or two copies of the predicted Type I RMS recognition motifs for UTI89 (5' GAAG (N₇) TGG -3'), K12 MG1655 (5' AAC (N₆) GTGC -3') and CFT073 (5' GAG (N₇) GTCA -3') were generated by PCR mediated insertion. Briefly, low copy number plasmid pACYC184 which does not have any of the three Type I recognition motifs was selected and primers designed to insert the desired recognition motif, followed by plasmid amplification. Amplified linear plasmid was then recircularized by restriction digestion to obtain compatible ends, followed by ligation. The two copies of each recognition site were inserted in the same location exactly 1500 bp apart. Primers used for plasmid generation are listed in Table 12.

2.4 Restriction Modification Assay

Plasmids bearing 1 and 2 copies of the bipartite Type I RMS motif in question were extracted from host strains possessing or lacking the corresponding functional RMS to obtain methylated and unmethylated plasmids, respectively. Plasmids were extracted using the GeneAll Hybrid-Q plasmid miniprep kit (South Korea) according to the manufacturer's instructions. Competent cells were prepared by diluting overnight cultures of the desired strains 1:100 in LB and growing till mid log

phase ($OD_{600} = 0.45 - 0.55$) at 37°C , followed by two washes with sterile water and one wash with sterile 10% glycerol, maintaining chilled conditions. Cells were resuspended in 1/100 original volume in 10% glycerol and stored in 50 μl aliquots. 100 ng of each plasmid was transformed into electrocompetent cells using 1mm electroporation cuvettes in a GenePulser XCELL system set at 400 Ω resistance, 25 μF capacitance and 1700V output voltage (Bio-Rad, Singapore). Cells were recovered in 1 ml LB at 37°C for 1 hour with shaking and plated on selective (LB-chloramphenicol) and non-selective (LB) plates to calculate colony forming units per ml (cfu/ml). Efficiency of Transformation (EoT) was calculated by dividing the cfu/ml obtained on selective vs non selective plates per unit amount of plasmid DNA.

2.5 Mouse Infections

Infections were performed using a well-established murine transurethral model of urinary tract infection [149]. Bacterial strains were cultured in Type I pili inducing conditions by growing in LB statically at 37°C for two 24 hour passages and normalized to $OD_{600} = 1$ with sterile cold PBS. Type I piliation for each strain was evaluated by Hemagglutination assay and Type I phase assay as described previously [150]. Seven to eight week old female C3H/HeN mice (InVivos, Singapore) were anaesthetized using isoflurane and 50 μl of inoculum (2×10^7 cfu) was transurethally instilled into the bladder with a catheter. Post infection, bladders and kidneys were harvested aseptically and homogenized in 1 ml and 0.8 ml sterile PBS respectively. Ten fold serial dilutions were plated on appropriate selective plates and cfu/ml calculated.

For co-infections, the inoculum consisted of a 1:1 mixture of two differentially marked antibiotic resistant strains; otherwise, the procedure was identical to that described for single infections above. For example, wild type UTI89 with Kan^R was co-inoculated with Mutant UTI89 with Chlor^R and plated on respective selective plates for enumeration. To reduce any bias, selection markers were reversed and UTI89 with Chlor^R was co-inoculated with Mutant UTI89 with Kan^R and data presented for each co-infection consist of infections of an equal number of mice infected

with both combinations of marked strains. Bacterial titres from each organ and starting inoculum were calculated and used to calculate the Competitive Index (CI) as follows: $CI = (\text{Output wild type} / \text{Output mutant}) / (\text{Input wild type} / \text{Input mutant})$. CI value was log transformed and plotted.

2.6 RNAseq

Bacterial strains were propagated statically for two serial 24 hour passages at 37°C to mimic the stationary phase inoculum used for murine infections. Log phase cells were obtained by diluting an overnight culture 1:100 in LB and growing to early log phase ($OD_{600} = 0.4 - 0.5$) at 37°C with shaking. RNA was extracted from three biological replicates for both log and stationary phase samples using 7×10^8 cells with the RNeasy Mini kit (Qiagen, Singapore). RNA quality was assessed using the Agilent RNA 6000 pico kit on an Agilent 2100 Bioanalyzer (Agilent Technologies, USA) and only samples with a RNA integrity number (RIN) above 7 were used for library preparation. Ribosomal RNA (rRNA) depletion was performed using the Ribo-Zero rRNA removal kit (Epicenter, USA) according to the manufacturer's recommended protocol and the quality of purified mRNA was checked using the Agilent RNA 6000 pico kit. Libraries were generated using the ScriptSeq v2 RNA-Seq library preparation kit (Epicenter, USA) according to the manufacturer's recommended protocol. Each uniquely indexed strand specific library was assessed for library size and amount using the Agilent DNA 1000 kit (Agilent Technologies, USA). After normalization and pooling, samples were sequenced on either Illumina HiSeq or NextSeq series sequencers with paired end reads, either 2 x 151 bp or 2 x 76 bp (Illumina, USA).

Raw sequencing reads were mapped to their respective reference genomes, for *E. coli* UTI89 (NCBI accession number NC_007946), *E. coli* K12 substr. MG1655 (NCBI accession number NC_000913.3) and *E. coli* CFT073 (NCBI accession number AE014075.1) using BWA-MEM (version 0.7.10) with default parameters [151]. HTseq was used to quantify sequencing reads mapping to predicted open reading frames (ORFs) [152]. Ribosomal RNA (rRNA) and Transfer RNA (tRNA) sequences (based on the Genbank annotation) were filtered out of the data

set. R (version 2.15.1) was used for differential expression analysis, using the Bioconductor package, edgeR [153]. Briefly, samples were normalized by TMM (trimmed median of means), common and tagwise dispersion factors were estimated using a negative binomial model, and then fold change values were calculated on these normalized counts as a proxy for expression. A cutoff of False Discovery Rate (FDR) <0.05 and log fold change >1.5 was used to identify the final set of differentially expressed genes.

2.7 Whole Genome Sequencing (WGS)

Genomic DNA was extracted from log phase bacterial cultures ($OD_{600} = 0.4 - 0.5$) using the QIAamp DNA mini kit (Qiagen, Singapore) and sheared to 350 bp sized fragments using a Covaris S220 focused-ultrasonicator (Covaris, USA) according to manufacturer's instructions. Sheared DNA was used to generate libraries for whole genome sequencing using the TruSeq Nano DNA LT library preparation kit (Illumina, USA). Library size and quantity was checked using the Agilent DNA 1000 kit (Agilent Technologies, USA). Individual libraries were pooled and sequencing was performed as described in Section 2.6 for RNA sequencing. Sequencing reads were mapped to their respective genomes (E. coli UTI89 (NCBI accession number NC_007946), E. coli K12 substr. MG1655 (NCBI accession number NC_000913.3) and E. coli CFT073 (NCBI accession number AE014075.1)) using BWA-MEM (version 0.7.10) with default parameters [151]. Lofreq was used to do local realignment, insertion of indel qualities, and insertion of alignment qualities prior to variant calling of Single Nucleotide Polymorphisms (SNPs) and indels using default parameters [154]. Sanger sequencing was performed to validate variants identified by Lofreq.

2.8 Single Molecule Real Time (SMRT) Sequencing

Genomic DNA was extracted from log phase bacterial cultures and quantified using a Qubit 2.0 Fluorometer (Life Technologies, USA) using the dsDNA HS kit according to the

manufacturer's protocol. 5 µg of DNA was sheared to 10 kbp size using a g-Tube (Covaris, USA) and a SMRTbell library was generated according to manufacturer's instructions using the SMRTbell template prep kit 1.0 (Pacific Biosciences, USA). Library quality and quantity was assessed using the Agilent DNA 12000 kit (Agilent Technologies, USA) and, after loading the library using Mag-Bead bound protocol for large insert libraries, sequenced on the PacBio RS II sequencer (Pacific Biosciences, USA). Sequencing was performed using a single SMRTCell and the P4-C2 enzyme chemistry with 180 mins of data acquisition. Reads were mapped back to the corresponding reference genome and methylated motifs identified using the 'RS_Modification_and_motif_analysis' algorithm in SMRT Analysis suite v2.3 using default parameters. Bases with a coverage of at least 25x and a kinetic score of at least 60 were identified as being methylated by SMRT sequencing.

2.9 Quantitative RT-PCR

Total RNA was extracted from relevant samples using the RNeasy Mini kit (Qiagen, Singapore). 1 µg of RNA was treated with DNase I, RNase-free (Thermo Scientific, USA) at 37°C for 1 hour to remove any residual genomic DNA, followed by RNA purification using RNeasy Mini kit according to the manufacturer's recommended protocol. Next, 500 ng of RNA was used for cDNA synthesis with SuperScript II Reverse Transcriptase and Random hexamers (Invitrogen, USA) according the manufacturer's recommended protocol. Amplified cDNA was diluted 1:4 for all target genes and 1:400 for the *rrsA* internal control. Real-time PCR was performed with the KAPA SYBR FAST qPCR kit (KAPA Biosystems, USA) on a LightCycler 480 instrument (Roche, Singapore) with the following cycle: 95°C for 5 mins for enzyme activation, followed by 40 cycles of 95°C for 30 seconds and 60°C for 30 seconds. Target specific primers for qRT-PCR were designed using the IDT PrimerQuest tool (IDT, Singapore) and are listed in Table 13. Relative fold change for target genes was calculated by the $\Delta\Delta\text{CT}$ method utilizing the 16S ribosomal gene *rrsA*

as the internal control. No reverse transcriptase controls and negative controls were included in each run.

2.10 Motility Assay

Motility assays were performed as described previously with slight modifications [29]. Strains were grown up to log phase ($OD_{600} = 0.4 - 0.5$) by sub-culturing overnight bacterial culture 1:100 in LB at 37°C and normalized to $OD_{600} = 0.4$ across all strains with sterile PBS. 0.25% LB agar plates were prepared and stabbed with each strain using sterile toothpicks. The soft agar plate was then incubated at 37°C for 7 – 8 hours depending on the strain. Motility was calculated by measuring the diameter of bacterial motility, normalizing the diameter measured for each test strain to its corresponding wild type and expressing motility as a percentage of wildtype motility.

2.11 Biofilm Assay

A 96-well Crystal Violet staining protocol for biofilm quantification was performed [155]. Briefly, an inoculum was prepared by growing bacterial strains up to log phase and normalizing to $OD_{600} = 0.4$. 96 well clear flat bottom poly vinyl chloride (PVC) plates (Costar, Singapore) were seeded with 200 μ l of sterile media (LB or YESCA) and 5 μ l of normalized inoculum. PVC plates were incubated for the desired duration at 26°C or 37°C in a moistened chamber. Plates were washed once with water and stained with Crystal Violet for 30 mins. Excess stain was removed by washing thrice with water and finally the stain was dissolved with 50% ethanol, with care taken to avoid disturbing the biofilm. The amount of biofilm was quantified by measuring OD at 590nm using a Sunrise 96 well microplate absorbance reader (Tecan, Switzerland). The biofilm produced by each test strain was normalized to that of its corresponding wild type strain and expressed as a percentage of the wild type biofilm production.

2.12 Growth Curves

Growth curves for bacterial strains were determined using the Bioscreen C instrument (Bioscreen, Finland). Briefly, bacterial strains were inoculated from single colonies into LB media and allowed to grow overnight at 37°C. Cultures were then diluted 1:100 in fresh LB and allowed to grow till mid-log phase ($OD_{600} = 0.5$). All strains were then normalised to $OD_{600} = 0.4$ using sterile PBS, 5 μ l of this normalised culture was inoculated into 145 μ l of the desired growth media in triplicates. LB (rich) medium and M9 (minimal) medium were used at 37°C and measurements were taken every 15 mins for 24 hours.

2.13 Phenotype Microarray

Phenotype microarrays are a high throughput screen utilizing specialized 96 well plates containing different sets of conditions (including various carbon, nitrogen and phosphorous sources, pH, osmolytes and chemical inhibitors) to rapidly test for gene function amongst other applications [156]. Phenotype Microarray (PM) assays were performed using PM plates 1 to 20 by the PM services group at Biolog, USA using their standard protocol for *E. coli*. Briefly, PM plates contain: PM1-2 Carbon sources, PM3 Nitrogen sources, PM4 Phosphorous and Sulphur sources, PM5 Nutrient supplements, PM6-8 Peptide nitrogen sources, PM9 Osmolytes, PM10 pH and PM11-20 Inhibitors for different metabolic pathways. Each strain is normalized and inoculated into all 20 PM plates with supplementation of Niacin, since UTI89 is auxotrophic for niacin. Cellular respiration causes reduction of a tetrazolium dye which results in a colour change measured colorimetrically by the OmniLog instrument. Plates are incubated at 37°C for 24 hours and the entire panel repeated twice. Output consists of growth curves for both wild type and test strains in each well of the twenty 96 well PM plates. These are analysed according to the recommendations of Biolog. Briefly, under each condition the two curves are superimposed, effectively normalising the test strain versus the wild type so as to identify any gained/lost phenotypes. The area under the curve for both wild type and test strain growth curves are compared for each well and the average

height difference in wells which show a difference in both replicates was calculated. The average height difference is an arbitrary quality score that is positive for an acquired phenotype or negative for a lost phenotype in the test strain. A static threshold (recommended by Biolog) was set at 150 above which phenotypic differences are considered significant; further emphasis is laid on phenotypes which show the largest magnitude, show a dose dependent response, or belong to similar metabolic pathways. Identified putative phenotypic differences are subsequently validated independently.

3. RESULTS

Elucidating the role of Type I restriction modification system mediated methylation in regulating Uropathogenic *E. coli* virulence and physiology

3.1 Single Molecule Real Time (SMRT) sequencing identifies a novel methylated motif in UTI89

To gain insight into the role of epigenetic DNA modification in Uropathogenic *Escherichia coli* (UPEC) virulence, we sequenced the cystitis UPEC isolate UTI89 using SMRT sequencing [Performed by Dr. Jacqueline Chee]. SMRT sequencing of unamplified log phase UTI89 genomic DNA allowed detection of 46,407 modified bases in the genome and 3 sequence motifs to which they belong (Table 1). The methylated palindromic motifs 5'-GATC-3' and 5'-CCWGG-3' belong to the previously identified orphan methyltransferases DNA adenine methyltransferase (*dam*) and DNA cytosine methyltransferase (*dcm*), respectively [138]. The third methylated motif, 5'-CCA(N₇)CTTC-3', is bipartite with a 7 bp degenerate sequence between the two specific DNA targets and a methylated adenine on both strands. This recognition sequence is characteristic of methyltransferases belonging to Type I RMSs and represents a novel Type I methylation motif.

Motif	Methylation	Total # of motifs in genome	# of motifs detected	% of motifs detected	Mean motif coverage
<u>G</u>A<u>T</u>C <u>C</u>T<u>A</u>G	6mA	41190	39916	96.9%	84.8X
<u>C</u>C<u>W</u>G<u>G</u> <u>G</u>G<u>W</u>C<u>C</u>	5mC	26482	487	1.8%	95.1X
<u>C</u>C<u>A</u>(N₇)<u>C</u>T<u>T</u>C <u>G</u>G<u>T</u>(N₇)<u>G</u>A<u>A</u>G	6mA	754	737	97.75%	88.3X
		754	749	99.3%	84.6X

Table 1: UTI89 methylated motifs. SMRT sequencing identified 3 methylated motifs in the UTI89 genome. *dam* (5'-GATC-3), *dcm* (5'-CCWGG-3') and a novel Type I RMS motif (5'-CCA(N₇)CTTC-3'). Methylated positions on each strand are underlined and in bold

The UTI89 genome contains only one putative Type I RMS, encoded by the host specificity determinant (*hsd*) locus, consisting of 3 genes. *hsdS* encodes a specificity determinant DNA binding protein, *hsdM* a methyltransferase and *hsdR* an endonuclease. The majority (97 – 99%) of the 754

recognition sites for this Type I RMS appear to be methylated, typical of functional RMSs. The low percentage of methylated *dcm* motifs (1.8%), can be attributed to the weak and dispersed kinetic signal emanating from 5-methylcytosine (5mC) in SMRT sequencing [79], the better detection of which requires conversion of 5mC to 5-hydroxymethylcytosine (5hmC) before sequencing [157]. However, since *dam* [91] and *dcm* [74, 116, 117] represent previously well characterized orphan methyltransferases, we decided to focus our study on characterizing the novel Type I RMS found in UTI89 and exhaustively identifying the functional consequences of this methylation on bacterial physiology and virulence.

3.2 Characterization of Type I Restriction Modification Systems

3.2.1 Novel UTI89 Type I methylation is mediated by a functional Restriction

Modification System (RMS)

To assign the novel Type I methylation motif (5'-CCA(N₇)CTTC-3') to the UTI89 *hsdSMR* Type I RMS and identify if both methylation and restriction functions are intact, we optimised a plasmid transformation based Restriction Modification (RM) assay. As depicted in Figure 3, the RM assay employs plasmids with either 0 (p0), 1 (p1) or 2 (p2) copies of the Type I RMS recognition site in question. These plasmids are extracted from a methylation capable or isogenic methylation incapable strain to obtain methylated or unmethylated versions, respectively, of the same plasmid. These plasmids are then transformed into the test strain bearing the putative Type I RMS (Figure 3.A and 3.B). The test strain transformed with plasmids bearing 0, 1 or 2 copies of the methylated recognition sites should show no difference in the efficiency of transformation (EOT) (Figure 3.A). However, if the same test strain bears a functional Type I RMS and is transformed with unmethylated preparations of the same plasmids, we expect to observe a decline in EOT with an increasing number of recognition sites (Figure 3.B). Finally, an otherwise isogenic test strain carrying a deletion of the Type I RMS transformed with the same set of methylated and unmethylated plasmids, should show no difference in EOT since the plasmid borne recognition site

cannot be recognized irrespective of methylation status (Figure 3.C). Greater the number of unmethylated recognition sites, lower is the expected EOT, since Type I RMSs translocate DNA between two adjacent recognition sites and cleave at a random position between them when translocation is blocked [55].

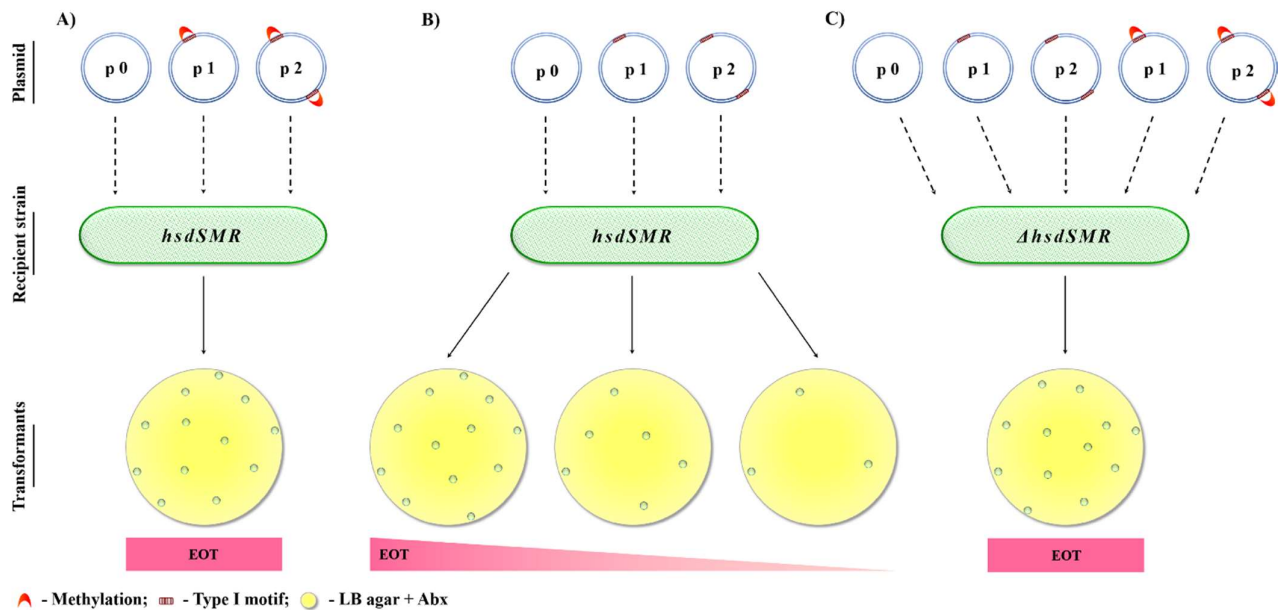


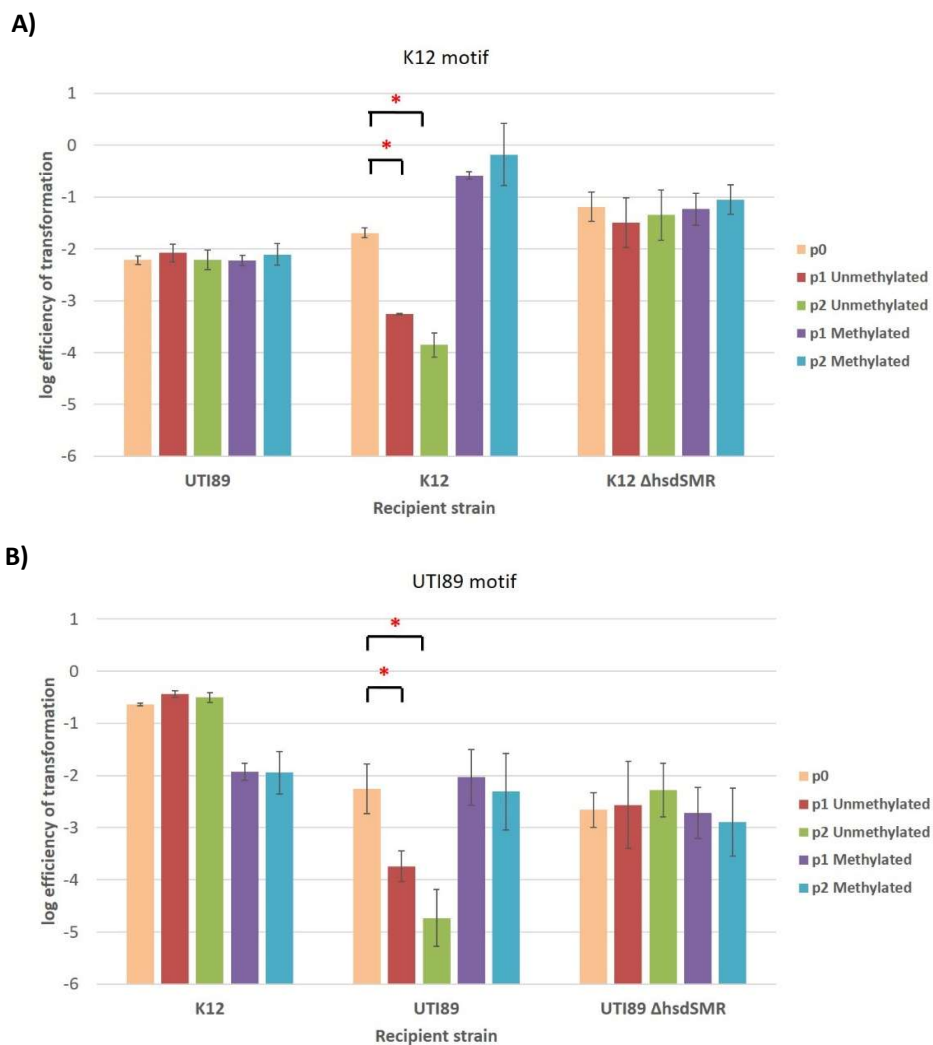
Figure 3: Principle of the Restriction Modification Assay. Plasmids bearing 0, 1 or 2 copies of the Type I RMS recognition site were transformed into a recipient strain with the corresponding Type I RMS and the number of transformants obtained was used to calculate the Efficiency of Transformation. Experiment was performed in a restriction-modification competent strain (*hsdSMR*) with methylated **A**) and unmethylated plasmids **B**). And repeated again with a restriction-modification incapable isogenic deletion strain (Δ *hsdSMR*) **C**). p=plasmid. EOT=Efficiency of Transformation. *hsdSMR*= host specificity determinant genes of Type I RMSs.

In order to test whether the RM assay works as designed, it was first tested using the K12 Type I RMS. The *E. coli* K12 Type I RMS with recognition sequence 5'-AAC(N₆)GTGC -3' has been

the workhorse of researchers studying RMSs in general and Type I systems in particular [54]. UTI89, K12 and K12 *AhdsSMR* (KSM2-102-7) were transformed with unmethylated and methylated variants of the plasmids p0 (pACYC184), p1 (pKSM3-42-1) and p2 (pKSM3-45-1), bearing the known K12 recognition sequence (Figure 4.A). Methylated and unmethylated versions of the plasmids were obtained by propagation in K12 and UTI89 respectively. Plasmids p1 and p2 show a 2 log drop in EOT only when unmethylated and transformed into K12, as shown in Figure 4.A (middle red and green bars). When the same plasmids are transformed into K12 *AhdsSMR* (right bars) or UTI89 (left bars), both lacking a Type I RMS capable of recognising the unmethylated plasmid recognition sites, the EOT is the same as the control (orange in right and left sets of bars). Based on the inverse correlation between the EOT and the number of unmethylated sites, we conclude that our RM assay performs as expected to properly detect both functions. The efficiency of transformation for unmethylated plasmids p1 and p2 transformed into K12 as shown in Figure 4.A (middle purple and blue bars) is higher than the EOT for p0 (middle orange bar). This is due to the fact that plasmid p0 was extracted from UTI89, while unmethylated p1 and p2 were extracted from K12 *AhdsSMR*. This observed difference in EOT is independent of the RMS, but rather due to the difference in plasmid quality when extracted from a cloning strain such as K12 or its mutants versus from a clinical isolate such as UTI89.

We then applied the RM assay to test the putative UTI89 Type I RMS (Figure 4.B). Methylated and unmethylated variants of plasmids p0 (pACYC184), p1 (pKSM3-16-1) and p2 (pKSM3-17-1), bearing the novel UTI89 recognition sequence were obtained by propagation in UTI89 and K12 respectively. Each plasmid was transformed into K12, UTI89 and UTI89 *AhdsSMR* (KSM2-102-4). EOT shows a progressive 2 log decrease from p0 to p2, as the number of unmethylated UTI89 Type I motifs increases from 0 to 2 (Figure 4.B, middle red and green bars). The EOT for the same transformations with methylated plasmids remains unchanged (middle purple and blue bars). Finally, when transformed into K12 and UTI89 *AhdsSMR*, the methylation status of the UTI89 Type I motif bearing plasmids has no effect on the EOT (Figure 4.B, left and right bars). Thus, the UTI89 methylation motif 5'-CCA(N₇)CTTC-3' is successfully assigned to

the *hsdSMR* genes, encoding a fully functional Type I RMS capable of methylating self DNA and restricting unmethylated non-self DNA. Again, the EOT for unmethylated plasmids p0, p1 and p2 extracted from K12 show a higher baseline EOT compared to methylated versions of the same plasmids extracted from UTI89 (Figure 4.B, left set of bars). It is important to note this repeated and inherent difference in baseline EOT when isolating identical plasmids from different strains, most obviously for clinical versus cloning host strains.



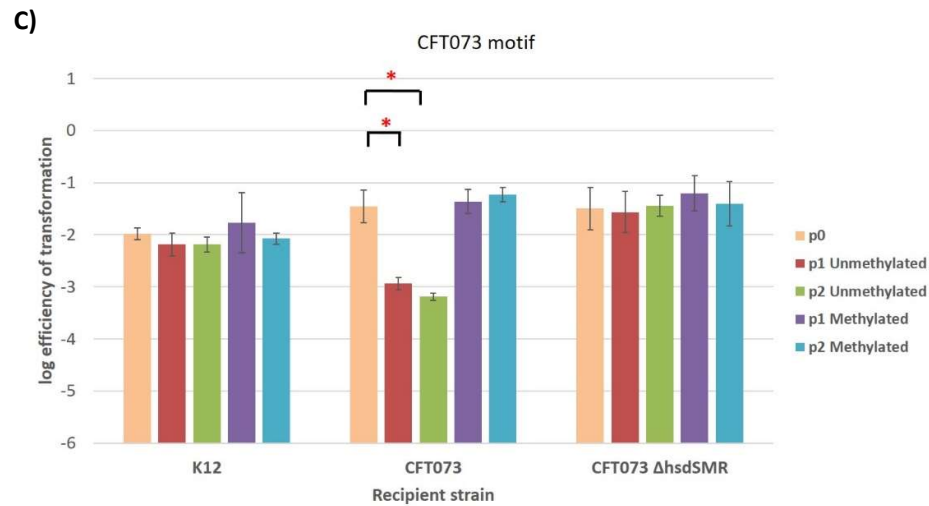


Figure 4: Restriction Modification assay for Type I Restriction Modification Systems.

Restriction Modification assay for **A)** K12, **B)** UTI89 and **C)** CFT073 Type I RMSs. The Type I motif present on the test plasmids is indicated above each figure. Log efficiency of transformation (EOT) for plasmids bearing 0, 1 or 2 copies of the recognition site when transformed into different test strains is plotted. EOT is the number of colonies obtained on selective vs. non-selective plates, per unit amount of DNA. Data represents mean and standard deviation from 3 independent biological replicates. * $p < 0.05$ by Student's T-test.

UPEC isolate CFT073 was isolated from the blood of a patient suffering from pyelonephritis [34] and represents another widely used UPEC model strain. CFT073 also possesses a Type I RMS gene cluster like K12 and UTI89, with yet another distinct methylation motif, 5'-GAG (N₇)GTCA-3' [138]. However, this motif is a prediction based on homology and the functionality of this system has not been experimentally verified. Plasmids p0 (pACYC184), p1 (pKSM4-95-4) and p2 (pKSM4-99-1) bearing the methylated and unmethylated CFT073 motifs were obtained by extracting them from CFT073 and CFT073 Δ hsdSMR (KSM7-15-1) respectively. Each plasmid was transformed into K12, CFT073 and CFT073 Δ hsdSMR to measure the EOT (Figure 4.C).

Transformation of unmethylated plasmids p1 and p2 results in a 2 log decrease in EOT with respect to p0 (Figure 4.C, middle red and green bars). No change in EOT between p0, p1, and p2 is seen, regardless of methylation status, when transformed into K12 and CFT073 Δ *hsdSMR*. Thus, CFT073 also has a functionally intact Type I RMS with the recognition sequence 5'-GAG (N₇)GTCA -3'. To control for the host strain dependent differences in the baseline EOT observed in Figures 4.A and 4.B, all plasmids used in Figure 4.C were extracted from clinical strain CFT073 or its isogenic methylation deletion mutant CFT073 Δ *hsdSMR*. As expected, Figure 4.C does not show any differences in the baseline EOT for any recipient strain.

3.2.2 Generating UTI89 strains bearing diverse *E. coli* Type I methylation motifs.

Type I RMSs have been classified into five families, designated A-E, based on genetic complementation of sub-units, DNA hybridization, serology using HsdM and HsdR, and most recently based on sequence homology [51, 158]. Systems belonging to the same family have >80% HsdM and HsdR amino acid identity and should be able to retain function after swapping individual subunits. Different Type I families have only approximately 20 – 30 % identity and cannot recruit sub-units belonging to another family [158]. Sequence comparisons between the K12 and UTI89 HsdM/R sequences reveal 99% amino acid identity for both proteins, thus both systems can be classified as Type IA RMSs. The UTI89 and CFT073 Type I systems, however, possess only 22.3% (HsdM) and 13.85% (HsdR) identity, indicating that they are from different families. Therefore, to introduce K12 methylation into UTI89, we exchanged the UTI89 *hsdS* with the K12 allele. To introduce CFT073 methylation into UTI89 we exchanged the entire *hsdSMR* operon with the CFT073 *hsdSMR* genes.

UTI89 *hsdS*^{K12} (KSM6-26-1), **UTI89 *hsdSMR*^{CFT073}** (KSM6-80-15) and **UTI89 *hsdS*^{UTI89}** (KSM3-95-4) were generated to introduce K12, CFT073 and UTI89 Type I methylation, respectively, in UTI89. **UTI89 *hsdS*^{K12}**, **UTI89 *hsdSMR*^{CFT073}** and **UTI89 *hsdS*^{UTI89}** shall henceforth be referred to as **UTI89 SWAP K12**, **UTI89 SWAP CFT073** and **UTI89 SWAP UTI89** respectively. The term ‘swap’ used here refers to the exchange of the wild type UTI89 *hsdS*

allele or entire *hsdSMR* Type I RMS for another. UTI89 SWAP UTI89 functions as a control as it has been subjected to the same recombineering as the other swap mutants, but the final result is a revertant strain with the native *hsdS* allele in the original wild type genomic context. Whole Genome Sequencing (WGS) was performed on all of these engineered mutant strains to confirm the expected genetic manipulations at the *hsd* locus and to verify that no off target mutations were inadvertently introduced during cloning.

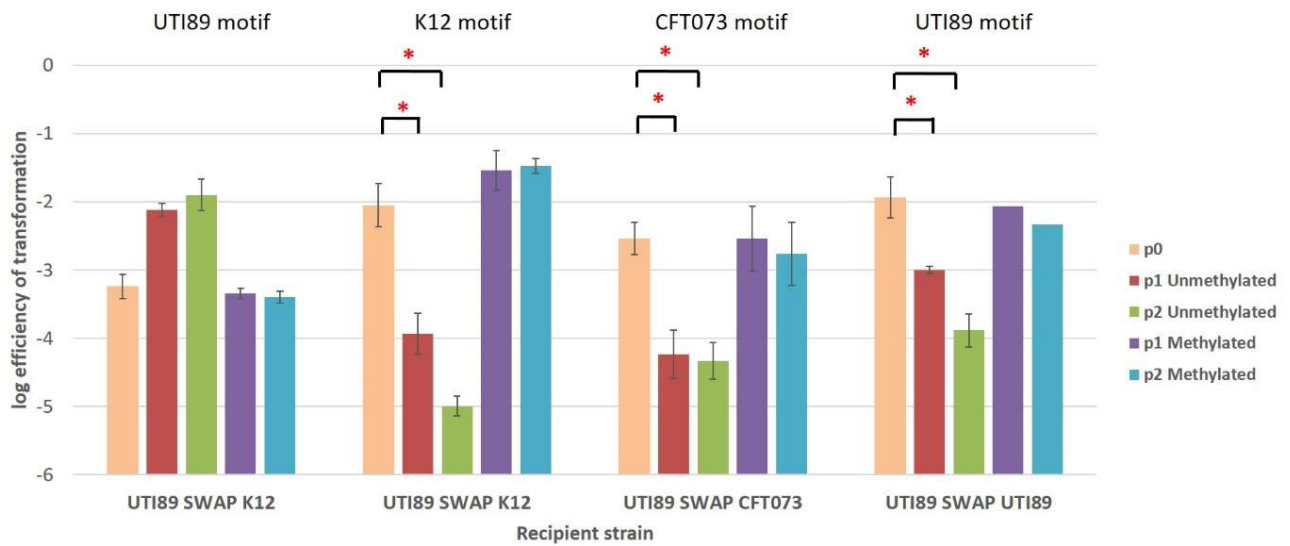


Figure 5: Restriction Modification Assay for UTI89 SWAP strains. Restriction Modification assay for UTI89 SWAP K12, UTI89 SWAP CFT073 and UTI89 SWAP UTI89 strains. Type I motif present on test plasmids is indicated above bars. Efficiency of transformation is number of colonies obtained on selective vs. non-selective plates, per unit amount of DNA. Data represents mean and standard deviation from 3 independent biological replicates. * $p < 0.05$ by Student's T-test.

When transformed with unmethylated and methylated plasmids bearing the UTI89 Type I motif, the UTI89 SWAP K12 strain no longer displays a restriction mediated drop in EOT (Figure 5, first set of bars). However, if transformed with unmethylated K12 Type I motif bearing plasmids, UTI89 SWAP K12 shows a progressive 2 log drop in EOT typical of a functional Type I RMS

(Figure 5, second set of bars). Similarly, the UTI89 SWAP CFT073 strain shows a drop in the EOT when transformed with plasmids bearing unmethylated CFT073 Type I motif and the UTI89 SWAP UTI89 revertant shows a decrease in the EOT for plasmids with unmethylated UTI89 Type I sites (Figure 5, third and fourth set of bars, respectively). Methylated K12, CFT073 and UTI89 site bearing plasmids were obtained by passaging the plasmids in the respective SWAP strain. The EOT for these methylated plasmids does not change (Figure 5, purple and blue in each set of bars), suggesting that both methylation and restriction functions of these Type I RMSs are intact, but the methylome is different from wild type UTI89. Thus, UTI89 strains were successfully generated to express fully functional K12, CFT073 and a UTI89 revertant Type I RMS, with both methylation and restriction functions still intact.

3.3 Effect of Type I methylation on UTI89 virulence *in vitro*

Before proceeding to the established *in vivo* transurethral murine model for ascending UTI, we decided to first utilise *in vitro* phenotypic assays for known UPEC virulence factors. Although these *in vitro* assays are by no means exhaustive and one could envision additional assays from the perspective of UTIs caused by UPEC, these assays represent powerful *in vitro* proxies for *in vivo* virulence. Growth kinetics, Type I pili mediated hemagglutination, motility and biofilm formation are some of the assays used here to gauge UPEC virulence regulated by Type I RMS methylation prior to *in vivo* tests.

3.3.1 Type I methylation does not affect UTI89 motility

Motility and chemotaxis are important bacterial processes required for immune evasion as well as nutrient acquisition and colonisation both *in vivo* and *in vitro* [159]. UPEC flagellar expression occurs only transiently *in vivo* and non-motile strains are primarily defective in kidney colonisation at specific late time points [29, 30]. UPEC motility and chemotaxis do contribute to fitness *in vivo*, but this contribution is only transient and is not absolutely required for virulence [31]. Neither removal of native Type I methylation, nor replacement with K12 or CFT073

methylation resulted in any alterations to UTI89 swimming motility (Figure 6). UTI89 $\Delta hsdS$ is the parental strain to UTI89 SWAP K12, SWAP CFT073 and SWAP UTI89 revertant strains and is therefore included as a control. UTI89 $\Delta fliC$ lacks flagellin and is completely non-motile, as expected.

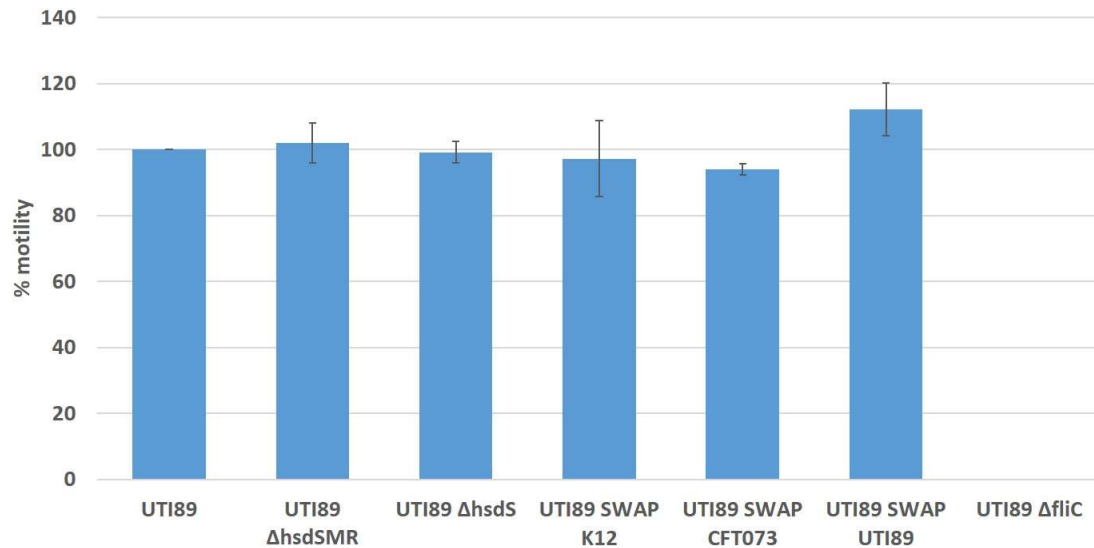


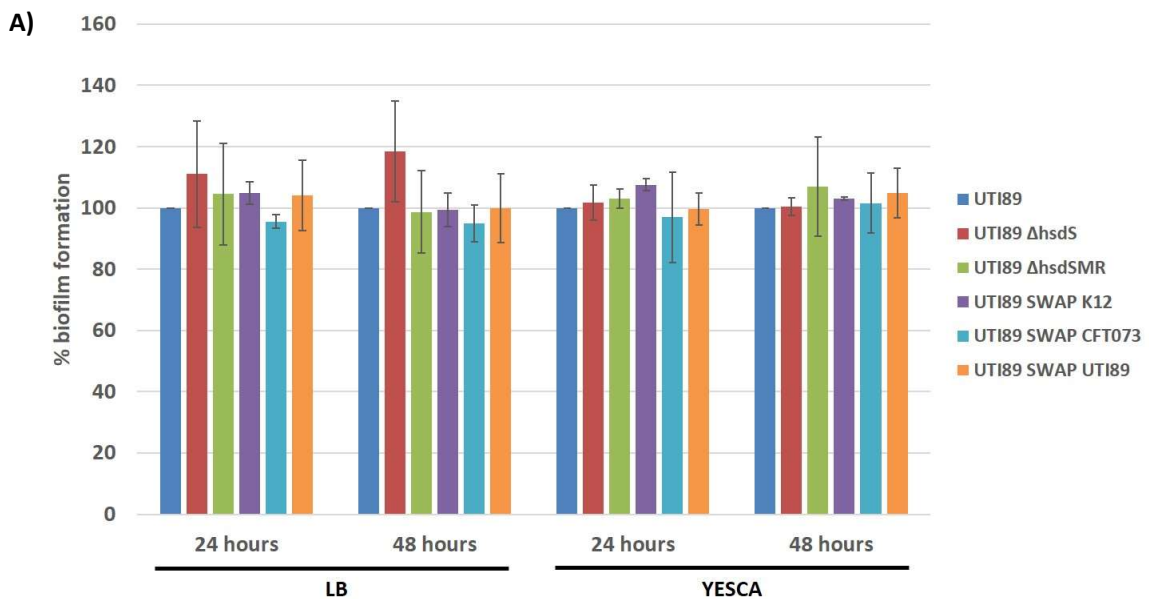
Figure 6: Motility of UTI89 mutants. Motility of UTI89 Type I methylation mutants relative to wild type using the soft agar motility assay. Data represents mean and standard deviation from biological triplicates. UTI89 $\Delta fliC$ is a non-motile control.

3.3.2 Type I methylation does not affect biofilm formation in UTI89

Biofilms represent complex, differentiated multicellular bacterial communities consisting of bacteria along with their extracellular products. Biofilm formation in UPEC confers significant advantages such as immune evasion and phenotypic resistance to antibiotics. Biofilm like intracellular bacterial communities (IBCs) are an important step in acute UTI, and catheter associated biofilms are an important source of inpatient nosocomial UTIs [2, 160]. *In vitro* defects in biofilm formation also manifest as *in vivo* attenuation, especially in reduced IBC formation

[155]. Therefore, we tested for an effect of Type I methylation on biofilm formation by UTI89 using a 96 well Crystal Violet biofilm assay.

Biofilms formed in LB media are multifactorial, but primarily believed to be Type I pili dependent; while biofilms in YESCA media are curli and cellulose dependent. However, depending on the attachment substrate, temperature and media conditions UPEC mediated biofilms are multifactorial phenotypes. To check biofilm formation mediated by UTI89 and its mutants, we tested LB and YESCA media at 26°C (ambient temperature) and 37°C (body temperature) over 48 hours (Figure 7). No significant difference was observed when comparing wild type UTI89 (dark blue bar) biofilms to the different mutants in LB or YESCA at 26°C (Figure 7.A). Biofilm formation at 37°C also did not vary significantly from wild type (Figure 7.B). Significant difference in biofilm formation was observed for the revertant strain in LB at 37°C (Figure 7.B, orange bars), and $\Delta hsdSMR$ and SWAP K12 strains in YESCA at 37°C after 48 hours (Figure 7.B, green and purple bars). UTI89 $\Delta hsdS$ serves as a control as it is the parental strain to all of the SWAP strains, and as expected phenotypically behaves similar to UTI89 $\Delta hsdSMR$.



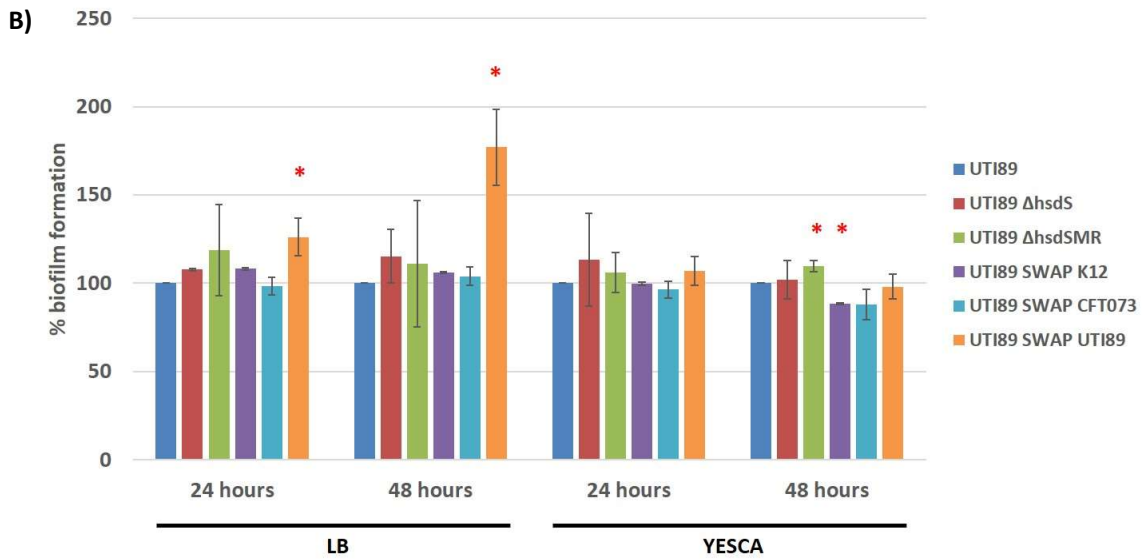


Figure 7: Biofilm formation by UTI89 mutants. Biofilm formation by UTI89 mutants relative to wild type both in LB and YESCA media at **A)** 26°C and **B)** 37°C. Biofilms formed in 96 well PVC plates, stained using Crystal Violet and quantified by measuring absorbance at 590nm. Data represents mean and standard deviation from three biological replicates. Significance measured compared to wild type UTI89 under matching conditions by student's T-test * $p < 0.05$

Thus, two important *in vitro* surrogates for fitness, motility and biofilm formation, reveal that removal or replacement of Type I methylation in UTI89 has no phenotypic consequences under the diverse set of conditions checked.

3.3.3 Type I RMS mediated methylation does not affect UTI89 growth

To test for a general effect of Type I methylation on growth, we tested growth in both rich (LB) and minimal (M9) media (Figure 8). UTI89 lacking native Type I methylation or expressing foreign Type I methylation from other *E. coli* strains does not show any significant alterations in growth in LB (Figure 8.A) or M9 (Figure 8.B) media over 24 hours when compared to wild type

UTI89. Although the final OD achieved by Type I methylation mutants is different in M9, these differences are not statistically significant over biological triplicates.

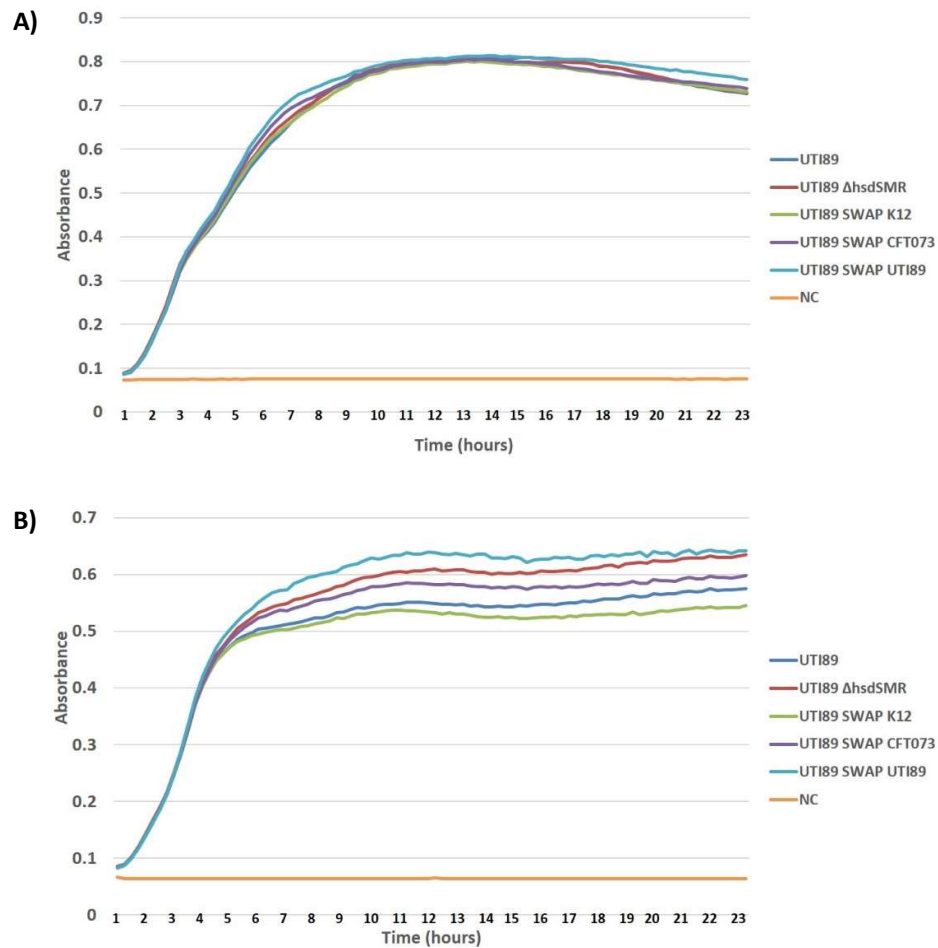


Figure 8: Growth curves for UTI89 methylation mutants. Growth curves were generated for UTI89 and its Type I methylation mutants in **A)** LB and **B)** M9 media. Experiment was performed at 37°C and curves represent the mean values from three biological replicates. Negative control (NC) contains media only. Significance measured by Student's T-test, with $p < 0.05$ considered significant.

Next, we tested the effect of native Type I methylation in other *E. coli*, namely K12 and CFT073. CFT073 growth in LB (Figure 9.A, green and purple curves) and M9 (Figure 9.B, green

and purple curves) remains unaffected. However, K12 $\Delta hsdSMR$ shows a significant growth defect compared to wild type K12 in LB, from 2 hours to 18 hours (Figure 9.A, blue and red curves). K12 and K12 $\Delta hsdSMR$ growth in M9 is not significant different (Figure 9.B, blue and red curves).

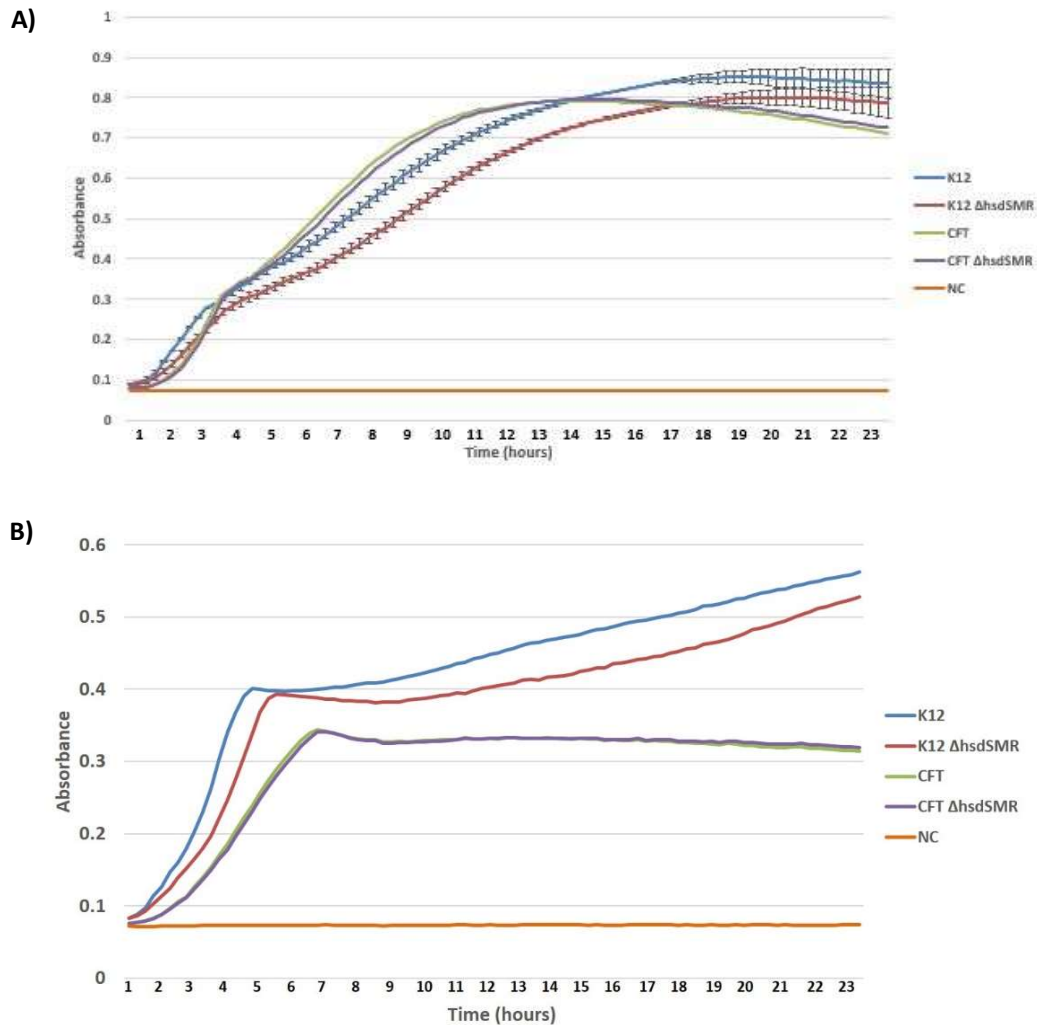


Figure 9: Growth curves for K12 and CFT073 methylation mutants. Growth curves for K12 and CFT073 along with their corresponding Type I RMS deletion mutants in **A)** LB and **B)** M9. Negative control (NC) contains media only. Means from three biological replicates are plotted and significance calculated using student's T-test $p < 0.05$. For K12 and K12 $\Delta hsdSMR$ in **A)** LB, error bars represent standard deviation from three biological replicates.

3.4 Effect of Type I methylation on UTI89 virulence *in vivo*

3.4.1 Loss of native UTI89 methylation does not affect virulence *in vivo*

In order to perform co-infections with UTI89 and its methylation mutants, we generated differentially marked strains for UTI89 wild type, UTI89 Δ *hsdSMR* and UTI89 SWAP K12 carrying both *neo* (Kan^R) and *cat* (Chlor^R) resistance genes. Kanamycin and Chloramphenicol resistance cassettes were added at the phage HK022 attachment site [161], a position known to maintain selection cassettes stably without selection pressure and with no measurable effect on virulence [162]. Type I pili, encoded by the *fim* operon, are filamentous surface structures expressed in UTI89 that mediate both attachment and invasion of bladder epithelial cells and, formation of intracellular bacterial communities [163]. To check if Type I methylation has any consequences on the expression of this important virulence factor before murine infection, we performed a Hemagglutination assay using guinea pig RBCs. As shown in Table 2, neither insertion of an antibiotic cassette at *att*_{HK022} or alteration of Type I methylation altered hemagglutination titres.

Strain	Genotype	Hemagglutination titre
UTI89	UTI89 wild type	7
KSM2-102-4	UTI89 <i>hsdSMR::neo</i>	7
KSM3-39-2	UTI89 <i>att</i> _{HK022} :: <i>neo</i>	7
KSM3-39-1	UTI89 <i>att</i> _{HK022} :: <i>cat</i>	7
KSM3-39-6	UTI89 <i>hsdSMR::FRT att</i> _{HK022} :: <i>neo</i>	7
KSM3-39-3	UTI89 <i>hsdSMR::FRT att</i> _{HK022} :: <i>cat</i>	7
KSM6-29-3	UTI89 <i>hsdS</i> ^{K12} <i>att</i> _{HK022} :: <i>neo</i>	7
KSM6-29-1	UTI89 <i>hsdS</i> ^{K12} <i>att</i> _{HK022} :: <i>cat</i>	7

Table 2: HA titres for UTI89 strains. HA titres were performed on 3 independent occasions and prior to each murine infection.

After verifying that there were no differences in growth or HA titres, we proceeded to *in vivo* infection tests. Female C3H/HeN mice were inoculated transurethrally with 2×10^7 cfu of either UTI89 or UTI89 Δ *hsdSMR* according to the well-established ascending urinary tract

infection (UTI) model [149]. After 24 hours, the bacterial load in both bladders and kidneys were determined. There were no statistically significant differences in the median bladder and kidney titres between wild type UTI89 and the Type I methylation deficient mutant (Figure 10.A).

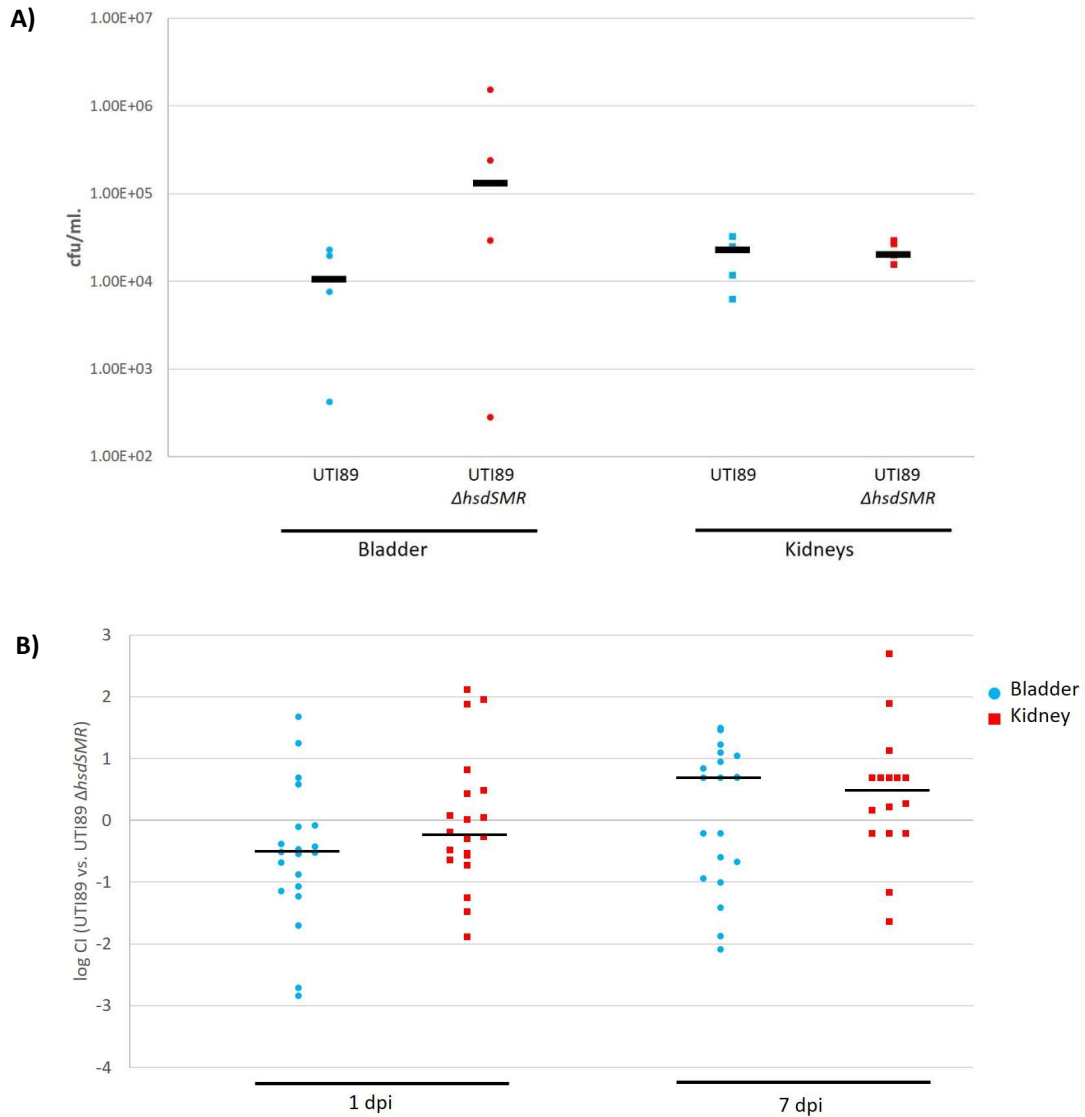


Figure 10: Deletion of the UTI89 Type I RMS has no effect on *in vivo* urinary tract infection.

A) Single infections with wild type UTI89 and UTI89 Δ hsdSMR. Female C3H/HeN mice (n=5 mice/strain) were transurethrally inoculated with the strains indicated on the x-axis. The log₁₀ of the bacterial loads in bladders (left two lanes) and kidneys (right two lanes) at 24 hpi were plotted

on the y-axis. Medians for each group are indicated by a black horizontal line. Mann-Whitney U test with $p < 0.05$, is considered significant. **B)** Competitive infection with wild type UTI89 and UTI89 Δ *hsdSMR* strains. Female C3H/HeN mice (n=20 mice/time point) were inoculated with equal mixtures of UTI89 and UTI89 Δ *hsdSMR* and sacrificed at the time point indicated on the x-axis. The \log_{10} of the competitive index between the two bacterial strains in bladders (blue dots) and kidneys (red dots) are plotted on the y-axis. Medians for each group are indicated by a black horizontal line. Wilcoxon signed rank test with $p < 0.05$, is considered significant.

Co-infection was performed with differentially marked strains, Figure 10.B i.e. UTI89 *att_{HK022}::neo* with UTI89 *hsdSMR::FRT att_{HK022}::cat* and UTI89 *att_{HK022}::cat* with UTI89 *hsdSMR::FRT att_{HK022}::neo*, to eliminate any bias due to selection markers and insertion at HK022 phage attachment site. Since single infection did not reveal any phenotype, infecting simultaneously with wild type and deletion mutant allows greater sensitivity in identifying any subtle phenotype. Moreover, infections were performed for 1 day and 7 days to represent the acute and chronic stages of infection respectively. Log Competitive Indices (CI) did not reveal any significant fitness advantage for either strain in the bladder or kidney, at 1 day or 7 days post infection, Figure 10.B. Thus, irrespective of the duration and site of infection, loss of native Type I methylation does not seem to alter UTI89 virulence *in vivo* at all.

3.4.2 Switching Type I methylation in UTI89 does not affect *in vivo* virulence.

The UTI89 SWAP K12 strain differs from UTI89 in the loss of UTI89 Type I methylation at 754 predicted sites and the gain of K12 Type I methylation at 668 predicted sites in the UTI89 genome. Even though the K12 Type I RMS belongs to the same family as the UTI89 Type I RMS, it is taken from a gut derived lab adapted non-pathogenic strain. Since methylation has the potential to alter gene expression, we hypothesised that this change in Type I methylation specificity could alter the virulence of UTI89. If true, the association of the UTI89 methylation

specificity with higher UTI virulence would suggest a novel potential epigenetic influence on UPEC evolution.

To test this hypothesis for K12 methylation, competitive infections between UTI89 and UTI89 SWAP K12 were performed with differentially marked strains as described in Section 3.4.1 (i.e. UTI89 *att_{HK022}::neo* with UTI89 *hsdS^{K12} att_{HK022}::cat* and UTI89 *att_{HK022}::cat* with UTI89 *hsdS^{K12} att_{HK022}::neo*; Figure 11). Infections were allowed to proceed for 1 day (acute) and 7 days (chronic). Interestingly the log CI was not significantly different from zero at either 1 day or 7 days post infection in either bladders or kidneys, indicating no significant difference between wild type UTI89 and UTI89 SWAP K12 (Figure 11). Therefore, switching the Type I methylation specificity in UTI89 had no detectable effect on *in vivo* virulence.

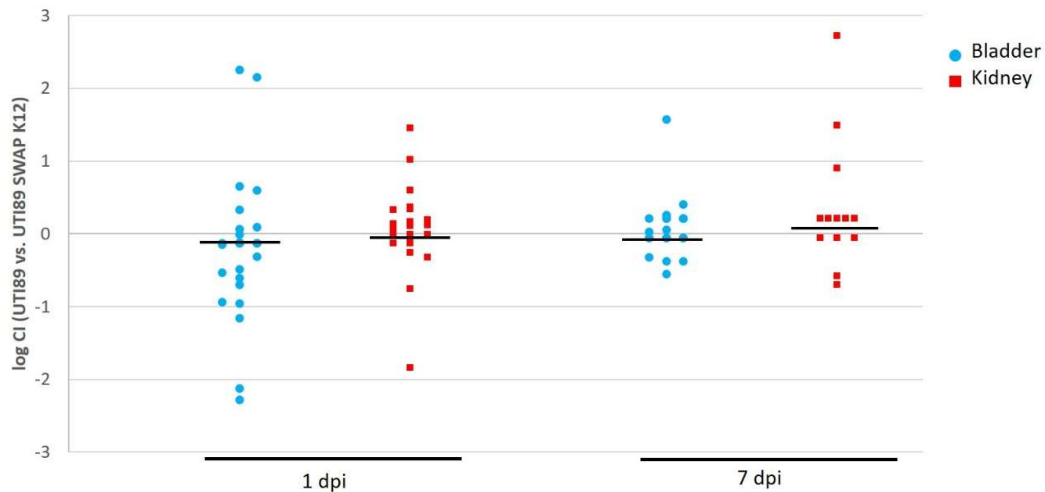


Figure 11. Co-infection with UTI89 SWAP K12 strain. Co-infections with wild type and UTI89 SWAP K12 strains at 1 (n=25 mice) and 7 (n=20 mice) days post infection. Log competitive indices for UTI89 vs. UTI89 SWAP K12 are plotted and median value marked with a horizontal line. Wilcoxon signed rank test used to evaluate if median value is significantly different from 0, with $p < 0.05$.

3.5 Effect of Type I methylation on *Escherichia coli* gene expression

3.5.1 RNA-seq reveals no consequence of altered Type I DNA methylation on UTI89

transcriptome

Loss of UTI89 Type I methylation or replacement with foreign K12 methylation did not affect virulence *in vivo*. Since DNA methylation mediated by orphan methyltransferases [91, 92] and certain phase variable RMSs [122] are capable of altering gene expression in other bacterial pathogens, we questioned whether the loss of native UTI89 Type I methylation or replacement with foreign K12 or CFT073 methylation has any influence on gene expression in UTI89. RNA was extracted from log phase cultures and cultures grown statically for two successive 24 hour incubations, mimicking the stationary phase cells used for murine infections (n=3 independent biological replicates for each). Strand specific RNA-seq was performed and compared to wild type UTI89 to identify differentially expressed genes.

A)

Gene ID	Gene	Annotation	log ₁₀ Fold Change	FDR	qPCR Fold Change
UTI89 vs. UTI89 Δ<i>hsd5MR</i>					
UTI89_C5053	<i>hsdR</i>	Type I RMS Endonuclease	-12.88	1.41E-47	
UTI89_C5051	<i>hsdM</i>	Type I RMS Methyltransferase	-11.71	1.95E-45	
UTI89_C5050	<i>hsdS</i>	Type I RMS Specificity Protein	-12.09	8.47E-44	
UTI89_C5049	<i>yjiW</i>	Endoribonuclease SymE	1.78	4.03E-10	
UTI89_C1127	-	Transposase	2.13	2.23E-09	1.099
UTI89 vs. UTI89 SWAP K12					
UTI89_C5050	<i>hsdS</i>	Type I RMS Specificity Protein	-4.64	0	
UTI89_C3217	-	Hypothetical protein	-1.76	0.043221	
UTI89 vs. UTI89 SWAP CFT073					
UTI89_C5051	<i>hsdM</i>	Type I RMS Methyltransferase	-7.23	1.53E-61	
UTI89_C5050	<i>hsdS</i>	Type I RMS Specificity Protein	-2.75	9.56E-14	
UTI89_C5053	<i>hsdR</i>	Type I RMS Endonuclease	-2.33	1.12E-08	
UTI89 vs. Revertant					
-	-	-	-	-	-

B)

Gene ID	Gene	Annotation	log ₁₀ Fold Change	FDR	qPCR Fold Change
UTI89 vs. UTI89 ΔhdsSMR					
UTI89_C5050	<i>hdsS</i>	Type I RMS Specificity Protein	-8.85	1.49E-34	
UTI89_C5051	<i>hdsM</i>	Type I RMS Methyltransferase	-7.43	1.43E-33	
UTI89_C5053	<i>hdsR</i>	Type I RMS Endonuclease	-7.35	5.96E-24	
UTI89_C1127	-	Transposase	4.29	2.54E-20	0.77
UTI89_C5049	<i>yjiW</i>	endoribonuclease SymE	2.54	7.15E-15	
UTI89_C5054	<i>mrr</i>	Mrr protein	2.43	3.26E-10	
UTI89 vs. UTI89 SWAP K12					
UTI89_C5050	<i>hdsS</i>	Type I RMS Specificity Protein	-5.29	7.32E-15	
UTI89_C2635	-	Hypothetical protein	2.40	4.82E-3	
UTI89 vs. UTI89 SWAP CFT073					
UTI89_C5051	<i>hdsM</i>	Type I RMS Methyltransferase	-8.05	2.47E-35	
UTI89_C5050	<i>hdsS</i>	Type I RMS Specificity Protein	-5.05	1.55E-17	
UTI89_C2635	-	Hypothetical protein	1.77	1.34E-2	
UTI89 vs. Revertant					
-	-	-	-	-	-

Table 3: Differentially expressed genes in UTI89 Type I methylation mutants. A) Log and **B)** Stationary phase cultures of UTI89 methylation mutants were compared to wild type. Genes knocked out or replaced are marked in red and serves as controls. Data is from three biological replicates. Significance is calculated by exact test assuming negative binomial distribution. False Discovery Rate (FDR) <0.05 and log Fold change >1.5 is considered significant.

As expected, the *hdsSMR* genes were all found to be significantly downregulated in the Δ *hdsSMR* mutant compared with the wild type. Two genes, *yjiW* and UTI89_C1127, were differentially expressed (both upregulated) in both log (Table 3.A) and stationary (Table 3.B) phase in UTI89 Δ *hdsSMR*. One gene, *mrr*, was upregulated in stationary phase only. *mrr* (Gene ID UTI89_C5049) and *yjiW* (Gene ID UTI89_C5054) are located on either side of the *hds* locus; the Δ *hdsSMR* mutation results in the introduction of a kanamycin cassette into the *hdsSMR* locus, which may lead to altered transcription of the adjacent genes. Therefore, the putative transposase gene, UTI89_C1127, appears to be the only dysregulated gene. However, subsequent qRT-PCR

assays in independent cultures did not validate the upregulation of this gene (Table 3). Therefore, I concluded that deletion of the *hsdSMR* locus in UTI89 had no effect on the transcription of other genes.

The UTI89 SWAP K12 and UTI89 SWAP CFT073 strains also show the expected downregulation of the UTI89 *hsdS* and *hsdSMR* genes, respectively. Transcription of a single hypothetical protein appears altered for either strain under either condition (Table 3). Finally, as expected, the UTI89 SWAP UTI89 revertant shows no significant changes in gene expression compared with the wild type. Therefore, removal of methylation from 754 predicted UTI89 methylation sites in UTI89 Δ *hsdSMR* and, replacement with methylation of either 668 predicted K12 methylation sites in UTI89 SWAP K12 or 784 predicted CFT073 methylation sites in UTI89 SWAP CFT073 appears to have no effect on UTI89 gene expression. This is in stark contrast to published reports of global gene expression changes due to mutation of some other bacterial DNA methyltransferases [91, 123].

3.5.2 RNA-seq reveals minimal Type I methylation mediated gene expression changes in K12 and CFT073

To confirm the surprising observation that Type I methylation does not affect UTI89 gene expression *in vitro* nor virulence *in vivo*, we extended our RNA-seq analysis to another UPEC strain, CFT073, and the K12 lab strain MG1655. I performed RNA-seq on CFT073 and the CFT073 Δ *hsdSMR* strain in log and stationary phase as described above for UTI89 (Section 3.5.1). Loss of CFT073 Type I methylation had no effect on stationary phase transcription, however three genes were dysregulated in log phase (Table 4). *yjiW* (Gene ID CFT073_c5422), which is adjacent to the *hsd* locus, was upregulated as in UTI89. I again attributed this to the insertion of the antibiotic cassette into the *hsdSMR* locus. *malK* and *lamB* were both downregulated in log phase. Subsequent qRT-PCR assays also failed to validate that these genes were downregulated in CFT073 Δ *hsdSMR* mutant (Table 4). Therefore, similar to UTI89, deletion of Δ *hsdSMR* in CFT073 also has no effect on the expression of any gene.

Gene ID	Gene	Annotation	log ₁₀ Fold Change	FDR	qPCR Fold Change
Log Phase : CFT073 vs. CFT073 ΔhsdSMR					
CFT073_c5424	<i>hsdM</i>	Type I RMS Methyltransferase	-7.37	1.95E-31	
CFT073_c5423	<i>hsdS</i>	Type I RMS Specificity Protein	-5.12	2.64E-26	
CFT073_c5425	<i>hsdR</i>	Type I RMS Endonuclease	-2.41	2.07E-06	
		Maltose/maltodextrin transporter			
CFT073_c5005	<i>malK</i>	ATP-binding protein	-5.07	0.0475	0.71
CFT073_c5006	<i>lamB</i>	Maltoporin	-5.42	0.0475	1.09
CFT073_c5422	<i>yjiW</i>	Endoribonuclease SymE	1.69	0.0475	
Stationary Phase : CFT073 vs. CFT073 ΔhsdSMR					
CFT073_c5423	<i>hsdS</i>	Type I RMS Specificity Protein	-5.15	6.24E-222	
CFT073_c5424	<i>hsdM</i>	Type I RMS Methyltransferase	-10.24	7.29E-216	
CFT073_c5425	<i>hsdR</i>	Type I RMS Endonuclease	-3.70	2.70E-76	

Table 4: Differentially expressed genes in CFT073 Δ hsdSMR. CFT073 Δ hsdSMR genes differentially expressed from triplicate log and stationary phase cultures are listed. Genes marked in red are knocked out and serve as controls. FDR <0.05 and log Fold change >1.5 was considered significant.

E. coli K12 substrain MG1655 is a commonly used lab cloning strain. Type I RMS genes are often deleted to eliminate the restriction barrier to exogenous DNA and allow more efficient cloning in several cloning strains such as DH10 β , DH5 α , TOP10 and BL21 (New England Biolabs). To further test whether Type I methylation affected gene expression, I also performed RNA-seq analysis on triplicate log and stationary phase samples for K12 and K12 Δ hsdSMR. Removal of Type I methylation resulted in 6 genes significantly upregulated in log phase (Table 5). Once again, the genes flanking the selection cassette were upregulated, *mrr* (Gene ID K12_b4351) and *symE* (Gene ID K12_b4347). I again attributed these to insertion of the kanamycin resistance cassette into the *hsdSMR* locus as before for UTI89 and CFT073. Interestingly, all 4 remaining genes belong to the CP4-44 cryptic prophage and this prophage

also has 4 Intragenic (IG) K12 methylation sites (Table 5). I used qRT-PCR to validate that two of these genes, *flu* and *yeeS*, were indeed upregulated in the K12 Δ *hsdSMR* mutant (Table 5). This increased CP4-44 phage transcription could either be due to alleviation of restriction pressure owing to the presence of multiple K12 methylation sites or due to direct methylation mediated regulatory suppression of CP4-44 transcription. Stationary phase transcriptional differences between K12 and K12 Δ *hsdSMR* consists of a single gene, *uhpA* differentially regulated (excluding polar effect *mrr*) (Table 5). *uhpA* is part of a two component regulatory system responsible for hexose phosphate utilization. Coincidentally *uhpA* also possesses an intragenic K12 methylation site, offering a possibility of direct methylation mediated regulation.

Gene ID	Gene	Annotation	log ₁₀ Fold Change	FDR	qPCR Fold Change	No. of IG motifs
Log Phase : K12 vs. K12 Δ<i>hsdSMR</i>						
K12_b4349	<i>hsdM</i>	Type I RMS Methyltransferase	-8.89	6.66E-79		
K12_b4348	<i>hsdS</i>	Type I RMS Specificity Protein	-8.81	1.59E-72		
K12_b4350	<i>hsdR</i>	Type I RMS Endonuclease	-9.70	6.40E-65		
K12_b4435	<i>isrC</i>	novel sRNA CP4-44 putative prophage remnant	3.68	3.21E-21		
K12_b2001	<i>yeeR</i>	CP4-44 prophage predicted membrane protein	2.64	4.63E-17		1
K12_b4351	<i>mrr</i>	methylated adenine and cytosine restriction protein	3.02	3.58E-14		1
K12_b2000	<i>flu</i>	CP4-44 prophage antigen 43 phase-variable biofilm formation autotransporter	2.86	1.87E-08	5.6	3
K12_b4347	<i>symE</i>	Toxic peptide regulated by antisense sRNA symR	2.15	3.40E-07		
K12_b2002	<i>yeeS</i>	CP4-44 prophage predicted DNA repair protein	4.14	0.0178	5.73	
Gene ID	Gene	Annotation	log ₁₀ Fold Change	FDR	qPCR Fold Change	No. of IG motifs
Stationary Phase : K12 vs. K12 Δ<i>hsdSMR</i>						
K12_b4350	<i>hsdR</i>	Type I RMS Endonuclease	-11.32	1.79E-60		
K12_b4348	<i>hsdS</i>	Type I RMS Specificity Protein	-10.12	6.52E-51		
K12_b4349	<i>hsdM</i>	Type I RMS Methyltransferase	-10.02	7.38E-47		
K12_b4351	<i>mrr</i>	Methylated adenine and cytosine restriction protein	2.47	1.25E-13		
K12_b3669	<i>uhpA</i>	DNA-binding response regulator in two-component regulatory system with UhpB	1.51	4.26E-06		1

Table 5. Differentially expressed genes in K12 *ΔhsdSMR*. Log and stationary phase gene expression changes in K12 *ΔhsdSMR* from biological triplicates are listed. Genes marked in red are deleted and hence downregulated. Fold change by qRT-PCR and number of intragenic K12 methylation sites in dysregulated genes are listed. FDR <0.05 and log Fold change >1.5 was considered significant.

Thus, global gene expression profiling of *hsdSMR* mutants of UTI89, CFT073, and K12 under two different growth conditions identified only a single prophage (CP4-44 in K12) and a single gene (*uhpA*) potentially upregulated in K12 that was not attributable to insertion of the selection cassette used to knock out the *hsdSMR* operon. The overall result is strikingly strong that Type I methylation has no effect on expression of any gene in both UPEC strains. For K12 substrain MG1655, there appear to be up to two loci affected, but this again is contrary to expectations based on other published reports of the effect of restriction methylation on global patterns of gene expression in other bacteria.

3.6 Phenotypic consequences of altered Type I DNA methylation in *E. coli*

3.6.1 High throughput phenotypic screen does not identify any differences owing to Type I DNA methylation

Phenotype microarray (PM) is a high throughput screen which allows a comprehensive analysis of cellular phenotypic profiles. Assigning genotype-phenotype associations, i.e. identifying the function of an unknown gene by simultaneously screening almost 2000 phenotypes, which is the primary function of PM utilised here. Using 96 well PM plates (01-20), with different conditions such as Carbon sources (PM plates 1 & 2), Nitrogen sources (PM plates 3, 6, 7 & 8), Phosphorus and Sulphur sources (PM plate 4), nutrient supplements (PM plate 5), pH (PM plate 9), osmolytes (PM plate 10) and chemical inhibitors (PM plates 11 to 20), it is

possible to rapidly and quantitatively evaluate a diverse set of phenotypes [156, 164]. Cellular respiration is used as a proxy for growth and growth curves for a wild type reference are compared to mutant test strains. If there is no phenotypic difference, the reference and test curves overlap (yellow area under curve). If the test strain outgrows the reference strain, the curve shows some green area, and this is interpreted as a gained phenotype/resistance in the test strain. Similarly, red area under the curve indicates a lost phenotype/resistance in the test strain. An average height difference is calculated for each pair of curves; as recommended by the manufacturer (Biolog), an average height difference above 150, seen across multiple concentrations of the same compound is considered significant.

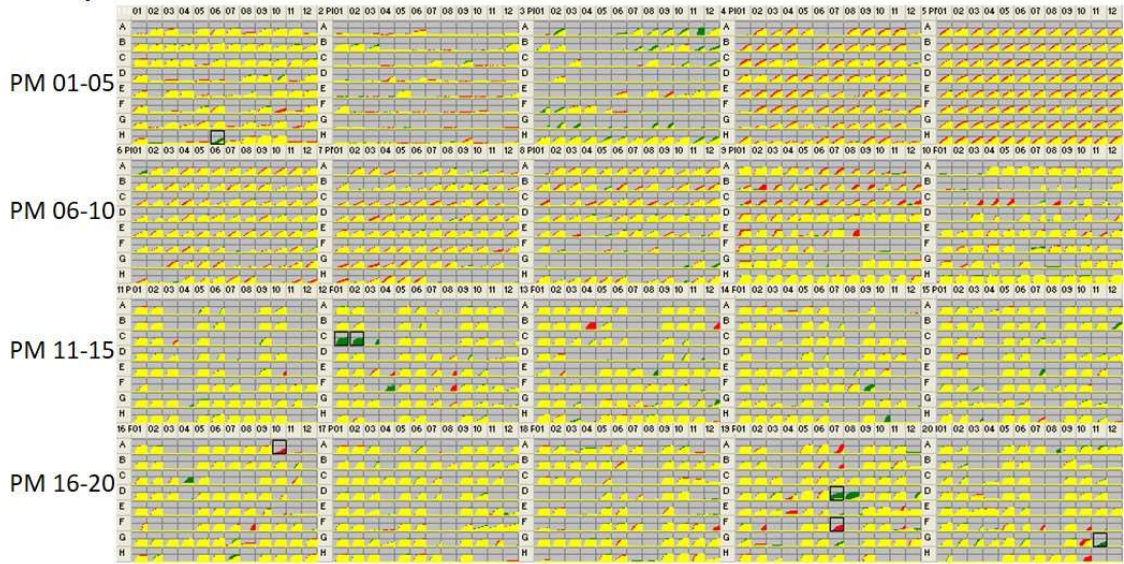
We performed PM to compare UTI89 wild type with UTI89 *ΔhdsSMR* and UTI89 SWAP UTI89 revertant. The entire panel of 20 PM plates were performed twice for each strain and the resulting PM plate data from one of the two runs is shown in Figure 12.A and B, with wells consistently different in both runs boxed. Most of the wells in all 20 PM plates show yellow curves signifying negligible phenotypic differences between UTI89 bearing or lacking its native methylation. The most prominent phenotype is the gain of resistance to Paromomycin by UTI89 *ΔhdsSMR* (Figure 12.A. PM 12 and Table 6.A). Paramomycin is an aminoglycoside antibiotic; the kanamycin resistance cassette I used to knock out *hdsSMR* is known to confer cross-resistance to paramomycin and therefore is the likely reason for this gained phenotype. As expected, the UTI89 methylation revertant does not show any phenotypic differences compared to wild type (Figure 12.B and Table 6.B). Finally, K12 *ΔhdsSMR* which shows transcriptional differences compared to wild type K12 in Section 3.5.2, does not show any additional phenotypic differences by PM (Figure 12.C). Once again the only difference is acquisition of resistance to Paromomycin and Geneticin sulfate (also an aminoglycoside antibiotic with cross-resistance mediated by the kanamycin resistance gene) which can be rationalised by the selection cassette present in this strain (Figure 12.C and Table 6.C).

Of note, there are some wells that visually have some fraction of red or green area under the growth curves; however, discussions with the manufacturer (Biolog) led to the conclusion that these apparent differences are within the range of technical variability of the assay and are not considered reliable. Therefore, based on the experience of Biolog, as well as the concordance with the lack of significant changes in the transcriptional data, I did not validate the growth phenotypes for these putatively unreliable differences.

PM plate	Wells	Quality Score	Compound	Category
A) UTI89 vs. UTI89 ΔhdsSMR				
Phenotypes gained:				
12	C01,C02	355	Paromomycin	Protein synthesis inhibitor, aminoglycoside
19	D07	172	INT	Respiration inhibitor
Phenotypes lost:				
-	-	-	-	-
B) UTI89 vs. UTI89 SWAP UTI89				
Phenotypes gained:				
-	-	-	-	-
Phenotypes lost:				
-	-	-	-	-
C) K12 vs. K12 ΔhdsSMR				
Phenotypes gained:				
12	C01,C02,C03	420	Paromomycin	Protein synthesis inhibitor, aminoglycoside
13	E07,E08	182	Geneticin disulfate (G418)	Protein synthesis inhibitor, aminoglycoside
Phenotypes lost:				
-	-	-	-	-

Table 6: Phenotype Microarray results for UTI89 and K12 mutants. Phenotypic differences observed by PM plates (PM01-20) for Type I mutants with respect to their corresponding wild type strains. Experiment was repeated twice and hits listed were reproducible in both runs, with their Average height difference as an arbitrary quality score (QS). Phenotype is significant if QS >150.

A) UTI89 vs. UTI89 Δ hdsMR



B) UTI89 vs. UTI89 SWAP UTI89



C) K12 vs. K12 Δ hdsMR

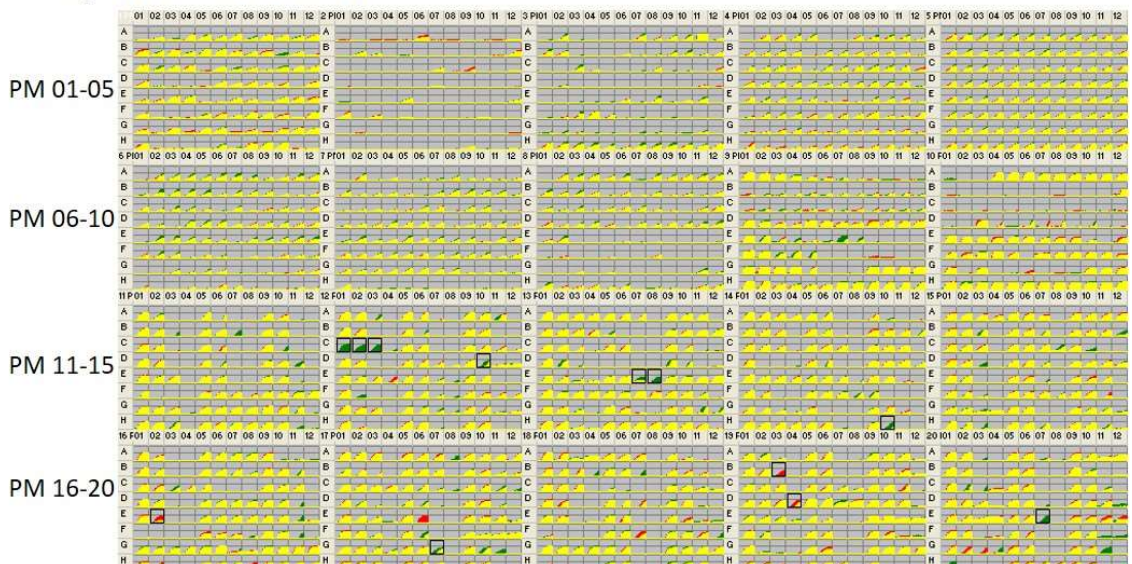


Figure 12: Phenotype Microarray panels for UTI89 and K12 mutants. Phenotype Microarray results panel for **A)** UTI89 vs. UTI89 Δ *hsdSMR*, **B)** UTI89 vs. UTI89 SWAP UTI89 and **C)** K12 vs. K12 Δ *hsdSMR*. Experiment was performed twice, with wells showing phenotypic difference in both runs boxed in all panels. Average height difference for acquired phenotype in mutant vs. reference is a green curve, lost phenotype in mutant vs. reference is a red curve and if both strains are identical under given condition, a yellow curve.

3.7 SUMMARY : Elucidating the role of Type I restriction modification system mediated methylation in regulating *E. coli* virulence and physiology

E. coli strains UTI89, CFT073 and K12 substrain MG1655 each harbour distinct Type I RMSs with both methylation and restriction functions intact. This epigenetic DNA methylation helps distinguish self DNA from non-self DNA, but also has the potential to alter gene expression. Deletion or replacement of the Type I methylation system in UTI89 appears to have no detectable effect in any assay I have used. This includes specific in vitro assays (growth rate, motility, biofilm formation, and HA titre) as well as several potentially more sensitive assays (in vivo infection, RNA-seq, and phenotype microarrays). I found a similar result in Type I methylation knockouts of CFT073 and MG1655, with the exception of one phage and one gene in MG1655 that are upregulated. Overall, the result is that Type I methylation generally has no effect on any phenotype in three different *E. coli* strains, which is a distinctly different conclusion from other published studies describing global (hundreds to thousands) gene expression changes controlled by RMS mediated methylation.

4. RESULTS

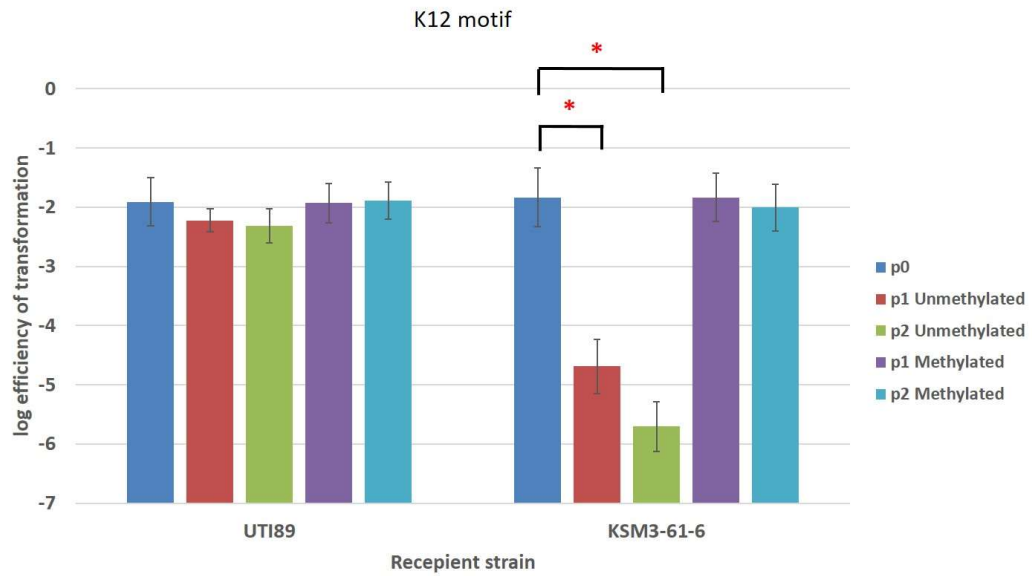
Identification and Characterization of a novel Ribonucleotide reductase gene mutant

In order to investigate the effect of foreign Type I methylation on UTI89 virulence, strain UTI89 *hsdS*^{K12} designated as KSM3-61-6 was generated. Chapter 4 characterises several interesting phenotypes identified in strain KSM3-61-6 and explains how these were attributed to a novel mutation in an *E. coli* ribonucleotide reductase gene and not Type I RMS mediated methylation.

4.1 UTI89 SWAP K12 strain KSM3-61-6 has a functional K12 Type I RMS

KSM3-61-6 Type I RMS function was tested using the Restriction Modification Assay (Figure 13). Using unmethylated and methylated plasmids p0 (0 K12 sites), p1 (1 K12 site) and p2 (2 K12 sites) to transform UTI89 and KSM3-61-6. Only unmethylated plasmids p1 and p2 showed a significant 3 log drop in efficiency of transformation (EOT) when transformed into KSM3-61-6 (Figure 13, right red and green bars). KSM3-61-6 when transformed with methylated preparations of the same plasmids, did not show a decrease in EOT (Figure 13, right purple and light blue bars). UTI89 transformation efficiency is unaffected by the methylation state of these plasmids as they bear K12 methylation motifs which are unrecognizable to wild type UTI89 (Figure 13, left bars). Thus, replacing UTI89 *hsdS* with the K12 allele successfully changed the Type I RMS and by extension methylation to that of K12. This matches data for the other UTI89 SWAP K12 strain KSM6-26-1 in Section 3.2.2.

Figure 13: Restriction modification assay for strain KSM3-61-6. Restriction Modification assay for KSM3-61-6 with plasmids bearing K12 Type I motif. Efficiency of transformation is number of colonies obtained on selective vs. non-selective plates, per unit amount of DNA. Data represents mean and standard deviation from 3 independent experiments. *p<0.05 by Student's T-test.



4.2 UTI89 SWAP K12 strain KSM3-61-6 is defective in kidney colonisation

KSM3-61-6 growth was compared to wild type UTI89 in LB at 37°C for 24 hours. KSM3-61-6 has a significant growth defect compared to wild type during late log phase, between 6 to 10 hours (Figure 14). This LB growth phenotype is absent for strain KSM6-26-1 as observed in Section 3.3.3, even though it is also a UTI89 SWAP K12 strain.

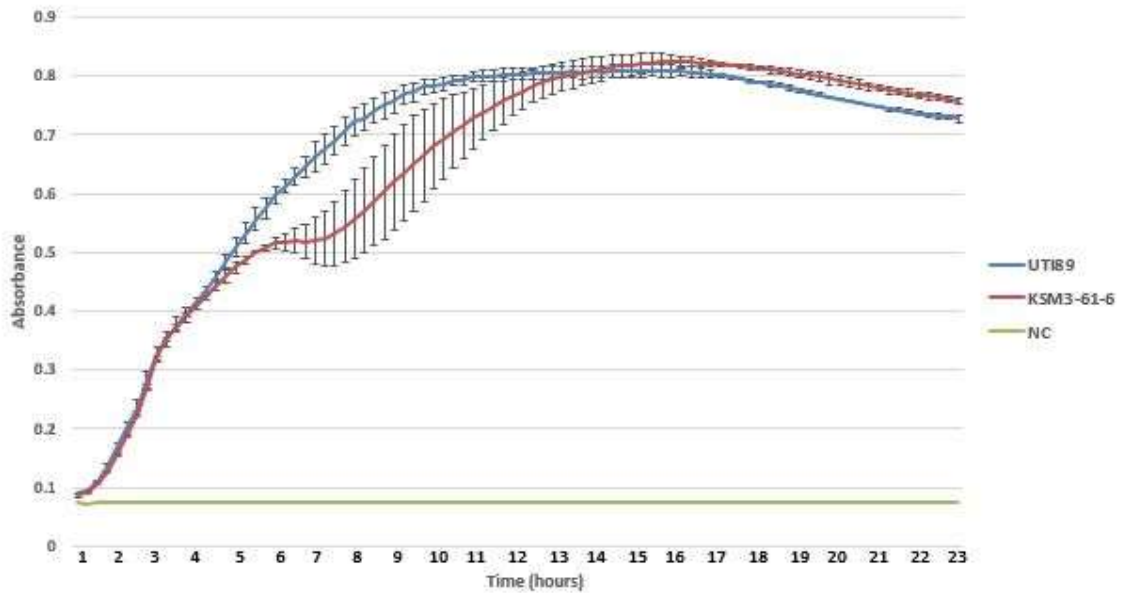


Figure 14: Growth curves for KSM3-61-6. Growth curves were generated for UTI89 and KSM3-61-6 in LB at 37°C. Curves represent the mean and bars represent the standard deviation from three biological replicates. Significance with respect to UTI89 was calculated by student's T-test $p < 0.05$. Negative control (NC) contains media only.

Hemagglutination assay did not reveal any difference between UTI89 and KSM3-61-6 in terms of Type I pilus expression. Competitive infection was performed with UTI89 wild type and UTI89 SWAP K12 strain KSM3-61-6. These co-infections used UTI89 and KSM3-61-6 derived differentially antibiotic resistance marked strains (i.e. UTI89 *att_{HK022}::neo* with UTI89 *hsdS^{K12} att_{HK022}::cat* and UTI89 *att_{HK022}::cat* with UTI89 *hsdS^{K12} att_{HK022}::neo*). Infection was allowed to proceed for 1 and 7 days post infection (dpi) to simulate acute and chronic stages of UTI. Log Competitive Indices (CI) for bladders and kidneys at both time points were determined (Figure 15). Median CI value for kidney infection at 1 dpi (0.62) is significantly different from 0. In other words, at 1 day post infection UTI89 wild type outcompetes UTI89 SWAP K12 strain KSM3-61-6 in the kidney. Since bladder CI at 1 dpi is unaffected, this phenotype reflects a defect for KSM3-61-6 in terms of kidney ascension. By 7 dpi, this kidney phenotype is lost. Once again, this *in vivo* phenotype is in contrast to the same experiment performed in Section 3.4.2 for strain KSM6-26-1.

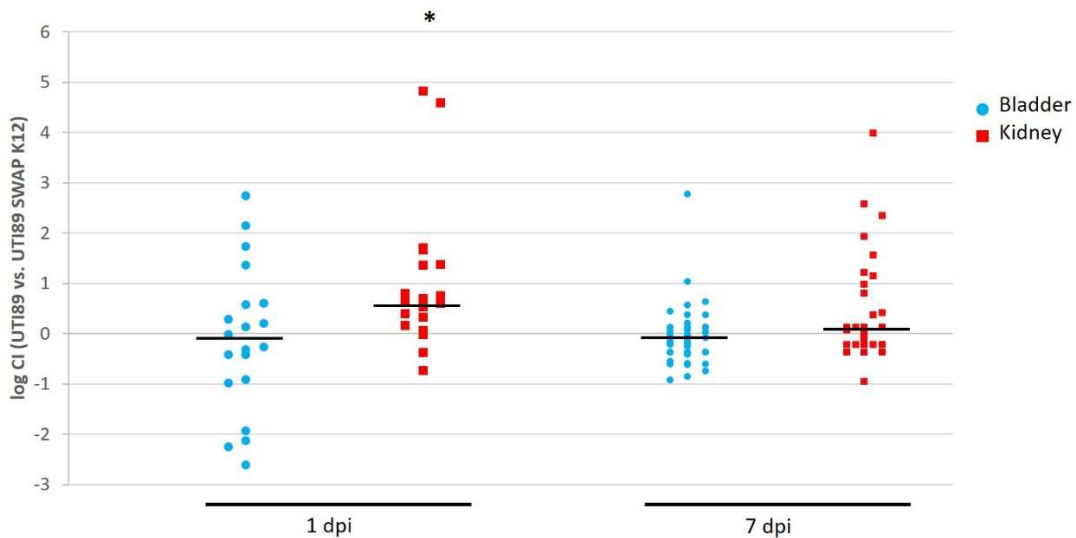


Figure 15: Co-infection with KSM3-61-6. Co-infection with wild type and KSM3-61-6 (UTI89 SWAP K12) strains at 1 (n=20 mice) and 7 (n=45 mice) days post infection. Female C3H/HeN mice were inoculated with equal mixtures of UTI89 and KSM3-61-6 and sacrificed at the time point indicated on the x-axis. The \log_{10} of the competitive index between the two bacterial strains in bladders (blue dots) and kidneys (red dots) are plotted on the y-axis. Medians for each group are indicated by a black horizontal line. Wilcoxon signed rank test with $*p < 0.05$, is considered significant.

4.3 RNA-seq reveals differentially expressed genes in strain KSM3-61-6

Presence of foreign K12 Type I methylation was scrutinised for its effect on the transcriptome of strain KSM3-61-6 using log and stationary phase cultures, as described in Section 3.5.1. Stationary phase transcription of KSM3-61-6 did not reveal any differentially expressed genes (Table 7). However, log phase transcription of KSM3-61-6 shows 47 genes significantly dysregulated with respect to wild type UTI89, with 36 genes downregulated and 11 genes upregulated (Table 8). Strikingly, 27 genes i.e. about 57% of all differentially expressed targets are part of the motility cascade and are all downregulated (Table 8, genes marked in blue). This hints towards a systematic reduction of bacterial motility gene expression. Also, amongst the top hits identified are genes *nrdAB* and *nrdHIEF* which encode the two UTI89 class I ribonucleotide reductase (RNR) systems and are both upregulated (Table 8). RNRs are essential enzyme systems responsible for conversion of ribonucleotides to deoxyribonucleotides. Since the transcriptional changes are observed for UTI89 SWAP K12 strain and not *AhsdSMR* strain, I looked for the presence of K12 methylation sites in the genes listed in Table 8. Three genes, *fliD*, *fliM*, and *mglA* possess one K12 Type I methylation site in their coding sequence and are all downregulated.

Gene ID	Gene	Annotation	log ₁₀ Fold Change	FDR	qPCR Fold Change
UTI89 vs. KSM3-61-6 (UTI89 SWAP K12)					
UTI89_C5050	hsdS	Type I RMS Specificity Protein	-4.59	7.80E-26	

Table 7: Stationary phase differentially expressed genes in KSM3-61-6. Gene expression changes between wild type UTI89 and KSM3-61-6 at stationary phase from triplicate experiments. Gene marked in red has been removed in KSM3-61-6. False discovery rate (FDR) <0.05 and log Fold change >1.5 is considered significant.

Sr. No.	Gene	log ₁₀ Fold Change	FDR
1	hsdS	-6.98	4.51E-27
2	-	2.16	4.76E-08
3	nrdF	2.90	2.49E-05
4	nrdE	2.86	2.56E-04
5	yfoE	1.67	3.33E-04
6	nrdI	3.41	6.37E-04
7	nrdA	1.68	3.07E-03
8	nrdB	1.66	5.18E-03
9	nrdH	2.46	8.61E-03
10	proW	1.52	9.06E-03
11	flgB	-2.35	0.010
12	fliA	-2.43	0.010
13	flgF	-2.68	0.010
14	flgI	-2.29	0.010
15	fliK	-2.74	0.011
16	flgD	-2.37	0.012
17	flgE	-2.56	0.012
18	flgJ	-2.56	0.012
19	yecR	-2.14	0.016
20	fliF	-2.16	0.016
21	flgA	-2.12	0.022
22	-	2.04	0.023
23	flgG	-2.28	0.023
24	-	-2.36	0.023
25	fliJ	-2.25	0.023
26	fliM	-2.05	0.023
27	fliN	-2.32	0.023

Sr. No.	Gene	log ₁₀ Fold Change	FDR
28	fliO	-2.04	0.023
29	-	-1.75	0.024
30	fliE	-1.92	0.027
31	flgC	-1.99	0.027
32	flgK	-2.37	0.027
33	fliE	-1.88	0.027
34	fliI	-2.30	0.027
35	fliH	-2.00	0.030
36	fliZ	-2.12	0.031
37	uidA	-1.96	0.036
38	fliT	-2.34	0.037
39	-	2.13	0.037
40	flgM	-1.90	0.043
41	ydeN	-1.76	0.044
42	mglA	-3.84	0.044
43	tar	-2.84	0.046
44	flgH	-1.90	0.047
45	fliD	-2.44	0.048
46	csgB	-1.50	0.049
47	yrbA	-1.83	0.049
48	fliG	-1.83	0.049

Table 8: Log phase differentially expressed genes in KSM3-61-6. Gene expression changes between wild type UTI89 and KSM3-61-6 at log phase from triplicate experiments. Genes marked in blue are part of the motility regulon and gene marked in red has been removed in KSM3-61-6. Upregulated genes have their fold change marked in green, downregulated in black. False discovery rate (FDR) <0.05 and log Fold change >1.5 is considered significant.

The gene expression changes observed for UTI89 SWAP K12 strain KSM3-61-6 are in contrast to data for KSM6-26-1 (Section 3.5.1), where no changes were observed. Since, both strains should be identical genotypically i.e UTI89 *hsdS*^{K12}, we validated transcript levels for 4 genes *nrda*, *flhC*, *fliA* and *fliC* by qRT-PCR (Figure 16). *flhC*, *fliA* and *fliC* are Class I, II and III genes in the flagellar cascade, respectively. *flhC* being the a master regulator of the entire regulon, *fliA* a flagellar specific sigma factor σ^{28} required for late gene expression and *fliC* encodes the main structural component, flagellin [165]. Of the three motility associated genes only *fliA* appears significantly downregulated in KSM3-61-6 by RNA-seq (Table 8). *nrda* encodes the alpha sub unit of the class Ia RNR system.

qRT-PCR validates RNAseq results for both UTI89 SWAP K12 strains; wherein *fliA* and *nrda* are down and upregulated respectively for KSM3-61-6 (Figure 16 and Table 8) and unaffected in KSM6-26-1 (Figure 16 and Section 3.5.1). Also, the vast majority of *E. coli* Class II and III motility genes are downregulated in KSM3-61-6 either by RNA-seq or qRT-PCR (*fliC*) (Figure 16 and Table 8). But not the master regulator *flhC*, a possible mechanistic target in KSM3-61-6. FliC expression is mediated by very efficient translation from a few transcripts [29], which potentially explains why there is a reduction in KSM6-26-1 *fliC* mRNA but no reduction in KSM6-26-1 motility (Section 3.3.1).

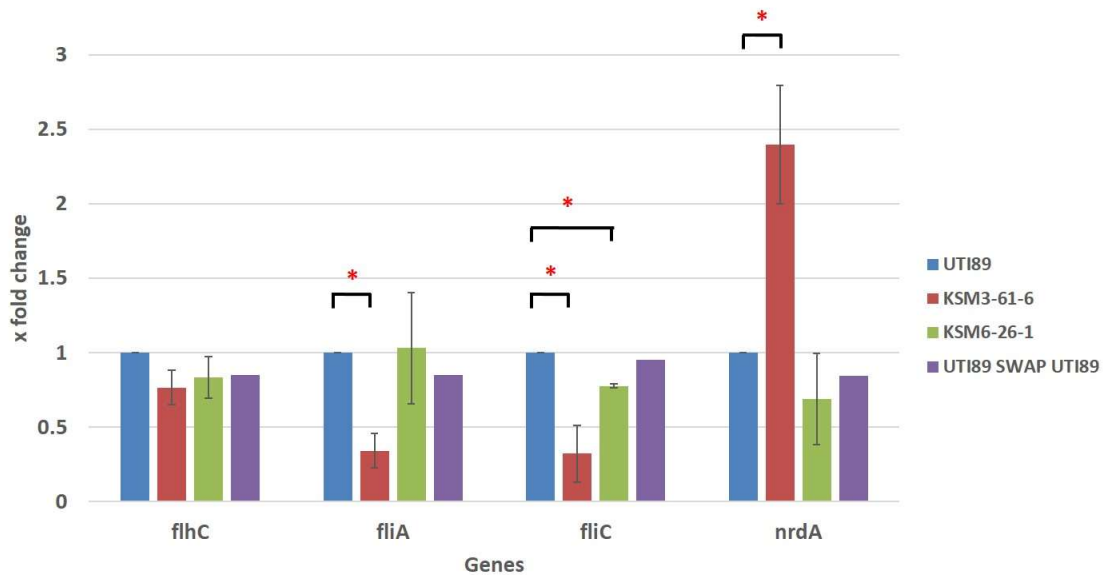


Figure 16: qRT-PCR for differentially expressed genes in KSM3-61-6. Fold change calculated by qRT-PCR using $\Delta\Delta C_t$ method and rRNA gene *rrsA* as internal control, for targets identified by RNA-seq. Experiment repeated thrice for UTI89, KSM3-61-6 and KSM6-26-1, but only once for UTI89 SWAP UTI89. Transcript abundance is considered significantly different for $*p < 0.05$ by student's T-test, compared to corresponding wild type.

Based on the gene expression profile of KSM3-61-6, it appears that this strain should demonstrate diminished motility. Soft agar motility assay demonstrated an approximately 25 % reduced motility for KSM3-61-6 compared to wild type UTI89 (Figure 17). This also potentially explains the *in vivo* defect in kidney colonisation observed for this strain at 1 dpi (Figure 15). It is expected that lack of motility causes a transient bladder and especially kidney colonisation defect around 1 to 3 dpi, which gradually diminishes and infection titres return to wild type levels at later time points [29, 31]. The subtle extent of the kidney colonisation defect observed in Figure 15 may stem from the fact that KSM3-61-6 has reduced motility as observed in Figure 17, but isn't completely non motile like the *AfliC* mutant in this paper [29]. Thus, the motility defect identified by RNA-seq in strain KSM3-61-6 is a genuine phenotype, which extends *in vivo* and results in a transient kidney colonisation defect along expected lines from published literature.

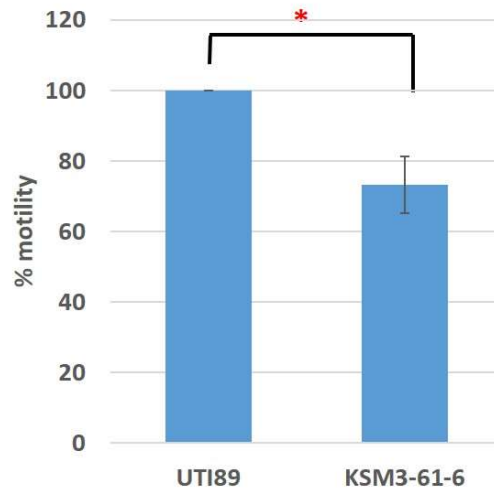


Figure 17: Motility of strain KSM3-61-6. Soft agar motility assay for comparing wild type and KSM3-61-6 motility. Experiment was performed thrice, with mean and standard deviation plotted. Student's T-test is used to determine significant difference at $*p < 0.05$.

4.4 Phenotype Microarray identifies additional phenotypic differences for KSM3-61-6

Phenotype Microarray (PM) was performed on UTI89 SWAP K12 strain KSM3-61-6 and revealed differences with respect to wild type UTI89 (Figure 18). A systematic negative difference was observed in PM plate 5 containing nutrient supplements and also PM plates 11 to 20 representing a wide range of chemical inhibitors (Figure 18). These differences primarily constitute marginal single dose differences which never achieved the Quality score cut-off of 150 and hence are considered as background. Instead, attention was focused on phenotypes which demonstrated a large ($QS > 150$) multi dose effect, listed in Table 9. There are 16 conditions under which KSM3-61-6 differs from UTI89. Majority of these represent nucleic acid analogues, DNA damage inducers and inhibitors of enzymes involved with DNA processing. One of the top hits, Azathioprine a purine analogue was selected for independent validation because it represents a common category of inhibitors identified here.

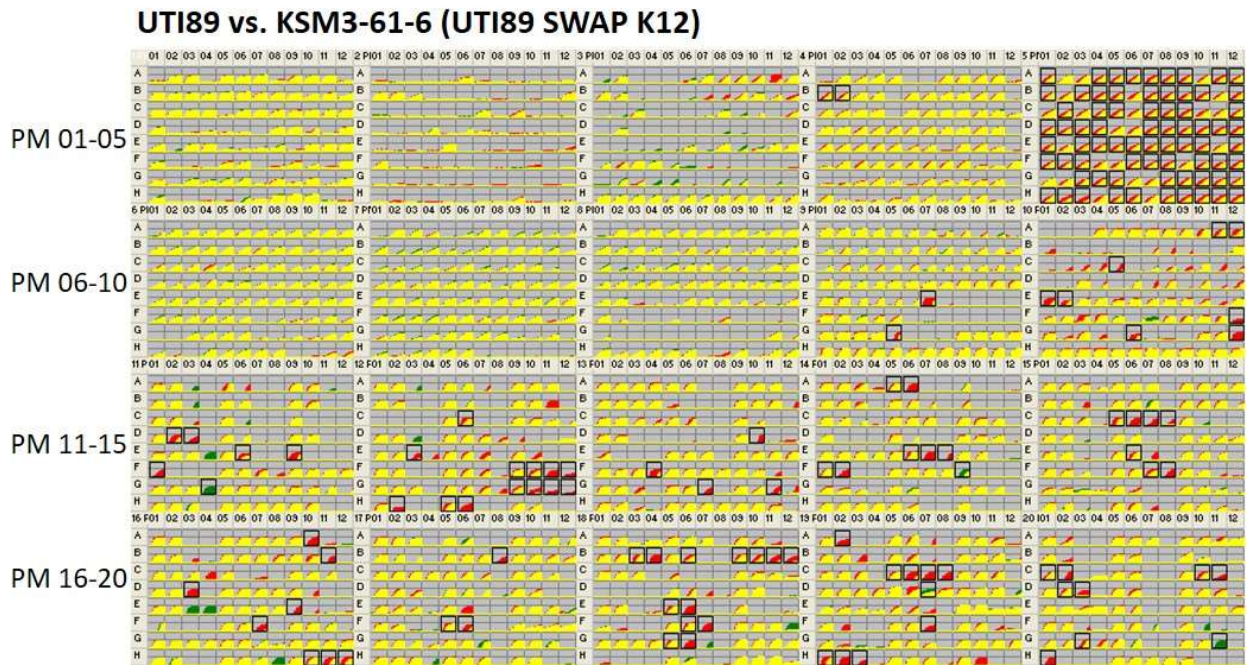


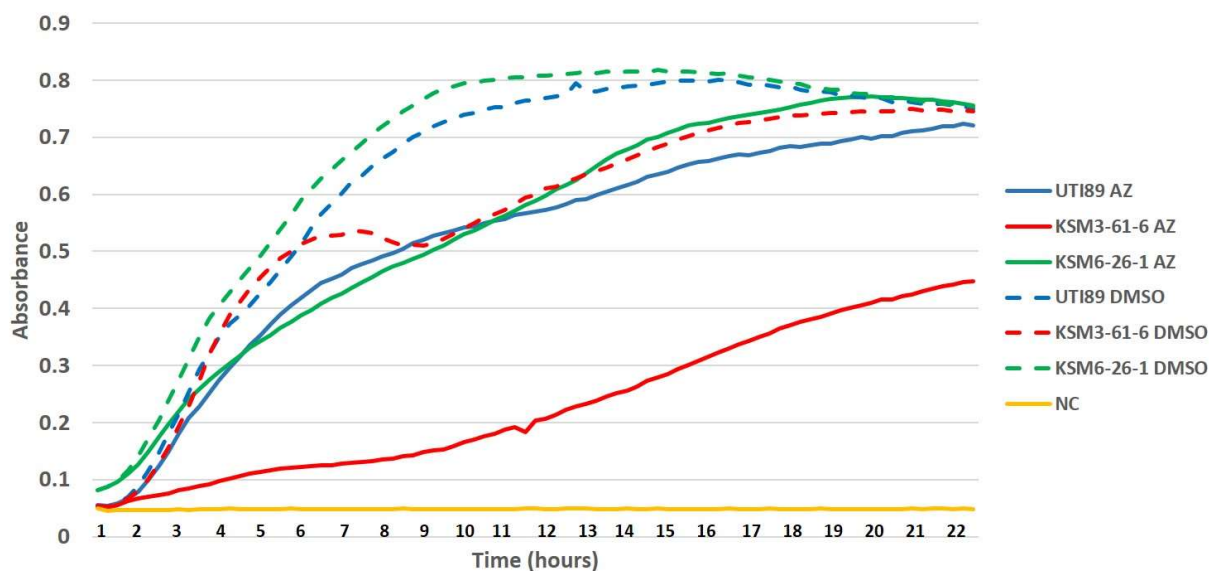
Figure 18: Phenotype Microarray panel for KSM3-61-6. Phenotype Microarray results panel for UTI89 vs. KSM3-61-6 (UTI89 SWAP K12). Experiment was performed twice, with wells showing phenotypic differences in both runs marked with boxes. Average height difference (or quality score) for yellow curves is low signifying no difference, positive for green curves signifying an acquired phenotype in mutant vs. reference strain and negative for red curves representing a lost phenotype in mutant vs. reference strain.

Table 9: Phenotype Microarray results for KSM3-61-6. List of phenotypic differences observed in PM plates (PM01-20) for KSM3-61-6 with respect to UTI89 wild type. Experiment was repeated twice and hits listed were reproducible in both runs, with their Average height difference as an arbitrary quality score (QS). Phenotype is considered significant if it displays a multi-dose effect for a particular condition/chemical inhibitor and a QS >150.

PM plate	Wells	Quality Score	Compound	Category / Target
UTI89 vs. KSM3-61-6 (UTI89 SWAP K12)				
19	C05,C06,C07,C08	-487	7-Hydroxycoumarin	DNA intercalator
15	C05,C06,C07,C08	-366	Fusidic acid, sodium salt	Protein synthesis, elongation factor
18	B09,B10,B11,B12	-341	Azathioprine	Nucleic acid analog, purine
19	H01,H02,H03	-321	Hexamine cobalt (III) chloride	DNA synthesis
12	F09,F10,F11,F12	-303	5-Fluoroorotic acid	Nucleic acid analog, pyrimidine
14	E06,E07,E08	-265	Nitrofurantoin	Nitro compound, oxidizing agent, DNA damage
12	G09,G10,G11,G12	-254	L-Aspartic-b-hydroxamate	tRNA synthetase
16	H10,H11,H12	-243	Sorbic acid	Respiration, ionophore, H+
18	B03,B04	-225	Trifluorothymidine	Nucleic acid analog, pyrimidine, DNA synthesis
12	H05,H06	-221	Rifampicin	RNA polymerase
18	E05,E06	-207	Sodium periodate	Toxic anion, oxidizing agent
11	D02,D03	-202	Capreomycin	Protein synthesis
14	A05,A06	-190	Furaltadone	Nitro compound, oxidizing agent, DNA damage
18	G05,G06	-190	3,5- Diamino-1,2,4-triazole (Guanazole)	Ribonucleotide DP reductase inhibitor
18	F06,F07	-185	Tinidazole	Nitro compound, oxidizing agent, DNA damage
15	F07,F08	-184	Oleandomycin, phosphate salt	Protein synthesis, 50S ribosomal subunit, macrolide

Azathioprine at a concentration of 1mM caused a reduction in UTI89 growth versus DMSO diluent control (Figure 19, blue solid and dash lines). The UTI89 SWAP K12 strain KSM6-26-1 behaves exactly like wild type, (Figure 19, green solid and dash lines). However, in agreement with Phenotype Microarray (Table 9), UTI89 SWAP K12 strain KSM3-61-6 shows a dramatic statistically significant Azathioprine induced reduction in growth (Figure 19, red solid line). Even in the presence of DMSO this strain shows a significant defect during late log phase (6 to 16 hours), reminiscent of its LB growth defect as observed in Figure 14.

Figure 19: Azathioprine sensitivity of UTI89 SWAP K12 mutants. Growth curves for UTI89, KSM3-61-6 and KSM6-26-1 in the presence of 1mM Azathioprine (AZ, solid lines) or diluent (DMSO, dash lines). Negative control (NC) contains media only. Curves represent mean values from three biological replicates, with significance calculated using student's T-test, $p < 0.05$.



4.5 Secondary mutation independent of methylation is responsible for KSM3-61-6

phenotypes

Even though strains KSM6-26-1 and KSM3-61-6 are genotypically identical i.e. UTI89 *hsdS^{K12}*, they behave differently in terms of growth kinetics, motility, kidney colonisation, gene expression and sensitivity to several chemical inhibitors (Sections 4.2 to 4.4). Whole genome sequencing (WGS) of KSM6-26-1 revealed no unexpected mutations, which allows confident assignment of the role of Type I DNA methylation. However, WGS of strain KSM3-61-6 revealed two nonsynonymous mutations acquired by this strain in the process of cloning (Table 10). Tetranucleotide insertion in *ybaL* results in a premature stop codon and consequently a protein which is 165 amino acids shorter. A single G to T SNP in gene *nrdA* results in a K584N mutation. Both mutations were independently confirmed by Sanger sequencing.

Position	Reference	KSM3-61-6	Mutation	Gene	Annotation
516083	T	TCTTC	Premature stop codon, 165 aa truncation	<i>ybaL</i>	Putative cation:proton antiport protein
2482923	G	T	K584N	<i>nrdA</i>	Ribonucleotide-diphosphate reductase

Table 10: Mutations present in KSM3-61-6. Mutations revealed by whole genome sequencing (WGS) of strain KSM3-61-6 with respect to wild type UTI89. Position, type and consequence of nonsynonymous mutations are listed. Excludes known manipulation of *hsdS* locus and each mutation independently confirmed by Sanger sequencing.

Based on sequence homology, *ybaL* is a putative cation-proton antiport system with no information available on the effects of *ybaL* deletion on *E. coli* physiology. *nrdA* is a class Ia RNR gene, which is essential in *E. coli* and has been the subject of previous mutational analysis; K584N however is a novel *nrdA* mutation. To correctly assign the altered phenotype of strain KSM3-61-6, the above *ybaL* and *nrdA* mutations were independently generated in UTI89 wild type. Two phenotypes, namely motility (Figure 17) and Azathioprine sensitivity (Figure 19) were selected to further narrow down the gene responsible for the phenotypic differences between the two UTI89 SWAP K12 strains. *ybaL* mutant does not show the motility defect (Figure 20) nor Azathioprine sensitivity (Figure 21) of strain KSM3-61-6. Mutation in this gene closely mimics wild type UTI89 in both assays. However, *nrdA* K584N mutation results in decreased motility (Figure 20) and Azathioprine sensitivity (Figure 21, solid green line), which very closely replicates the phenotype observed in KSM3-61-6.

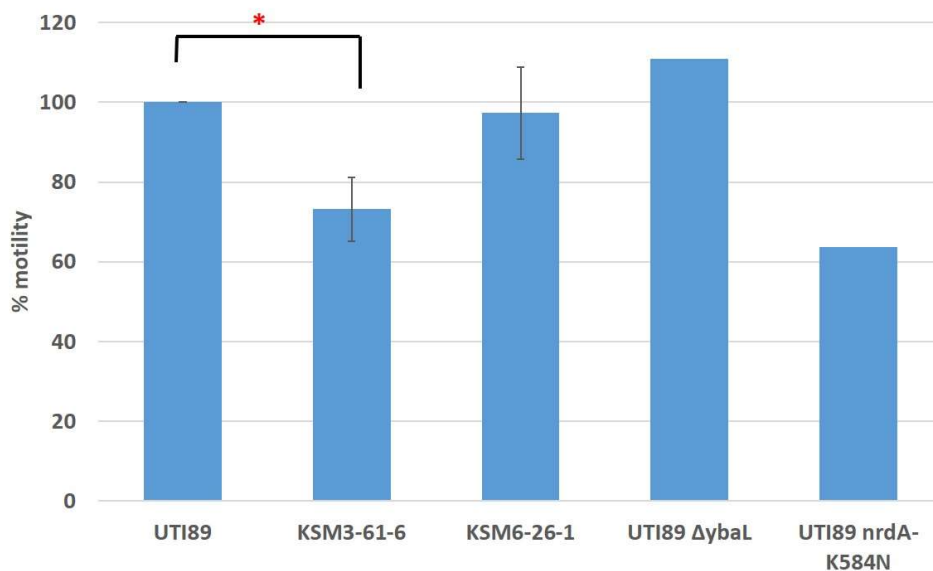


Figure 20: Motility of UTI89 *ybaL* and *nrdA* mutants. Motility of different UTI89 mutants assessed by soft agar motility assay. Experiment for UTI89, KSM3-61-6 and KSM6-26-1 performed thrice and data represents mean and standard deviation. Experiment for UTI89 *ybaL* and *nrdA*-K584N mutants performed only once. Significance for data with biological replicates was evaluated using student's T-test, * $p < 0.05$

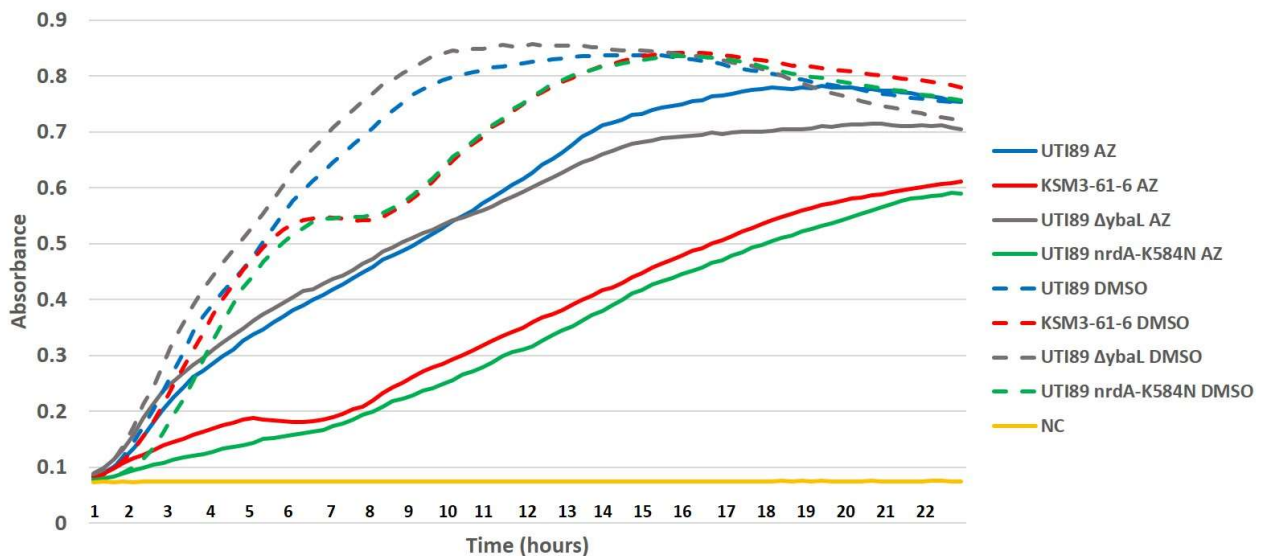


Figure 21: Azathioprine sensitivity of UTI89 *ybaL* and *nrdA* mutants. Growth curves for UTI89, KSM3-61-6, $\Delta ybaL$ and *nrdA*-K584N in the presence of 1mM Azathioprine (AZ, solid lines) or diluent (DMSO, dash lines). Negative control (NC) contains media only. Curves represent mean values from three independent experiments for UTI and KSM3-61-6, and one experiment for $\Delta ybaL$ and *nrdA*-K584N UTI89 single mutants. Significance calculated using student's T-test, $p < 0.05$.

4.6 SUMMARY: Identification and Characterization of a novel Ribonucleotide reductase gene mutant

Ribonucleotide reductase (RNR) systems are responsible for the *de novo* synthesis of deoxyribonucleotides from their ribonucleotide precursors. RNR systems play an important role in regulating DNA replication and repair, with a single system capable of generating all four deoxyribonucleotides (dATP, dCTP, dTTP and dGTP) [166]. RNR systems are subject to complex allosteric regulation, wherein a balanced cellular dNTP pool is constantly maintained by sampling the dNTP pool and accordingly synthesizing the appropriate dNTP as well as alternating between active and inactive quaternary structures when required [167, 168]. Any imbalance can be lethal as it can result in aberrant mutation rates [169, 170]. Owing to their essential role in a central cellular metabolic process, namely DNA synthesis and their ubiquitous nature in all living organisms, RNR systems are attractive drug targets. Chemotherapy drugs targeting RNR systems as antiproliferatives already exist and efforts are underway to develop high throughput screening strategies for RNR inhibitors, an otherwise lab-intensive process, so as to identify novel therapeutics including antibiotics [171].

nrdAB represents the primary *E. coli* RNR system, functional under aerobic conditions and unlike the other RNR systems present in the genome (*nrdHIEF* and *nrdDG*), it is essential [172]. WGS of a Type I methylation mutant identified a mutation within *nrdA*, namely K584N. Amino acid K584 does not encompass the catalytic or any of the allosteric sites [172] and based on literature it is a novel mutation. Phenotypically, this mutation shows reduced growth at late log and log to stationary phase transition, which is when wild type *nrdA* expression is maximal [173]. *nrdA*-K584N demonstrates sensitivity to nucleotide analogues, DNA damage inducers and most strikingly, RNR inhibitor Guanazole by phenotype microarray; Table 9 Section 4.4. Increased expression of *nrdAB* and *nrdHIEF* as well as sensitivity to known RNR inhibitors suggests that K584N is a hypomorph, confirmation of which requires complementation to validate and direct quantification of the cellular dNTP pool. Most importantly, this is one of the first reports

phenotypically linking a RNR system and motility [174]. *E. coli nrdAB* is known to be regulated by a complex regulatory network which includes DnaA, Ici, Fnr, Fur, Fis, NrdR and H-NS, several of which are global transcription regulators [175]. Several of these are independently known to regulate motility either by directly binding to the promoters of motility genes, regulating secondary messengers or by unknown mechanisms [174, 176-178]. K584N motility phenotype first requires to be validated by complementation, after which it represents a useful mutant to try and identify the genes involved in this regulatory crosstalk, either by a candidate based approach or an unbiased mutagenesis screen.

Leveraging the strengths of next generation sequencing, classical microbiological assays and appropriate controls; it was possible to identify and distinguish phenotypes associated with a novel mutation in RNR *nrdA*, which was independent of Type I DNA methylation. This previously unknown lysine to asparagine mutation in *nrdA*, represents a potentially useful tool for improving our understanding of the regulation and functioning of these essential enzyme systems. UTI89 *nrdA*-K584N was discovered as a spontaneous mutant in the course of elucidating the role of Type I methylation in *E. coli* virulence, it is not the primary focus of this thesis and the data presented in chapter 4 is preliminary at best warranting further scrutiny.

5. DISCUSSION

Methylation of DNA is a well-known epigenetic mechanism with the ability to impart to the genome an additional layer of functionally relevant information. Moreover, this epigenetic signature of the genome is amenable to rapid, heritable alterations in response to external stimuli. First studied in the context of mammalian development and disease [171, 172], the relevance of DNA methylation in several bacterial physiological processes quickly emerged [94]. The phenomena of DNA methylation in bacteria was first identified in the context of host defence over half a century ago [51]. Methyltransferases associated with bacterial Restriction Modification Systems (RMSs) methylate DNA in specific sequence contexts to help differentiate self from non-self DNA, protecting against harmful phages and stabilising the genome by preventing unregulated horizontal gene transfer. Subsequently, highly conserved DNA methyltransferases have been identified independent of the classical RMS endonucleases and termed as orphan methyltransferases [91]. These orphan methyltransferases caught the attention of researchers as their conservation [76] and in certain cases essentiality [91, 110, 111] implied that an alternative function must exist. Orphan methyltransferases such as *dam*, *ccrM* and *dcm* have been the subject of intense scrutiny and *dam* methylation does in fact play an important role in bacterial virulence [73, 95, 99, 101-105, 107-112, 114, 179, 180]. *dam* and *ccrM* even regulate basic cellular processes such as DNA replication, cell division, mismatch repair and controlling transposons [94].

RMS associated methyltransferases have not been studied as extensively, because their primary function is believed to be host defence. However, by virtue of methylating DNA to identify host genomic material, these methyltransferases can function via mechanisms similar to those utilised by orphan methyltransferases. Publications exist elucidating the role of phasevariable RMSs, mainly Type III capable of reversible ON/OFF switching of methyltransferase expression [124-128, 130-132] and complex multi-specificity Type I RMSs

capable of rapidly and reversibly replacing one methylation mark with another [62, 63]. In both these cases these RMSs confer upon the host genome the ability to rapidly change phenotypic traits to allow better colonisation of a niche or immune evasion, contributing to virulence. Publications mentioning such functions for classical RMSs, which lack the ability to reversibly change the host methylome are rare [58, 88]. This paucity of information regarding the potential role of methylation mediated by classical, fully functional RMSs in virulence is either because such systems in fact do not deviate from their primary roles or have not been subjected to proper systematic scrutiny.

Urinary tract infections (UTIs) are one of the most common infections, both in the community as well as nosocomial settings [2]. Although usually self-limiting and treatable by antibiotic therapy, the high incidence and rate of recurrence means that UTIs are a significant cause of morbidity and public healthcare expenditure [36]. Uropathogenic *E. coli* (UPEC), which is the primary causative agent for UTIs, undergoes a complex life cycle involving intracellular and extracellular stages and is adept at evading host immunity and treatment. The repeated and prolonged use of antibiotics for UTIs carries an additional risk of causing the emergence of antibiotic resistance [1]. Attempts to characterise a universal set of virulence factors specific to UPEC have not succeeded. In fact, UPEC strains often encode diverse types and numbers of virulence factors, making characterisation and targeting of urinary tract specific virulence factors difficult [27]. However UPEC virulence factors such as Type I pili, P pili and Antigen 43 do show epigenetic regulatory mechanisms, either due to an invertible promoter element or *dam* methylation [99, 100, 181]. It thus stands to reason that epigenetics, especially DNA methylation, plays an important role in regulating UPEC virulence. In addition to the few currently known, locus specific effects mediated by orphan methyltransferase *dam*, whose underlying molecular mechanisms have been elucidated [99, 151]. The overall objective of this study is to systematically elucidate the functions of the identified methyltransferases irrespective of the type of enzyme, in regulating virulence as well as bacterial physiology.

5.1 UTI89 encodes a novel Type I methylation motif mediated by a functional RMS

Based on sequence homology to known methyltransferases, UTI89 is predicted to possess four distinct methylated motifs [79, 138]. However, in order to find which methyltransferases are functional, we utilised Single Molecule Real Time (SMRT) sequencing. SMRT sequencing identified three motifs methylated to varying extents in the UTI89 genome. Two of these 5'-GATC-3' and 5'-CCWGG-3' belong to known *E. coli* orphan methyltransferases *dam* and *dcm* respectively. Although majority of *dam* methylation sites (96.9%) were methylated, only 1.8% of *dcm* sites could be detected as methylated. This low proportion of methylation is probably owing to the weak and diffuse signal detected for 5-methylcytosine (5mC) in SMRT sequencing [157] and also because majority of *dcm* sites are methylated during stationary phase only [74]. However, the UTI89 genome did display a novel bipartite methylated sequence 5'-CCA(N₇)CTTC-3' (methylated N6-methyladenine residues on both strands are underlined). This methylation sequence consists of two specific parts, separated by a degenerate sequence of fixed length, which is characteristic of Type I RMSs. Indeed, the UTI89 genome does possess a previously uncharacterised Type I RMS *hsd* (host specificity determinant) gene cluster. This RMS consists of three genes *hsdS* (Specificity determining DNA binding component), *hsdM* (Methyltransferase) and *hsdR* (Endonuclease). The *hsd* gene cluster is located on a hypervariable region of the *E. coli* chromosome called the immigration control region, which is also present in other *E. coli* strains such as lab adapted commensal strain K12 substrain MG1655 and pyelonephritis causing strain CFT073 [54, 138]. The K12 Type I RMS called EcoKI is the first one discovered, in the 1960s and has served as the model for all other Type I RMSs. This K12 RMS has a known specificity (5'-AAC (N₆)GTGC -3') [51], while the CFT073 Type I RMS is predicted to methylate (5'-GAAG (N₇)TGG -3') [138].

A considerable amount of data exists regarding the roles of *dam* and *dcm* in regulating important cellular processes in other *E. coli* as well as other bacterial species. We therefore

decided to systematically investigate the effect of epigenetic DNA methylation mediated by the putative Type I RMS. RMS associated methyltransferases are found in conjunction with a partner endonuclease, however due to acquired mutations such as truncations or premature stop codons these endonuclease genes could be silenced. Therefore a perceived RMS could actually just be an orphan methyltransferase whose endonuclease component is in the process of evolutionary removal. However, studies pertaining to RMS associated methyltransferases do not always check both methylation and restriction functions. To identify the genes responsible for unique UTI89 Type I methylation and demonstrate the functional status of the responsible RMS, we utilised an adapted plasmid transformation based Restriction Modification (RM) Assay [182]. The RM assay was first validated using the previously well characterised K12 Type I RMS. Using UTI89 wild type and Type I RMS deletion mutant, the novel bipartite methylated motif 5'-CCA(N₇)CTTC-3' was successfully assigned to *hsdSMR*. Moreover, not only was this Type I RMS capable of methylating the target site in plasmid DNA, but also distinguishing methylated and unmethylated copies of the site and selectively restricting unmethylated plasmids only. Similarly, the CFT073 Type I RMS was also tested by the RM assay. CFT073 Type I RMS also has both restriction as well as modification functions intact and recognises the sequence 5'-GAAG(N₇)TGG-3' as predicted. Thus, each of the three *E. coli* strains encode one fully functional Type I RMS, each with a distinct methylation specificity.

Type I RMS methylation specificities are difficult to determine by classical methods owing to the fact that they cleave DNA at variable distances from their recognition site. However, based on a mixture of sequence based predictions as well as biochemical confirmation, there are approximately 1047 putative bacterial Type I RMS specificities [138]. Type I RMSs belonging to the same family show a high degree of homology within *hsdMR*, differences mainly exist in the *hsdS* gene which determines DNA specificity. Type I RMS specificity diversification frequently occurs by exchanging either the entire or partial *hsdS* gene by intra family homologous recombination [51, 58]. Also, most RMSs are encoded on mobile genetic elements and although rarer, transfer of entire RMSs can also add to the host RMS repertoire [65, 146]. Owing to the

modular nature of Type I and III RMSs and the fact that the endonuclease component does not encode the target recognition domain (TRD) for DNA binding, loss or alterations in the TRD does not put the host at immediate risk for endonucleolytic attack. Therefore, to simulate the acquisition of a foreign Type I methylation specificity in UTI89, the native Type I RMS genes were replaced with those from K12 and CFT073. K12 is a gut derived commensal isolate, but its RMS belongs to the same Type IA family and hence merely the chromosomal *hsdS*^{UTI89} allele was replaced with *hsdS*^{K12}. CFT073 is another UPEC strain, but its Type I RMS belongs to a different family and hence it necessitated replacing the entire native *hsdSMR*^{UTI89} locus with *hsdSMR*^{CFT073}. To validate whether these altered RMSs actually function, the RM assay was used. As expected, the UTI89 SWAP K12 strain KSM6-26-1 expressed a fully functional K12 Type I RMS in the UTI89 genomic context. Also, UTI89 SWAP CFT073 strain KSM6-80-15 expressed a fully functional CFT073 Type I RMS in UTI89. This is one of the first studies to chromosomally swap Type I RMS mediated methylation patterns so as to evaluate the consequences of foreign methylation on host phenotype.

5.2 Altered UTI89 Type I methylation does not affect virulence

Based on literature, DNA methylation can alter gene expression by direct mechanisms such as influencing the binding affinity of regulatory proteins to the promoters of genes [90, 99, 106] and also by unknown indirect mechanism via methylated sites not present at or near the affected genes [72, 86]. Also, evidence exists for both orphan as well as RMS associated methyltransferases exerting an effect on gene expression. We thus hypothesised that removal of 754 native UTI89 Type I methylation sites as well as the introduction of 668 (K12) or 784 (CFT073) foreign methylation sites should manifest phenotypically in UTI89. Phenotypes such as biofilm formation, motility, Type I pilus expression and growth kinetics are important determinants of UTI89 virulence [152] and can be assayed *in vitro* under precise conditions. Type I pili are required for attachment and invasion of target urothelial cells lining the lumen of the

bladder as well as the subsequent formation of biofilm-like intracellular bacterial communities [8, 9, 183]. Biofilm formation acts as a surrogate for the ability to form IBCs. Motility is required transiently for both UPEC adhesion to target cells and ascension to the kidneys [29, 30]. Using these versatile *in vitro* proxies for virulence, UTI89 Type I methylation deletion and swap mutants were evaluated for differences. Surprisingly, UTI89 mutants did not differ from wild type or the revertant in any of the *in vitro* phenotypes tested. Based on a limited array of UPEC specific virulence assays, it appears that contrary to expectations Type I RMS methylation did not alter UTI89 phenotype.

Although informative, *in vitro* assays can check only specific parameters and by no means represent an exhaustive scrutiny of possible phenotypes nor replicate complex host environments. Competitive infections comparing both UTI89 lacking native Type I methylation and bearing K12 Type I methylation, showed no difference from wild type. Lack of a virulence phenotype cannot be due to the locus and type of inserted selection marker, as this was controlled for. This result is in agreement with *in vitro* data and although unexpected is convincing.

5.3 Type I methylation does not affect global gene expression patterns or phenotypes in multiple *E. coli* strains

Although DNA methylation can potentially alter expression patterns so as to influence virulence, data with respect to UTI89 and its Type I methylation mutants convincingly proves otherwise. Therefore in order to strengthen the alternative hypothesis that methylation mediated by a fully functional classical Type I RMS such as the UTI89 RMS does not perturb global gene expression, we employed two approaches. Firstly, we decided to systematically screen using broad high throughput approaches for any changes mediated by Type I methylation using RNA-seq and Phenotype Microarray. Secondly, we extended the hypothesis to include two other Type I RMSs similar to the one in UTI89, one in uropathogen CFT073 and another in lab strain K12. It is important to reiterate the similarities between all three Type I RMSs, in that they are all fully

functional and methylate less than 760 sites each in their respective genomes (754, 595 and 427 for UTI89, K12 and CFT073 Type I RMSs respectively). SMRT sequencing data shows that 99% of the UTI89 Type I methylation sites are methylated, as would be expected for a RMS with a functional restriction component. However, consistently unmethylated sites such as those protected from methylation by the binding of regulatory proteins could exist for all three *E. coli* strains. For example, unmethylated GATC sites are found in the promoter of the *dnaA* gene due to binding by the SeqA repressor protein [94].

Strand specific RNAseq was performed to investigate the transcriptional patterns in stationary and exponential phases. Comparisons were made using the entire panel of UTI89 Type I methylation mutants including a deletion mutant, UTI89 bearing K12 methylation, UTI89 bearing CFT073 methylation mutant and a revertant as control. Comparisons were also made between K12 and CFT073 with their respective Type I RMS deletion mutants. We thus hoped to be thorough enough to be able to detect any possible effect of Type I RMS methylation on gene expression. None of the UTI89 Type I methylation mutants showed any changes in transcript levels relative to wild type in exponential or stationary phase. Similarly, lack of CFT073 Type I methylation did not alter log or stationary phase gene expression either. Stationary phase transcriptomes for K12 Type I RMS deletion mutant remained unaffected, except for the subtle borderline upregulation of *uhpA*, encoding a two component sugar transport system. However, during log phase K12 showed upregulation of four genes, all four belonging to the same prophage.

K12 cryptic prophage CP4-44 gene upregulation was validated by qRT-PCR and represents a single small locus specific effect. Interestingly, four K12 methylation sites were identified in the coding sequence of the CP4-44 phage genes identified as upregulated. This raises the possibility of several direct mechanisms for RMS mediated regulation, including alleviation of restriction pressure exerted due to the presence of multiple Type I sites or direct methylation mediated regulatory suppression. Thus one of the first steps would be to delineate and identify if

phage expression is due to methylation or restriction. This can be tested by comparing CP4-44 upregulation in K12 *ΔhsdSMR* (methylation and restriction null mutant) versus K12 *ΔhsdR* (restriction null mutant). Coincidentally K12 Type I RMS deletion mutant also shows a growth defect in LB during log phase. Consistent with our data, Type I RMS mediated methylation does not result in global gene expression changes.

Although RNA-seq is a frequently used and powerful genomic tool to study subtle changes in transcription, it provides information only under the specific conditions tested. Most studies either check a particular growth phase or under a specific, relevant *in vivo* condition. In selecting exponential phase cultures we consciously selected an actively replicating bacterial population. Similarly, bacterial cultures passaged for two successive 24 hour periods were selected to replicate the stationary phase inoculum used in the murine model for UTI. However, we acknowledge that this still represents a limited set of conditions and to conclusively prove that Type I RMS methylation bears no consequence on the host, we need to cast the widest net possible. To this end, we decided to employ a high throughput phenotypic screen called Phenotype Microarray by Biolog Inc. Phenotype Microarray (PM) allows the simultaneous analysis in a convenient 96-well format of almost 2000 distinct phenotypes [156].

In the absence of any prior information about a particular protein, PM can provide useful clues as to the relevant biological function of the protein. In this case, we use PM with the intent of elucidating any previously neglected phenotypes for Type I RMS methylation. UTI89 and its revertant were included as a negative control and indeed not a single phenotypic difference was observed. But, UTI89 and K12 Type I RMS deletion mutants also did not show a single phenotypic difference across the entire PM panel, confirming our hypothesis and the rest of the data. K12 lacking native Type I methylation which showed reduced growth in LB (Section 3.3.3) and increased transcription of prophage genes (Section 3.5.2), did not show any hits via PM.

5.4 Conclusions

To summarise, we utilised an array of *in vitro* and *in vivo* UPEC specific methods to test our initial hypothesis that DNA methylation mediated by Type I RMS can modulate the virulence of this important human pathogen. DNA methylation in the context of *dam* methylation does in fact prove this theory, based on prior publications. However, our exhaustive search with UTI89 strains with altered Type I methylation revealed no effect. Although disappointing, this supported the alternative hypothesis that Type I RMS mediated methylation does not have any consequences on *E. coli* physiological processes at all. To prove such a theory, we had to rationally employ non-specific high throughput methods such as RNA-seq and Phenotype Microarray, so as to comprehensively screen for the presence of any effects due to Type I methylation. To lend further credence to this theory, we decided to extend our scrutiny to other pathogenic (CFT073) as well as non-pathogenic (K12) *E. coli* Type I RMSs. It is important to note that all three RMSs are fully functional with 427 to 754 sites per corresponding genome, however they differ in their methylation specificity, genomic context and the clinical manifestations of their host bacterium. Results from PM, a technique which can simultaneously screen a diverse set of nutritional and resistance phenotypes, supported this theory. Barring cryptic phage transcription in K12 during log phase, no other differences in gene expression were observed. These represent small gene expression changes (only 4 genes) at the same genomic locus and the biological consequences of which remain to be seen. Based on our thorough systematic approach using an extensive set of simple specific as well as high throughput molecular assays, we can convincingly say that Type I RMS mediated methylation does not have any consequences on UPEC model organism UTI89. This observation holds true to a large extent for Type I RMS mediated methylation in other *E. coli* as well.

On the basis of current literature and our study pertaining to the effects of DNA methylation on bacterial cellular processes, and specifically the effects of orphan *dam* methylation versus Type I RMS mediated methylation in UPEC, certain patterns emerge. **1.** DNA methylation has

the ability to alter bacterial transcription and hence phenotype. This is not a eukaryotic specific mechanism [92]. **2.** The vast majority of information pertaining to methylation affecting global gene expression is in the context of orphan methyltransferases. Orphan methyltransferases are highly conserved and sometimes even essential genes, with no role in host defence [75]. RMSs on the other hand, not only show high prevalence among bacterial and archaeal species, but are also highly diversified in their methylation specificities [53, 138]. Amongst Type II methyltransferases, orphans outnumber RMS associated genes [76]. **3.** RMS associated methyltransferases can also affect gene expression and virulence, for example Type III phasevarions and complex Type I phasevariable systems of *S. pneumoniae*. However, these systems harbour the unique ability to rapidly and reversibly alter methylation, from a simple binary ON/OFF system to a complex six-phase switching of methylation patterns, respectively. Instances of classical RMSs similarly regulating virulence are extremely rare, such as *E. coli* O104:H4 strain harbouring a phage encoded classical RMS [88]. However, this system was very recently acquired via phage transduction and since RMS site avoidance correlates with the lifespan of that particular system [52], the long term evolutionary consequences of this classical RMS methylation are unknown. **4.** The primary molecular mechanism involved in methylation mediated gene regulation is the differential binding of regulatory proteins to methylated versus unmethylated sites in the promoters of target genes. Based on this information, studies have looked at the localisation of methylation target sites in gene regulatory elements and the presence of consistently unmethylated target sites, potentially indicative of regulatory protein binding. A recent study involving 230 different prokaryotic genomes, revealed that methylation target sites in regulatory regions and unmethylated target sites are a frequent feature of orphan Type II methyltransferases and not Type II RMSs [76]. **5.** Global gene expression changes involving large numbers of genes correlates with the number of methylation sites in the genome. For example, Type II (palindromic) and III (short asymmetric) sequences occur more frequently than Type I (bipartite) sequences in bacterial genomes. Correspondingly, gene expression changes due to Type II (for example *dam* [97] and phage encoded Type II RMS in *E. coli* O104:H4 [88]) and

Type III (for example phasevarions [123]) methyltransferases are more diverse and include several genes belonging to diverse operons located across the genome. Type I methylation mediated effects are fewer in number and locus specific ([58] and this study).

Due to the paucity of information especially regarding Type I RMSs, the above points are merely observations and by no means conclusive biological rules. Our study aims to address this lack of information regarding any additional roles for Type I RMSs using Uropathogenic *E. coli* as a model. Also, in order to simulate the sudden acquisition of a foreign methylation, for example via phage transduction, we included UTI89 strains expressing K12 and CFT073 Type I methylation patterns. As most studies classify methyltransferases as orphan or RMS associated on the basis of gene content which may not be accurate, we chose to functionally test the existence of both restriction and modification functions of the Type I RMSs in question. The rational and thorough approach of this study sheds some light on the question of whether epigenetic DNA modification mediated by Type I RMSs regulates Uropathogenic *E. coli* virulence, besides its importance for host defence.

Table 11: List of Strains

Strain	Genotype/Relevant characteristic	Source/Reference
UTI89	Clinical isolate	[150]
K12 substrain MG1655	Gut derived lab strain	[184]
CFT073	Clinical isolate	[34]
KSM2-102-4	UTI89 <i>hsdSMR::neo</i>	This study
KSM2-102-7	K12 <i>hsdSMR::neo</i>	This study
KSM7-15-1	CFT073 <i>hsdSMR::neo</i>	This study
KSM6-21-4	UTI89 <i>hsdS::neo-P_{rhaB}-relE</i>	This study
KSM6-26-1	UTI89 <i>hsdS^{K12}</i>	This study
KSM3-95-4	UTI89 <i>hsdS^{UTI89}</i>	This study
KSM6-74-1	UTI89 <i>hsdSMR::neo-P_{rhaB}-relE</i>	This study
KSM6-80-15	UTI89 <i>hsdSMR^{CFT073}</i>	This study
KSM3-39-2	UTI89 <i>att_{HK022}::neo</i>	This study
KSM3-39-1	UTI89 <i>att_{HK022}::cat</i>	This study
KSM3-39-6	UTI89 <i>hsdSMR::FRT att_{HK022}::neo</i>	This study
KSM3-39-3	UTI89 <i>hsdSMR::FRT att_{HK022}::cat</i>	This study
KSM6-29-3	UTI89 <i>hsdS^{K12} att_{HK022}::neo</i>	This study
KSM6-29-1	UTI89 <i>hsdS^{K12} att_{HK022}::cat</i>	This study
KSM5-24-7	UTI89 <i>fliC::neo</i>	This study
KSM3-52-1	UTI89 <i>hsdS::neo-P_{rhaB}-relE</i>	This study
KSM3-61-6	UTI89 <i>hsdS^{K12}</i>	This study
KSM5-72-4	UTI89 <i>hsdS^{K12} att_{HK022}::neo</i>	This study
KSM5-72-5	UTI89 <i>hsdS^{K12} att_{HK022}::cat</i>	This study
KSM7-10-1	UTI89 <i>ybaL::neo</i>	This study
KSM7-12-1	UTI89 <i>yfaL::neo-P_{rhaB}-relE</i>	This study
KSM7-13-2	UTI89 <i>nrdA-K584N</i>	This study

Table 12: List of Plasmids

Plasmid	Genotype/Relevant characteristic	Source/Reference
pKM208	Red recombinase	[185]
pKD3	<i>cat</i> cassette	[147]
pKD4	<i>neo</i> cassette	[147]
pSLC217	<i>neo-P_{rhaB}-relE</i> cassette	[148]
pACYC184	Cloning vector, No Type I sites	New England Biolabs
pKSM3-42-1	1 copy K12 Type I site 5'-AAC (N) ₆ GTGC -3'	This study
pKSM3-45-1	2 copies K12 Type I site 5'-AAC (N) ₆ GTGC -3'	This study
pKSM3-16-1	1 copy UTI89 Type I site 5'-GAAG (N) ₇ TGG -3'	This study
pKSM3-17-1	2 copies UTI89 Type I site 5'-GAAG (N) ₇ TGG -3'	This study
pKSM4-95-4	1 copy CFT073 Type I site 5'-GAG (N) ₇ GTCA -3'	This study
pKSM4-99-1	2 copies CFT073 Type I site 5'-GAG (N) ₇ GTCA -3'	This study

Table 13: List of Primers

Sr. No.	Sequence	Purpose
1	5'TCAGGACTTTTTACGCGAGGCTTTTTACCCCCGCTGGCTGCGCGTTC AGGCTTGCAGTGGGCTTACATG	UTI89 <i>hsdS</i> knockout with <i>neo-P_{rhaB}-relE</i> cassette
2	5'ATGAGTGCTGGGAAATTGCCGAGGGGTGGGAACAGATTGAAATA GGCGATCAGAGCAGGATCGACGTCC	
3	5' TCTGCTTCGAACGGTTGTGC	UTI89 <i>hsdS</i> test
4	5' CGCCGAAGAGACGGAAATTG	
5	5'CCGCATCGCGCTAAATACCTGGATATATCATCAGTAAATACAGGGAA AGTCCAGCTAAAAATAGAATAAAATGGG	Amplification of K12 <i>hsdS</i> with flanking UTI89 homology
6	5'AGCGATGGGTGAACTGGTACAGGCGCTGTCTGAACTGGATGCGCTG ATGCGTGAAGTGGGGGCGAGCGATGAGGC	
7	5'TCAGGACTTTTTACGCGAGGCTTTTTACCCCCGCTGGCTGCGCGTTC AGGCTTGCAGTGGGCTTACATG	UTI89 <i>hsdSMR</i> knockout with <i>neo-P_{rhaB}-relE</i> cassette
8	5'ATGAATAAATCCAACCTCGAATTCCTGAAGGGCGTCAACGACTTCAC TTATTCAGAGCAGGATCGACGTCC	
9	5' TCTGCTTCGAACGGTTGTGC	UTI89 <i>hsdSMR</i> test
10	5' GTCATTGCCCGGAAAGGTAC	
11	5'GCGTAAAAATAATAATAGCGCACCAGAAAGGTGCGCCAGAAAATA ATGATTAGTTTATCGCCGCTCAGTGAGTG	Amplification of CFT073 <i>hsdSMR</i> with flanking UTI89 homology
12	5'GGGCCTAAATATTTGGACAGGCCACACAGCAATGGATTAATAACA ATGATGGCGGAACTGAACCTAAGTAACC	
13	5'CAGGACTTTTTACGCGAGGCTTTTTACCCCCGCTGGCTGCGCGTTCA GCGTGTAGGCTGGAGCTGCTTC	UTI89 <i>hsdSMR</i> knockout with <i>neo</i> cassette
14	5'TGAATAAATCCAACCTCGAATTCCTGAAGGGCGTCAACGACTTCACT TATCATATGAATATCCTCCTTAG	
15	5'ATAAAAAGCGCACCGGAAAGGTGCGCCAGAAAATAATGTTTCAGGAT TTTTGTGTAGGCTGGAGCTGCTTC	K12 <i>hsdSMR</i> knockout with <i>neo</i> cassette
16	5'TGATGAATAAATCCAATTTTGAATTCCTGAAGGGCGTCAACGACTTC ACTCATATGAATATCCTCCTTAG	
17	5' TTAGCGATGCGGGTGTGTTG	K12 <i>hsdSMR</i> test
18	5' GGTCATTGCCCGGAAAGGTA	

19	5'TTAGTTTATCGCCGCGTCAGTGAGTGCCTGCAAGGTGCAGTTGG GTTTGTGTAGGCTGGAGCTGCTC	CFT073 <i>hsdSMR</i> knockout with <i>neo</i> cassette
20	5'ATGGCGGAACTGAACCTAAGTAACCTGACGGAAGCAGACATCATT CCAACATATGAATATCCTCCTAG	
21	5' GTTATTTGCCGGGAGACTTCTGCTTC	CFT073 <i>hsdSMR</i> test
22	5' CGGTTCTGTAACCTTCTGTTTATCTGTC	
23	5'TCAGAATCAGTGCCAAAGAGAACTAATCCCAGCAATAACAGGCT GTATGTGTAGGCTGGAGCTGCTC	UTI89 <i>att_{HK022}</i> insertion with <i>neo</i> and <i>cat</i> cassette
24	5'CGAGCGGCGGATATGTTGCGGTCGGCATTTCGCGTCATGACTA AAACATATGAATATCCTCCTAG	
25	5' TAAACCACGCGCCAGAGGAT	UTI89 <i>att_{HK022}</i> test
26	5' ACGAAATCCCCTGGTGACA	
27	5'TTAACCTGCAGCAGAGACAGAACCTGCTGCGGTACCTGGTTGGCTT TTGGTGTAGGCTGGAGCTGCTC	UTI89 <i>fliC</i> knockout with <i>neo</i> cassette
28	5'ATGGCACAAGTCATTAATACCAACAGCCTCTCGCTGATCACTCAAAA TAACATATGAATATCCTCCTAG	
29	5' AATTTGGCGTTGCCGTCAGT	UTI89 <i>fliC</i> test
30	5' AGCGGAATAAGGGGCAGAG	
31	5'CGCGTCTTATCAGGCCTACATATCCGCGCCTATTATCCGGAACCACC GTGTGTAGGCTGGAGCTGCTC	UTI89 <i>ybaL</i> knockout with <i>neo</i> cassette
32	5'TCGATTATGCTCAACCCGGTACTGTTGCACTACTGGAGAAATATCT GGCCATATGAATATCCTCCTAG	
33	5' CTCAATGCCTGATGCGACG	UTI89 <i>ybaL</i> test
34	5' CTTCTGTTTTCCGATCCGTGC	
35	5'TATCCCGAGGCGCTGTTATTTTCATCAATCATCACACCATTGTTTTGG TCGCTTGCAGTGGGCTTACATG	UTI89 <i>yfaL</i> knockout with <i>neo-P_{rhaB}-relE</i> cassette
36	5'TGATGATGCTTTTAATAATAACCAGGCATATACATCAACAAGTTATA GTGTCAGAGCAGGATCGACGTCC	
37	5' ACGGCAGCGCCAACAACCTGACGCA	UTI89 <i>yfaL</i> test & amplification of <i>nrdA</i> -K584N with flanking UTI89 <i>yfaL</i> homology
38	5' ATGGCACTTATGTGCTATCCG	
39	5'ATTGCAGCGCCGCGAAGGCCTACATGGATTTATTATTCTGCGAAG TG	pACYC184 UTI89 1st Type I site insertion

40	5' AGCTATGCGGCCGCATAAAATCCTGGTGTCCCTGTTGATAACC	
41	5'ATTGCAGCGGCCGCGAAGACGTTTCATGGCCGGACGCATCGTGGCCG GCATCA	pACYC184 UTI89 2nd Type I site insertion
42	5' AGCTATGCGGCCGCCGTAGAGGATCCACAGGACGGGTGTGG	
43	5'ATTGCAGCGGCCGCAACGGTACCGTGCATTTATTTATTCTGCGAAGT G	pACYC184 K12 1st Type I site insertion
44	5' AGCTATGCGGCCGCATAAAATCCTGGTGTCCCTGTTGATAACC	
45	5'ATTGCAGCGGCCGCAACGGTACCGTGCCCGGACGCATCGTGGCCGG CATCA	pACYC184 K12 2nd Type I site insertion
46	5'AGCTATGCGGCCGCCGTAGAGGATCCACAGGACGGGTGTGG	
47	5'ATTGCAGCGGCCGCGAGCATATGAGTCAATTTATTTATTCTGCGAAG TG	pACYC184 CFT073 1st Type I site insertion
48	5' AGCTATGCGGCCGCATAAAATCCTGGTGTCCCTGTTGATAACC	
49	5'ATTGCACTCGAGGAGACGCGTAGTCAACTTCGGGCTCATGAGCGCT TGTTTC	pACYC184 CFT073 2nd Type I site insertion
50	5' AGCTATGCGGCCGCCGTAGAGGATCCACAGGACGGGTGTGG	
51	5' CGGTCTGGTTATAGGTACATTGAGC	Sequencing pACYC184 1st Type I site insertion
52	5' AGAGATTACGCGCAGACCAAAACG	Sequencing pACYC184 2nd Type I site insertion
53	5' CTCTTGCCATCGGATGTGCCCA	UTI89/K12 <i>rrsA</i> qPCR
54	5' CCAGTGTGGCTGGTCATCCTCTC	
55	5' GGTGATCAAAGCCTACCGTTTA	UTI89 <i>fliC</i> qPCR
56	5' CAACAAACCGCACCAATGTC	
57	5' TGAAGTGGCACAGGCAATAG	UTI89 <i>fliA</i> qPCR
58	5' TCGGCAATATCGATCCCTAAAC	
59	5' CCTTCATGCTGATGTGGGTAA	UTI89 <i>fliC</i> qPCR
60	5' GCTTCCACCGATGGTGATATT	
61	5' CTATGAGCTGCTGTGGGAAA	UTI89 <i>nrdA</i> qPCR
62	5' GACGGATCGTAGTTGGTGTAG	
63	5' CAGACGCCGACACTGAAAT	UTI89 C1127 qPCR

64	5' AACCAGCCATCCCACATAAC	
65	5' GTGGGTACAGCCTTCTGTTATC	K12 <i>flu</i> qPCR
66	5' GTCCAGTGATGTGCCATTCT	
67	5' GAATTCCGGGTGCTGTATCT	K12 <i>yeoS</i> qPCR
68	5' ACAGGGCGCGTTTAATCA	
69	5' ACGCGGACAAACAATATCC	K12 <i>prfF</i> qPCR
70	5' CTCCCAGTCGGCACATAAATAC	
71	5' GTTGCCGAAGTCGAGACATTA	K12 <i>yhaV</i> qPCR
72	5' GACCGTGATATGCTCCTCAATC	
73	5' GCCGGACGTCCCAATAAA	K12 <i>emrD</i> qPCR
74	5' CAGACATTCATCACGCCAAAC	
75	5' GGAAGGACGCACACTCTTT	K12 <i>yjhH</i> qPCR
76	5' CTGAACATCACGCCACTCTT	
77	5' TTTCAGTCTTACGCGCTCTATC	CFT073 <i>malK</i> qPCR
78	5' ACTCGCTGGTTAATCACTTCTT	
79	5' ACCGGCGACAAGAACAAT	CFT073 <i>lamB</i> qPCR
80	5' TAGGTTGCGAAGACACGAATAG	

REFERENCES

1. Flores-Mireles, A.L., et al., *Urinary tract infections: epidemiology, mechanisms of infection and treatment options*. Nat Rev Microbiol, 2015. **13**(5): p. 269-84.
2. Foxman, B., *The epidemiology of urinary tract infection*. Nat Rev Urol, 2010. **7**(12): p. 653-60.
3. Bower, J.M., D.S. Eto, and M.A. Mulvey, *Covert operations of uropathogenic Escherichia coli within the urinary tract*. Traffic, 2005. **6**(1): p. 18-31.
4. Foxman, B., *Epidemiology of urinary tract infections: incidence, morbidity, and economic costs*. Dis Mon, 2003. **49**(2): p. 53-70.
5. Glover, M., et al., *Recurrent urinary tract infections in healthy and nonpregnant women*. Urol Sci, 2014. **25**(1): p. 1-8.
6. Russo, T.A., et al., *Chromosomal restriction fragment length polymorphism analysis of Escherichia coli strains causing recurrent urinary tract infections in young women*. J Infect Dis, 1995. **172**(2): p. 440-5.
7. Ejrnaes, K., *Bacterial characteristics of importance for recurrent urinary tract infections caused by Escherichia coli*. Dan Med Bull, 2011. **58**(4): p. B4187.
8. Mulvey, M.A., et al., *Induction and evasion of host defenses by type 1-piliated uropathogenic Escherichia coli*. Science, 1998. **282**(5393): p. 1494-7.
9. Martinez, J.J., et al., *Type 1 pilus-mediated bacterial invasion of bladder epithelial cells*. EMBO J, 2000. **19**(12): p. 2803-12.
10. Dhakal, B.K. and M.A. Mulvey, *Uropathogenic Escherichia coli invades host cells via an HDAC6-modulated microtubule-dependent pathway*. J Biol Chem, 2009. **284**(1): p. 446-54.
11. Eto, D.S., et al., *Clathrin, AP-2, and the NPXY-binding subset of alternate endocytic adaptors facilitate FimH-mediated bacterial invasion of host cells*. Cell Microbiol, 2008. **10**(12): p. 2553-67.
12. Song, J., et al., *TLR4-mediated expulsion of bacteria from infected bladder epithelial cells*. Proc Natl Acad Sci U S A, 2009. **106**(35): p. 14966-71.
13. Ulett, G.C., et al., *Uropathogenic Escherichia coli virulence and innate immune responses during urinary tract infection*. Curr Opin Microbiol, 2013. **16**(1): p. 100-7.
14. Justice, S.S., et al., *Differentiation and developmental pathways of uropathogenic Escherichia coli in urinary tract pathogenesis*. Proc Natl Acad Sci U S A, 2004. **101**(5): p. 1333-8.
15. Anderson, G.G., et al., *Intracellular bacterial communities of uropathogenic Escherichia coli in urinary tract pathogenesis*. Trends Microbiol, 2004. **12**(9): p. 424-30.
16. Anderson, G.G., et al., *Intracellular bacterial biofilm-like pods in urinary tract infections*. Science, 2003. **301**(5629): p. 105-7.
17. Mulvey, M.A., J.D. Schilling, and S.J. Hultgren, *Establishment of a persistent Escherichia coli reservoir during the acute phase of a bladder infection*. Infect Immun, 2001. **69**(7): p. 4572-9.
18. Eto, D.S., J.L. Sundsbak, and M.A. Mulvey, *Actin-gated intracellular growth and resurgence of uropathogenic Escherichia coli*. Cell Microbiol, 2006. **8**(4): p. 704-17.
19. Mysorekar, I.U. and S.J. Hultgren, *Mechanisms of uropathogenic Escherichia coli persistence and eradication from the urinary tract*. Proc Natl Acad Sci U S A, 2006. **103**(38): p. 14170-5.
20. Hannan, T.J., et al., *Host-pathogen checkpoints and population bottlenecks in persistent and intracellular uropathogenic Escherichia coli bladder infection*. FEMS Microbiol Rev, 2012. **36**(3): p. 616-48.
21. Nielubowicz, G.R. and H.L. Mobley, *Host-pathogen interactions in urinary tract infection*. Nat Rev Urol, 2010. **7**(8): p. 430-41.
22. Garofalo, C.K., et al., *Escherichia coli from urine of female patients with urinary tract infections is competent for intracellular bacterial community formation*. Infect Immun, 2007. **75**(1): p. 52-60.

23. Rosen, D.A., et al., *Detection of intracellular bacterial communities in human urinary tract infection*. PLoS Med, 2007. **4**(12): p. e329.
24. Bahrani-Mougeot, F.K., et al., *Type 1 fimbriae and extracellular polysaccharides are preeminent uropathogenic Escherichia coli virulence determinants in the murine urinary tract*. Mol Microbiol, 2002. **45**(4): p. 1079-93.
25. Connell, I., et al., *Type 1 fimbrial expression enhances Escherichia coli virulence for the urinary tract*. Proc Natl Acad Sci U S A, 1996. **93**(18): p. 9827-32.
26. Garcia, E.C., A.R. Brumbaugh, and H.L. Mobley, *Redundancy and specificity of Escherichia coli iron acquisition systems during urinary tract infection*. Infect Immun, 2011. **79**(3): p. 1225-35.
27. Wiles, T.J., R.R. Kulesus, and M.A. Mulvey, *Origins and virulence mechanisms of uropathogenic Escherichia coli*. Exp Mol Pathol, 2008. **85**(1): p. 11-9.
28. Wiles, T.J., et al., *Inactivation of host Akt/protein kinase B signaling by bacterial pore-forming toxins*. Mol Biol Cell, 2008. **19**(4): p. 1427-38.
29. Wright, K.J., P.C. Seed, and S.J. Hultgren, *Uropathogenic Escherichia coli flagella aid in efficient urinary tract colonization*. Infect Immun, 2005. **73**(11): p. 7657-68.
30. Lane, M.C., et al., *Expression of flagella is coincident with uropathogenic Escherichia coli ascension to the upper urinary tract*. Proc Natl Acad Sci U S A, 2007. **104**(42): p. 16669-74.
31. Lane, M.C., et al., *Role of motility in the colonization of uropathogenic Escherichia coli in the urinary tract*. Infect Immun, 2005. **73**(11): p. 7644-56.
32. Rasko, D.A., et al., *The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates*. J Bacteriol, 2008. **190**(20): p. 6881-93.
33. Brzuszkiewicz, E., et al., *How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic Escherichia coli strains*. Proc Natl Acad Sci U S A, 2006. **103**(34): p. 12879-84.
34. Welch, R.A., et al., *Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli*. Proc Natl Acad Sci U S A, 2002. **99**(26): p. 17020-4.
35. Rodriguez-Siek, K.E., et al., *Comparison of Escherichia coli isolates implicated in human urinary tract infection and avian colibacillosis*. Microbiology, 2005. **151**(Pt 6): p. 2097-110.
36. Barber, A.E., et al., *Urinary tract infections: current and emerging management strategies*. Clin Infect Dis, 2013. **57**(5): p. 719-24.
37. Blango, M.G. and M.A. Mulvey, *Persistence of uropathogenic Escherichia coli in the face of multiple antibiotics*. Antimicrob Agents Chemother, 2010. **54**(5): p. 1855-63.
38. Schilling, J.D., R.G. Lorenz, and S.J. Hultgren, *Effect of trimethoprim-sulfamethoxazole on recurrent bacteriuria and bacterial persistence in mice infected with uropathogenic Escherichia coli*. Infect Immun, 2002. **70**(12): p. 7042-9.
39. Greene, S.E., et al., *Pilicide ec240 disrupts virulence circuits in uropathogenic Escherichia coli*. MBio, 2014. **5**(6): p. e02038.
40. Cusumano, C.K., et al., *Treatment and prevention of urinary tract infection with orally active FimH inhibitors*. Sci Transl Med, 2011. **3**(109): p. 109ra115.
41. Totsika, M., et al., *A FimH inhibitor prevents acute bladder infection and treats chronic cystitis caused by multidrug-resistant uropathogenic Escherichia coli ST131*. J Infect Dis, 2013. **208**(6): p. 921-8.
42. Blango, M.G., et al., *Forced resurgence and targeting of intracellular uropathogenic Escherichia coli reservoirs*. PLoS One, 2014. **9**(3): p. e93327.
43. Langermann, S., et al., *Prevention of mucosal Escherichia coli infection by FimH-adhesin-based systemic vaccination*. Science, 1997. **276**(5312): p. 607-11.
44. Langermann, S., et al., *Vaccination with FimH adhesin protects cynomolgus monkeys from colonization and infection by uropathogenic Escherichia coli*. J Infect Dis, 2000. **181**(2): p. 774-8.
45. Sivick, K.E. and H.L. Mobley, *Waging war against uropathogenic Escherichia coli: winning back the urinary tract*. Infect Immun, 2010. **78**(2): p. 568-85.

46. Mathers, A.J., G. Peirano, and J.D. Pitout, *The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant Enterobacteriaceae*. Clin Microbiol Rev, 2015. **28**(3): p. 565-91.
47. Davis, B.M., M.C. Chao, and M.K. Waldor, *Entering the era of bacterial epigenomics with single molecule real time DNA sequencing*. Curr Opin Microbiol, 2013. **16**(2): p. 192-8.
48. Jones, P.A., *Functions of DNA methylation: islands, start sites, gene bodies and beyond*. Nat Rev Genet, 2012. **13**(7): p. 484-92.
49. Heyn, H. and M. Esteller, *An Adenine Code for DNA: A Second Life for N6-Methyladenine*. Cell, 2015. **161**(4): p. 710-3.
50. Smith, Z.D. and A. Meissner, *DNA methylation: roles in mammalian development*. Nat Rev Genet, 2013. **14**(3): p. 204-20.
51. Loenen, W.A., *Tracking EcoKI and DNA fifty years on: a golden story full of surprises*. Nucleic Acids Res, 2003. **31**(24): p. 7059-69.
52. Ershova, A.S., et al., *Role of Restriction-Modification Systems in Prokaryotic Evolution and Ecology*. Biochemistry (Mosc), 2015. **80**(10): p. 1373-86.
53. Vasu, K. and V. Nagaraja, *Diverse functions of restriction-modification systems in addition to cellular defense*. Microbiol Mol Biol Rev, 2013. **77**(1): p. 53-72.
54. Murray, N.E., *Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle)*. Microbiol Mol Biol Rev, 2000. **64**(2): p. 412-34.
55. Studier, F.W. and P.K. Bandyopadhyay, *Model for how type I restriction enzymes select cleavage sites in DNA*. Proc Natl Acad Sci U S A, 1988. **85**(13): p. 4677-81.
56. Loenen, W.A., et al., *Type I restriction enzymes and their relatives*. Nucleic Acids Res, 2014. **42**(1): p. 20-44.
57. Kim, J.S., et al., *Crystal structure of DNA sequence specificity subunit of a type I restriction-modification enzyme and its functional implications*. Proc Natl Acad Sci U S A, 2005. **102**(9): p. 3248-53.
58. Furuta, Y., et al., *Methylome diversification through changes in DNA methyltransferase sequence specificity*. PLoS Genet, 2014. **10**(4): p. e1004272.
59. Dybvig, K., R. Sitaraman, and C.T. French, *A family of phase-variable restriction enzymes with differing specificities generated by high-frequency gene rearrangements*. Proc Natl Acad Sci U S A, 1998. **95**(23): p. 13923-8.
60. Sitaraman, R. and K. Dybvig, *The hsd loci of Mycoplasma pulmonis: organization, rearrangements and expression of genes*. Mol Microbiol, 1997. **26**(1): p. 109-20.
61. Cerdeno-Tarraga, A.M., et al., *Extensive DNA inversions in the B. fragilis genome control variable gene expression*. Science, 2005. **307**(5714): p. 1463-5.
62. Li, J., et al., *Epigenetic Switch Driven by DNA Inversions Dictates Phase Variation in Streptococcus pneumoniae*. PLoS Pathog, 2016. **12**(7): p. e1005762.
63. Manso, A.S., et al., *A random six-phase switch regulates pneumococcal virulence via global epigenetic changes*. Nat Commun, 2014. **5**: p. 5055.
64. Pingoud, A., G.G. Wilson, and W. Wende, *Type II restriction endonucleases--a historical perspective and more*. Nucleic Acids Res, 2014. **42**(12): p. 7489-527.
65. Kobayashi, I., *Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution*. Nucleic Acids Res, 2001. **29**(18): p. 3742-56.
66. Morozova, N., et al., *Temporal dynamics of methyltransferase and restriction endonuclease accumulation in individual cells after introducing a restriction-modification system*. Nucleic Acids Res, 2016. **44**(2): p. 790-800.
67. Rao, D.N., D.T. Dryden, and S. Bheemanaik, *Type III restriction-modification enzymes: a historical perspective*. Nucleic Acids Res, 2014. **42**(1): p. 45-55.
68. Tock, M.R. and D.T. Dryden, *The biology of restriction and anti-restriction*. Curr Opin Microbiol, 2005. **8**(4): p. 466-72.

69. Gawthorne, J.A., et al., *Origin of the diversity in DNA recognition domains in phasevarion associated modA genes of pathogenic Neisseria and Haemophilus influenzae*. PLoS One, 2012. **7**(3): p. e32337.
70. Loenen, W.A. and E.A. Raleigh, *The other face of restriction: modification-dependent enzymes*. Nucleic Acids Res, 2014. **42**(1): p. 56-69.
71. Ishikawa, K., E. Fukuda, and I. Kobayashi, *Conflicts targeting epigenetic systems and their resolution by cell death: novel concepts for methyl-specific and other restriction systems*. DNA Res, 2010. **17**(6): p. 325-42.
72. Chao, M.C., et al., *A Cytosine Methyltransferase Modulates the Cell Envelope Stress Response in the Cholera Pathogen [corrected]*. PLoS Genet, 2015. **11**(11): p. e1005666.
73. Campellone, K.G., et al., *Increased adherence and actin pedestal formation by dam-deficient enterohaemorrhagic Escherichia coli O157:H7*. Mol Microbiol, 2007. **63**(5): p. 1468-81.
74. Kahramanoglou, C., et al., *Genomics of DNA cytosine methylation in Escherichia coli reveals its role in stationary phase transcription*. Nat Commun, 2012. **3**: p. 886.
75. Seshasayee, A.S., P. Singh, and S. Krishna, *Context-dependent conservation of DNA methyltransferases in bacteria*. Nucleic Acids Res, 2012. **40**(15): p. 7066-73.
76. Blow, M.J., et al., *The Epigenomic Landscape of Prokaryotes*. PLoS Genet, 2016. **12**(2): p. e1005854.
77. Frommer, M., et al., *A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands*. Proc Natl Acad Sci U S A, 1992. **89**(5): p. 1827-31.
78. Eid, J., et al., *Real-time DNA sequencing from single polymerase molecules*. Science, 2009. **323**(5910): p. 133-8.
79. Flusberg, B.A., et al., *Direct detection of DNA methylation during single-molecule, real-time sequencing*. Nat Methods, 2010. **7**(6): p. 461-5.
80. Roberts, R.J., M.O. Carneiro, and M.C. Schatz, *The advantages of SMRT sequencing*. Genome Biol, 2013. **14**(7): p. 405.
81. Chin, C.S., et al., *Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data*. Nat Methods, 2013. **10**(6): p. 563-9.
82. Sanchez-Romero, M.A., I. Cota, and J. Casadesus, *DNA methylation in bacteria: from the methyl group to the methylome*. Curr Opin Microbiol, 2015. **25**: p. 9-16.
83. Pirone-Davies, C., et al., *Genome-wide methylation patterns in Salmonella enterica Subsp. enterica Serovars*. PLoS One, 2015. **10**(4): p. e0123639.
84. Murray, I.A., et al., *The methylomes of six bacteria*. Nucleic Acids Res, 2012. **40**(22): p. 11450-62.
85. Xu, S.Y., et al., *Characterization of type II and III restriction-modification systems from Bacillus cereus strains ATCC 10987 and ATCC 14579*. J Bacteriol, 2012. **194**(1): p. 49-60.
86. Anjum, A., et al., *Phase variation of a Type IIG restriction-modification enzyme alters site-specific methylation patterns and gene expression in Campylobacter jejuni strain NCTC11168*. Nucleic Acids Res, 2016. **44**(10): p. 4581-94.
87. Krebes, J., et al., *The complex methylome of the human gastric pathogen Helicobacter pylori*. Nucleic Acids Res, 2014. **42**(4): p. 2415-32.
88. Fang, G., et al., *Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing*. Nat Biotechnol, 2012. **30**(12): p. 1232-9.
89. Zhu, L., et al., *Precision methylome characterization of Mycobacterium tuberculosis complex (MTBC) using PacBio single-molecule real-time (SMRT) technology*. Nucleic Acids Res, 2016. **44**(2): p. 730-43.
90. Kozdon, J.B., et al., *Global methylation state at base-pair resolution of the Caulobacter genome throughout the cell cycle*. Proc Natl Acad Sci U S A, 2013. **110**(48): p. E4658-67.

91. Casadesus, J. and D. Low, *Epigenetic gene regulation in the bacterial world*. Microbiol Mol Biol Rev, 2006. **70**(3): p. 830-56.
92. Wion, D. and J. Casadesus, *N6-methyl-adenine: an epigenetic signal for DNA-protein interactions*. Nat Rev Microbiol, 2006. **4**(3): p. 183-92.
93. Coffin, S.R. and N.O. Reich, *Modulation of Escherichia coli DNA methyltransferase activity by biologically derived GATC-flanking sequences*. J Biol Chem, 2008. **283**(29): p. 20106-16.
94. Low, D.A. and J. Casadesus, *Clocks and switches: bacterial gene regulation by DNA adenine methylation*. Curr Opin Microbiol, 2008. **11**(2): p. 106-12.
95. Cohen, N.R., et al., *A role for the bacterial GATC methylome in antibiotic stress survival*. Nat Genet, 2016. **48**(5): p. 581-6.
96. Heusipp, G., S. Falker, and M.A. Schmidt, *DNA adenine methylation and bacterial pathogenesis*. Int J Med Microbiol, 2007. **297**(1): p. 1-7.
97. Oshima, T., et al., *Genome-wide analysis of deoxyadenosine methyltransferase-mediated control of gene expression in Escherichia coli*. Mol Microbiol, 2002. **45**(3): p. 673-95.
98. Lobner-Olesen, A., M.G. Marinus, and F.G. Hansen, *Role of SeqA and Dam in Escherichia coli gene expression: a global/microarray analysis*. Proc Natl Acad Sci U S A, 2003. **100**(8): p. 4672-7.
99. Hernday, A., et al., *Self-perpetuating epigenetic pili switches in bacteria*. Proc Natl Acad Sci U S A, 2002. **99** Suppl 4: p. 16470-6.
100. Lim, H.N. and A. van Oudenaarden, *A multistep epigenetic switch enables the stable inheritance of DNA methylation states*. Nat Genet, 2007. **39**(2): p. 269-75.
101. Heithoff, D.M., et al., *An essential role for DNA adenine methylation in bacterial virulence*. Science, 1999. **284**(5416): p. 967-70.
102. Garcia-Del Portillo, F., M.G. Pucciarelli, and J. Casadesus, *DNA adenine methylase mutants of Salmonella typhimurium show defects in protein secretion, cell invasion, and M cell cytotoxicity*. Proc Natl Acad Sci U S A, 1999. **96**(20): p. 11578-83.
103. Balbontin, R., et al., *DNA adenine methylation regulates virulence gene expression in Salmonella enterica serovar Typhimurium*. J Bacteriol, 2006. **188**(23): p. 8160-8.
104. Badie, G., et al., *Altered levels of Salmonella DNA adenine methylase are associated with defects in gene expression, motility, flagellar synthesis, and bile resistance in the pathogenic strain 14028 but not in the laboratory strain LT2*. J Bacteriol, 2007. **189**(5): p. 1556-64.
105. Jakomin, M., et al., *Regulation of the Salmonella enterica std fimbrial operon by DNA adenine methylation, SeqA, and HdfR*. J Bacteriol, 2008. **190**(22): p. 7406-13.
106. Cota, I., et al., *OxyR-dependent formation of DNA methylation patterns in OpvABOFF and OpvABON cell lineages of Salmonella enterica*. Nucleic Acids Res, 2016. **44**(8): p. 3595-609.
107. Cota, I., et al., *Epigenetic Control of Salmonella enterica O-Antigen Chain Length: A Tradeoff between Virulence and Bacteriophage Resistance*. PLoS Genet, 2015. **11**(11): p. e1005667.
108. Julio, S.M., et al., *DNA adenine methylase is essential for viability and plays a role in the pathogenesis of Yersinia pseudotuberculosis and Vibrio cholerae*. Infect Immun, 2001. **69**(12): p. 7610-5.
109. Pouillot, F., C. Fayolle, and E. Carniel, *A putative DNA adenine methyltransferase is involved in Yersinia pseudotuberculosis pathogenicity*. Microbiology, 2007. **153**(Pt 8): p. 2426-34.
110. Falker, S., et al., *Overproduction of DNA adenine methyltransferase alters motility, invasion, and the lipopolysaccharide O-antigen composition of Yersinia enterocolitica*. Infect Immun, 2007. **75**(10): p. 4990-7.
111. Falker, S., M.A. Schmidt, and G. Heusipp, *DNA methylation in Yersinia enterocolitica: role of the DNA adenine methyltransferase in mismatch repair and regulation of virulence factors*. Microbiology, 2005. **151**(Pt 7): p. 2291-9.
112. Erova, T.E., et al., *DNA adenine methyltransferase influences the virulence of Aeromonas hydrophila*. Infect Immun, 2006. **74**(1): p. 410-24.

113. Wu, H., et al., *Inactivation of DNA adenine methyltransferase alters virulence factors in Actinobacillus actinomycetemcomitans*. Oral Microbiol Immunol, 2006. **21**(4): p. 238-44.
114. Mehling, J.S., H. Lavender, and S. Clegg, *A Dam methylation mutant of Klebsiella pneumoniae is partially attenuated*. FEMS Microbiol Lett, 2007. **268**(2): p. 187-93.
115. Militello, K.T., et al., *Conservation of Dcm-mediated cytosine DNA methylation in Escherichia coli*. FEMS Microbiol Lett, 2012. **328**(1): p. 78-85.
116. Militello, K.T., et al., *5-azacytidine induces transcriptome changes in Escherichia coli via DNA methylation-dependent and DNA methylation-independent mechanisms*. BMC Microbiol, 2016. **16**(1): p. 130.
117. Militello, K.T., et al., *Cytosine DNA methylation influences drug resistance in Escherichia coli through increased sugE expression*. FEMS Microbiol Lett, 2014. **350**(1): p. 100-6.
118. Gonzalez, D., et al., *The functions of DNA methylation by CcrM in Caulobacter crescentus: a global approach*. Nucleic Acids Res, 2014. **42**(6): p. 3720-35.
119. Shell, S.S., et al., *DNA methylation impacts gene expression and ensures hypoxic survival of Mycobacterium tuberculosis*. PLoS Pathog, 2013. **9**(7): p. e1003419.
120. Kumar, R., et al., *Comparative transcriptomics of H. pylori strains AM5, SS1 and their hpyAVIBM deletion mutants: possible roles of cytosine methylation*. PLoS One, 2012. **7**(8): p. e42303.
121. Chernov, A.V., et al., *Depletion of CG-Specific Methylation in Mycoplasma hyorhinitis Genomic DNA after Host Cell Invasion*. PLoS One, 2015. **10**(11): p. e0142529.
122. Fox, K.L., Y.N. Srikhanta, and M.P. Jennings, *Phase variable type III restriction-modification systems of host-adapted bacterial pathogens*. Mol Microbiol, 2007. **65**(6): p. 1375-9.
123. Srikhanta, Y.N., K.L. Fox, and M.P. Jennings, *The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes*. Nat Rev Microbiol, 2010. **8**(3): p. 196-206.
124. Fox, K.L., et al., *Haemophilus influenzae phasevarions have evolved from type III DNA restriction systems into epigenetic regulators of gene expression*. Nucleic Acids Res, 2007. **35**(15): p. 5242-52.
125. Attack, J.M., et al., *A biphasic epigenetic switch controls immunoevasion, virulence and niche adaptation in non-typeable Haemophilus influenzae*. Nat Commun, 2015. **6**: p. 7828.
126. Srikhanta, Y.N., et al., *Phasevarions mediate random switching of gene expression in pathogenic Neisseria*. PLoS Pathog, 2009. **5**(4): p. e1000400.
127. Kwiatek, A., et al., *Type III Methyltransferase M.NgoAX from Neisseria gonorrhoeae FA1090 Regulates Biofilm Formation and Interactions with Human Cells*. Front Microbiol, 2015. **6**: p. 1426.
128. Jen, F.E., K.L. Seib, and M.P. Jennings, *Phasevarions mediate epigenetic regulation of antimicrobial susceptibility in Neisseria meningitidis*. Antimicrob Agents Chemother, 2014. **58**(7): p. 4219-21.
129. Seib, K.L., et al., *Specificity of the ModA11, ModA12 and ModD1 epigenetic regulator N(6)-adenine DNA methyltransferases of Neisseria meningitidis*. Nucleic Acids Res, 2015. **43**(8): p. 4150-62.
130. Srikhanta, Y.N., et al., *Phasevarion mediated epigenetic gene regulation in Helicobacter pylori*. PLoS One, 2011. **6**(12): p. e27569.
131. Blakeway, L.V., et al., *ModM DNA methyltransferase methylome analysis reveals a potential role for Moraxella catarrhalis phasevarions in otitis media*. FASEB J, 2014. **28**(12): p. 5197-207.
132. Skoglund, A., et al., *Functional analysis of the M.HpyAIV DNA methyltransferase of Helicobacter pylori*. J Bacteriol, 2007. **189**(24): p. 8914-21.
133. Lawrence, J.G., *Gene transfer in bacteria: speciation without species?* Theor Popul Biol, 2002. **61**(4): p. 449-60.
134. Jeltsch, A., *Maintenance of species identity and controlling speciation of bacteria: a new function for restriction/modification systems?* Gene, 2003. **317**(1-2): p. 13-6.

135. Croucher, N.J., et al., *Diversification of bacterial genome content through distinct mechanisms over different timescales*. Nat Commun, 2014. **5**: p. 5471.
136. Sorek, R., V. Kunin, and P. Hugenholtz, *CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea*. Nat Rev Microbiol, 2008. **6**(3): p. 181-6.
137. Goldfarb, T., et al., *BREX is a novel phage resistance system widespread in microbial genomes*. EMBO J, 2015. **34**(2): p. 169-83.
138. Roberts, R.J., et al., *REBASE--a database for DNA restriction and modification: enzymes, genes and genomes*. Nucleic Acids Res, 2015. **43**(Database issue): p. D298-9.
139. Rocha, E.P., A. Danchin, and A. Viari, *Evolutionary role of restriction/modification systems as revealed by comparative genome analysis*. Genome Res, 2001. **11**(6): p. 946-58.
140. Rusinov, I., et al., *Lifespan of restriction-modification systems critically affects avoidance of their recognition sites in host genomes*. BMC Genomics, 2015. **16**: p. 1084.
141. Cockfield, J.D., et al., *Rapid determination of hospital-acquired meticillin-resistant Staphylococcus aureus lineages*. J Med Microbiol, 2007. **56**(Pt 5): p. 614-9.
142. Waldron, D.E. and J.A. Lindsay, *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, 2006. **188**(15): p. 5578-85.
143. Roberts, G.A., et al., *Impact of target site distribution for Type I restriction enzymes on the evolution of methicillin-resistant Staphylococcus aureus (MRSA) populations*. Nucleic Acids Res, 2013. **41**(15): p. 7472-84.
144. Corvaglia, A.R., et al., *A type III-like restriction endonuclease functions as a major barrier to horizontal gene transfer in clinical Staphylococcus aureus strains*. Proc Natl Acad Sci U S A, 2010. **107**(26): p. 11954-8.
145. Nandi, T., et al., *Burkholderia pseudomallei sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles*. Genome Res, 2015. **25**(4): p. 608.
146. Forde, B.M., et al., *Lineage-Specific Methyltransferases Define the Methylome of the Globally Disseminated Escherichia coli ST131 Clone*. MBio, 2015. **6**(6): p. e01602-15.
147. Datsenko, K.A. and B.L. Wanner, *One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products*. Proc Natl Acad Sci U S A, 2000. **97**(12): p. 6640-5.
148. Khetrpal, V., et al., *A set of powerful negative selection systems for unmodified Enterobacteriaceae*. Nucleic Acids Res, 2015. **43**(13): p. e83.
149. Hung, C.S., K.W. Dodson, and S.J. Hultgren, *A murine model of urinary tract infection*. Nat Protoc, 2009. **4**(8): p. 1230-43.
150. Chen, S.L., et al., *Positive selection identifies an in vivo role for FimH during urinary tract infection in addition to mannose binding*. Proc Natl Acad Sci U S A, 2009. **106**(52): p. 22439-44.
151. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform*. Bioinformatics, 2009. **25**(14): p. 1754-60.
152. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data*. Bioinformatics, 2015. **31**(2): p. 166-9.
153. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data*. Bioinformatics, 2010. **26**(1): p. 139-40.
154. Wilm, A., et al., *LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets*. Nucleic Acids Res, 2012. **40**(22): p. 11189-201.
155. Hadjifrangiskou, M., et al., *Transposon mutagenesis identifies uropathogenic Escherichia coli biofilm factors*. J Bacteriol, 2012. **194**(22): p. 6195-205.
156. Bochner, B.R., P. Gadzinski, and E. Panomitros, *Phenotype microarrays for high-throughput phenotypic testing and assay of gene function*. Genome Res, 2001. **11**(7): p. 1246-55.
157. Clark, T.A., et al., *Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation*. BMC Biol, 2013. **11**: p. 4.

158. Titheradge, A.J., et al., *Families of restriction enzymes: an analysis prompted by molecular and genetic data for type ID restriction and modification systems*. Nucleic Acids Res, 2001. **29**(20): p. 4195-205.
159. Anderson, J.K., T.G. Smith, and T.R. Hoover, *Sense and sensibility: flagellum-mediated gene regulation*. Trends Microbiol, 2010. **18**(1): p. 30-7.
160. Tenke, P., et al., *Update on biofilm infections in the urinary tract*. World J Urol, 2012. **30**(1): p. 51-7.
161. Rossignol, M., L. Moulin, and F. Boccard, *Phage HK022-based integrative vectors for the insertion of genes in the chromosome of multiply marked Escherichia coli strains*. FEMS Microbiol Lett, 2002. **213**(1): p. 45-9.
162. Eshaghi, M., K.S. Mehershahi, and S.L. Chen, *Brighter Fluorescent Derivatives of UTI89 Utilizing a Monomeric vGFP*. Pathogens, 2016. **5**(1).
163. Schilling, J.D., M.A. Mulvey, and S.J. Hultgren, *Structure and function of Escherichia coli type 1 pili: new insight into the pathogenesis of urinary tract infections*. J Infect Dis, 2001. **183 Suppl 1**: p. S36-40.
164. Bochner, B.R., *New technologies to assess genotype-phenotype relationships*. Nat Rev Genet, 2003. **4**(4): p. 309-14.
165. Soutourina, O.A. and P.N. Bertin, *Regulation cascade of flagellar expression in Gram-negative bacteria*. FEMS Microbiol Rev, 2003. **27**(4): p. 505-23.
166. Nordlund, P. and P. Reichard, *Ribonucleotide reductases*. Annu Rev Biochem, 2006. **75**: p. 681-706.
167. Ando, N., et al., *Structural interconversions modulate activity of Escherichia coli ribonucleotide reductase*. Proc Natl Acad Sci U S A, 2011. **108**(52): p. 21046-51.
168. Hofer, A., et al., *DNA building blocks: keeping control of manufacture*. Crit Rev Biochem Mol Biol, 2012. **47**(1): p. 50-63.
169. Ahluwalia, D., R.J. Bienstock, and R.M. Schaaper, *Novel mutator mutants of E. coli nrdAB ribonucleotide reductase: insight into allosteric regulation and control of mutation rates*. DNA Repair (Amst), 2012. **11**(5): p. 480-7.
170. Gon, S., et al., *Increase in dNTP pool size during the DNA damage response plays a key role in spontaneous and induced-mutagenesis in Escherichia coli*. Proc Natl Acad Sci U S A, 2011. **108**(48): p. 19311-6.
171. Tholander, F. and B.M. Sjöberg, *Discovery of antimicrobial ribonucleotide reductase inhibitors by screening in microwell format*. Proc Natl Acad Sci U S A, 2012. **109**(25): p. 9798-803.
172. Brignole, E.J., et al., *The prototypic class Ia ribonucleotide reductase from Escherichia coli: still surprising after all these years*. Biochem Soc Trans, 2012. **40**(3): p. 523-30.
173. Torrents, E., et al., *NrdR controls differential expression of the Escherichia coli ribonucleotide reductase genes*. J Bacteriol, 2007. **189**(14): p. 5012-21.
174. Dreux, N., et al., *Ribonucleotide reductase NrdR as a novel regulator for motility and chemotaxis during adherent-invasive Escherichia coli infection*. Infect Immun, 2015. **83**(4): p. 1305-17.
175. Herrick, J. and B. Sclavi, *Ribonucleotide reductase and the regulation of DNA replication: an old story and an ancient heritage*. Mol Microbiol, 2007. **63**(1): p. 22-34.
176. Cendra Mdel, M., et al., *H-NS is a novel transcriptional modulator of the ribonucleotide reductase genes in Escherichia coli*. J Bacteriol, 2013. **195**(18): p. 4255-63.
177. Kelly, A., et al., *A global role for Fis in the transcriptional control of metabolism and type III secretion in Salmonella enterica serovar Typhimurium*. Microbiology, 2004. **150**(Pt 7): p. 2037-53.
178. Kim, E.A. and D.F. Blair, *Function of the Histone-Like Protein H-NS in Motility of Escherichia coli: Multiple Regulatory Roles Rather than Direct Action at the Flagellar Motor*. J Bacteriol, 2015. **197**(19): p. 3110-20.

179. Watson, M.E., Jr., J. Jarisch, and A.L. Smith, *Inactivation of deoxyadenosine methyltransferase (dam) attenuates Haemophilus influenzae virulence*. Mol Microbiol, 2004. **53**(2): p. 651-64.
180. Wallecha, A., et al., *Dam- and OxyR-dependent phase variation of agn43: essential elements and evidence for a new role of DNA methylation*. J Bacteriol, 2002. **184**(12): p. 3338-47.
181. Schwan, W.R., *Regulation of fim genes in uropathogenic Escherichia coli*. World J Clin Infect Dis, 2011. **1**(1): p. 17-25.
182. Kasarjian, J.K., M. Iida, and J. Ryu, *New restriction enzymes discovered from Escherichia coli clinical strains using a plasmid transformation method*. Nucleic Acids Res, 2003. **31**(5): p. e22.
183. Wright, K.J., P.C. Seed, and S.J. Hultgren, *Development of intracellular bacterial communities of uropathogenic Escherichia coli depends on type 1 pili*. Cell Microbiol, 2007. **9**(9): p. 2230-41.
184. Riley, M., et al., *Escherichia coli K-12: a cooperatively developed annotation snapshot--2005*. Nucleic Acids Res, 2006. **34**(1): p. 1-9.
185. Murphy, K.C. and K.G. Campellone, *Lambda Red-mediated recombinogenic engineering of enterohemorrhagic and enteropathogenic E. coli*. BMC Mol Biol, 2003. **4**: p. 11.

PUBLICATIONS

1. Nandi T, Holden MT, Didelot X, **Mehershahi K**, Boddey JA, Beacham I, Peak I, Harting J, Baybayan P, Guo Y, Wang S, How LC, Sim B, Essex-Lopresti A, Sarkar-Tyson M, Nelson M, Smither S, Ong C, Aw LT, Hoon CH, Michell S, Studholme DJ, Titball R, Chen SL, Parkhill J, Tan P. *Burkholderia pseudomallei* sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles. **Genome Research. (2015)**
2. Khetrapal V, **Mehershahi K**, Rafée S, Chen S, Lim CL, Chen SL. A set of powerful negative selection systems for unmodified Enterobacteriaceae. **Nucleic Acids Research. (2015)**
3. **Mehershahi KS**, Abraham SN, Chen SL. Complete Genome Sequence of Uropathogenic *Escherichia coli* Strain CI5. **Genome Announcements. (2015)**
4. **Mehershahi KS**, Hsu LY, Koh TH, Chen SL. et al. Complete Genome Sequence of *Streptococcus agalactiae* Serotype III, Multilocus Sequence Type 283 Strain SG-M1. **Genome Announcements. (2015)**
5. Khetrapal V, **Mehershahi KS**, Chen S, Chen SL. (Co-first authors) Application and optimization of *relE* as a negative selection marker in uropathogenic *Escherichia coli* strain UTI89. **Pathogens. (2016)**
6. Eshaghi M, **Mehershahi KS**, Chen SL. Brighter fluorescent derivatives of UTI89 utilizing a monomeric vGFP. **Pathogens. (2016)**
7. An epidemic of severe *Streptococcus agalactiae* ST283 infections associated with the consumption of raw freshwater fish. (Manuscript submitted to **Clinical Infectious Diseases**)
8. **Mehershahi KS**, Chen S.L. Complete Genome Sequence of Uropathogenic *Escherichia coli* Strain NU14. **Genome Announcements. (2017)**
9. Khetrapal V, **Mehershahi KS**, Chen S.L. Complete genome sequence of the original *Escherichia coli* isolate, strain NCTC86. **Genome Announcements. (2017)**

CONFERENCES

Poster: “Role of Epigenetic Modification in regulating Uropathogenic *E. coli* (UPEC) Virulence”.

Gordon Research Seminar (25-26 July 2015) and Gordon Research Conference (27-30 July 2015)
on Microbial Adhesion and Signal Transduction.

Burkholderia pseudomallei sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles

Tannistha Nandi,¹ Matthew T.G. Holden,^{2,13} Xavier Didelot,³ Kurosh Mehershahi,⁴ Justin A. Boddey,^{5,14} Ifor Beacham,⁵ Ian Peak,⁵ John Harting,⁶ Primo Baybayan,⁶ Yan Guo,⁶ Susana Wang,⁶ Lee Chee How,⁶ Bernice Sim,¹ Angela Essex-Lopresti,⁷ Mitali Sarkar-Tyson,⁷ Michelle Nelson,⁷ Sophie Smither,⁷ Catherine Ong,⁸ Lay Tin Aw,⁸ Chua Hui Hoon,¹ Stephen Michell,⁹ David J. Studholme,⁹ Richard Titball,^{9,10} Swaine L. Chen,^{1,4} Julian Parkhill,² and Patrick Tan^{1,11,12}

¹Genome Institute of Singapore, Singapore, 138672, Republic of Singapore; ²The Wellcome Trust Sanger Institute, Cambridge, CB10 1SA, United Kingdom; ³Department of Infectious Disease Epidemiology, Imperial College London, W2 1PG, United Kingdom; ⁴Department of Medicine, National University of Singapore, Singapore, 119074 Republic of Singapore; ⁵Institute for Glycomics, Griffith University (Gold Coast Campus), Southport, Queensland, QLD 4222, Australia; ⁶Pacific Biosciences, Menlo Park, California 94025, USA; ⁷Defence Science and Technology Laboratory, Porton Down, Salisbury, SP4 0JQ, United Kingdom; ⁸Defense Medical and Environmental Research Institute, DSO National Laboratories, Singapore, 117510, Republic of Singapore; ⁹Biosciences, University of Exeter, Exeter, EX4 4QD, United Kingdom; ¹⁰Faculty of Infectious and Tropical Diseases, Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, WC1E 7HT, United Kingdom; ¹¹Duke-NUS Graduate Medical School Singapore, Singapore, 169857, Republic of Singapore; ¹²Cancer Science Institute of Singapore, National University of Singapore, 117599, Republic of Singapore

Burkholderia pseudomallei (Bp) is the causative agent of the infectious disease melioidosis. To investigate population diversity, recombination, and horizontal gene transfer in closely related Bp isolates, we performed whole-genome sequencing (WGS) on 106 clinical, animal, and environmental strains from a restricted Asian locale. Whole-genome phylogenies resolved multiple genomic clades of Bp, largely congruent with multilocus sequence typing (MLST). We discovered widespread recombination in the Bp core genome, involving hundreds of regions associated with multiple haplotypes. Highly recombinant regions exhibited functional enrichments that may contribute to virulence. We observed clade-specific patterns of recombination and accessory gene exchange, and provide evidence that this is likely due to ongoing recombination between clade members. Reciprocally, interclade exchanges were rarely observed, suggesting mechanisms restricting gene flow between clades. Interrogation of accessory elements revealed that each clade harbored a distinct complement of restriction-modification (RM) systems, predicted to cause clade-specific patterns of DNA methylation. Using methylome sequencing, we confirmed that representative strains from separate clades indeed exhibit distinct methylation profiles. Finally, using an *E. coli* system, we demonstrate that Bp RM systems can inhibit uptake of non-self DNA. Our data suggest that RM systems borne on mobile elements, besides preventing foreign DNA invasion, may also contribute to limiting exchanges of genetic material between individuals of the same species. Genomic clades may thus represent functional units of genetic isolation in Bp, modulating intraspecies genetic diversity.

[Supplemental material is available for this article.]

Burkholderia pseudomallei (Bp) is the causative agent of melioidosis, a serious infectious disease of humans and animals and a leading cause of community-acquired sepsis and pneumonia in endemic regions (Currie et al. 2010). Initially thought to be confined to Southeast Asia and Northern Australia, the prevalence of Bp appears to be spreading (Wiersinga et al. 2012), and Bp has been designated a biothreat select agent in the United

States. Bp can persist in extreme environmental conditions and can infect several plant and animal hosts, including birds, dolphins, and humans (Wuthiekanun et al. 1995; Howard and Inglis 2003; Sprague and Neubauer 2004; Larsen et al. 2013). Treatment of clinical melioidosis is challenging because the bacterium is inherently resistant to many antibiotics, and Bp infections can persist in humans for more than a decade (Hayden et al. 2012; Wiersinga et al. 2012).

Present addresses: ¹³School of Medicine, University of St. Andrews, St. Andrews, KY16 9TF, UK; ¹⁴Division of Infection and Immunity, The Walter and Eliza Hall Institute of Medical Research, Parkville, 3052, Victoria, Australia.

Corresponding author: tanbop@gis.a-star.edu.sg

Article published online before print. Article, supplemental material, and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.177543.114>.

© 2015 Nandi et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <http://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

The Bp genome comprises one of the largest and most complex bacterial genomes sequenced to date. Consisting of two large circular replicons (chromosomes) with a combined 7.2-Mb genome size (Holden et al. 2004), it contains a rich arsenal of genes related to virulence (e.g., Type III and Type VI secretion systems, polysaccharide biosynthesis clusters), metabolic pathways, and environmental adaptation (Wiersinga et al. 2012). Besides conserved regions, accessory genes on mobile elements and genomic islands may also contribute to phenotypic and clinical differences in microbial behavior (Currie et al. 2000; Sim et al. 2008). Analysis of the Bp genome has revealed previously unknown toxins and mechanisms of antibiotic resistance (Chantratita et al. 2011; Cruz-Migoni et al. 2011).

Most large-scale studies of Bp genetic diversity to date have analyzed strains using multilocus sequence typing (MLST). These studies have suggested a high degree of genetic variability between Bp strains and related *Burkholderia* species (Cheng et al. 2008), and have shown that Bp strains belonging to different sequence types (STs) can often coexist in the same locale and sometimes even within the same sample (Pitt et al. 2007; Wuthiekanun et al. 2009). However, due to the limited number of genes analyzed by MLST, these studies cannot comment on the global proportion of genetic material shared between strains of different STs nor on the relative contribution of recombination, mutation, and horizontal gene transfer on intraspecies genetic diversity. Moreover, although previous studies have applied whole-genome sequencing (WGS) to study global patterns of Bp genetic heterogeneity and evolution, earlier Bp WGS reports have been confined to a limited number of isolates (10–12) derived from diverse geographical regions (Nandi et al. 2010), where geophysical barriers likely limit the propensity of the analyzed strains to exchange genetic material. To achieve a comprehensive understanding of genetic variation among closely related Bp strains, WGS analysis of much larger strain panels, ideally performed on strains isolated from a common region and belonging to the same (or closely related) ST groups, is required.

In this study, we attempted to fill this important knowledge gap by performing WGS on 106 Bp strains drawn from a restricted Asian locale (Singapore and Malaysia). The WGS data, exceeding previous Bp WGS studies by 10-fold, enabled us to identify specific genomic clades of Bp, molecular features of Bp recombination at the whole-genome level, and accessory genome features contributing to recombination and horizontal gene transfer. We found a consistent pattern of genetic separation correlating with MLST, recombination haplotypes, shared accessory genes, and restriction modification (RM) systems. We provide evidence that restriction modification, beyond its role as defense against foreign DNA invasion, may have also partitioned the Bp species by restricting gene flow, resulting in the other observed correlations. Because RM systems are widely dispersed through the bacterial kingdom, it is possible that similar principles may apply to other bacterial species, implicating a potential role for epigenetic barriers as a driver of early incipient speciation.

Results

Bp genome sequencing

We analyzed 106 Bp strains, including 97 strains from Singapore and Malaysia (87/10) and nine strains from Thailand (Supplemental Table S1). The Singaporean and Malaysian strains were isolated from various clinical, animal, and environmental sources

over a 10 yr period (1996–2005) (see Methods). MLST classified the strains into 22 sequence types (ST). Supporting their close phylogenetic relationships, 20 STs belonged to clonal complex CC48 (Supplemental Fig. S1A,B). The majority of strains were of ST51 (43 strains) and ST423 (16 strains).

Due to the high GC-bias of the Bp genome, we initially found that conventional Illumina sequencing protocols resulted in uneven genome coverage and suboptimal assemblies (median N_{50} : 2907 bp) (Supplemental Table S2). We overcame this problem by applying a PCR amplification-free strategy (Kozarewa et al. 2009), resulting in markedly improved genome coverage and assemblies (median depth 100 \times ; median N_{50} : 102,577 bp). In total, we predicted 84,846 high quality SNPs in the WGS panel compared to the K96243 reference (Chr I: 43,829 and Chr II: 41,017). We validated the technical accuracy of the WGS data by Sanger sequencing of 50 randomly selected SNPs. Of the predicted SNPs, all 50 were confirmed by Sanger sequencing.

Whole-genome phylogenetic analysis resolves genomic clades

We excluded SNPs associated with regions of recombination as previously described (Croucher et al. 2011), resulting in a set of 10,314 SNPs representing mutations inherited by vertical descent along different lineages (“lineage SNPs” [L-SNPs]). Maximum likelihood phylogenies using the L-SNPs identified three major clades (“genomic clades”) containing all the Singapore and Malaysian strains, clustering apart from Thailand strains (Fig. 1A). Strains of the same ST grouped together within the same genomic clade, indicating strong similarities between phylogenies based on WGS and MLST. However, compared to MLST, the WGS phylogenies provided increased resolving power. For example, although MLST indicated a high degree of relatedness between ST50 and ST414 strains, WGS revealed that ST50 is more related to ST46 (Genomic Clade C), with ST414 being a more distant group (Fig. 1A). WGS also subdivided the ST51 strains into two distinct subclades—ST51a (39 strains) and ST51b (four strains) (Fig. 1B)—distinguished by \sim 342 distinct L-SNPs (Supplemental Table S3). Notably, all three clades comprised a heterogeneous intermingling of strains from different isolation sources (e.g., clinical, animal, and environmental), arguing against the existence of a genetically distinct Bp subpopulation preferentially associated with human disease. The three clades also contained strains isolated during similar time periods (1996–2005), suggesting that genetically distinct Bp strains from different clades can coexist in the same region over many years.

L-SNPs occurred at a \sim 1.2-fold higher frequency on Chr II compared to Chr I (6.1×10^{-3} SNPs per site for Chr I versus 7.5×10^{-3} for Chr II, $P = 3.08 \times 10^{-14}$, $\chi^2 = 57.68$, χ^2 test) (Supplemental Table S4a), suggesting a preferential accumulation of genetic mutations on Chr II during evolution. The majority of L-SNPs corresponded to C/G \rightarrow T/A transitions (Fig. 1C; for Clade ST51a), in the context of CG dinucleotides ($P = 3 \times 10^{-10}$, binomial test, Fig. 1D), likely reflecting the tendency of methylated cytosines to form thymines (Kahramanoglou et al. 2012). For both chromosomes, L-SNPs preferentially localized to intergenic regions (Chr I: $P < 2.2 \times 10^{-16}$, $\chi^2 = 101.42$; Chr II: $P = 4.196 \times 10^{-14}$, $\chi^2 = 57.04$, χ^2 test), and one-third of L-SNPs occurring within genes were nonsynonymous (Supplemental Table S4b). The d_N/d_S ratio (proportion of rate of nonsynonymous substitutions per site to rate of synonymous substitutions per site) for the major STs (e.g., ST51a, ST84) ranged between 0.17 and 0.64 per genome. Similar results were obtained when we analyzed a more restricted subset of 8035 L-SNPs associated

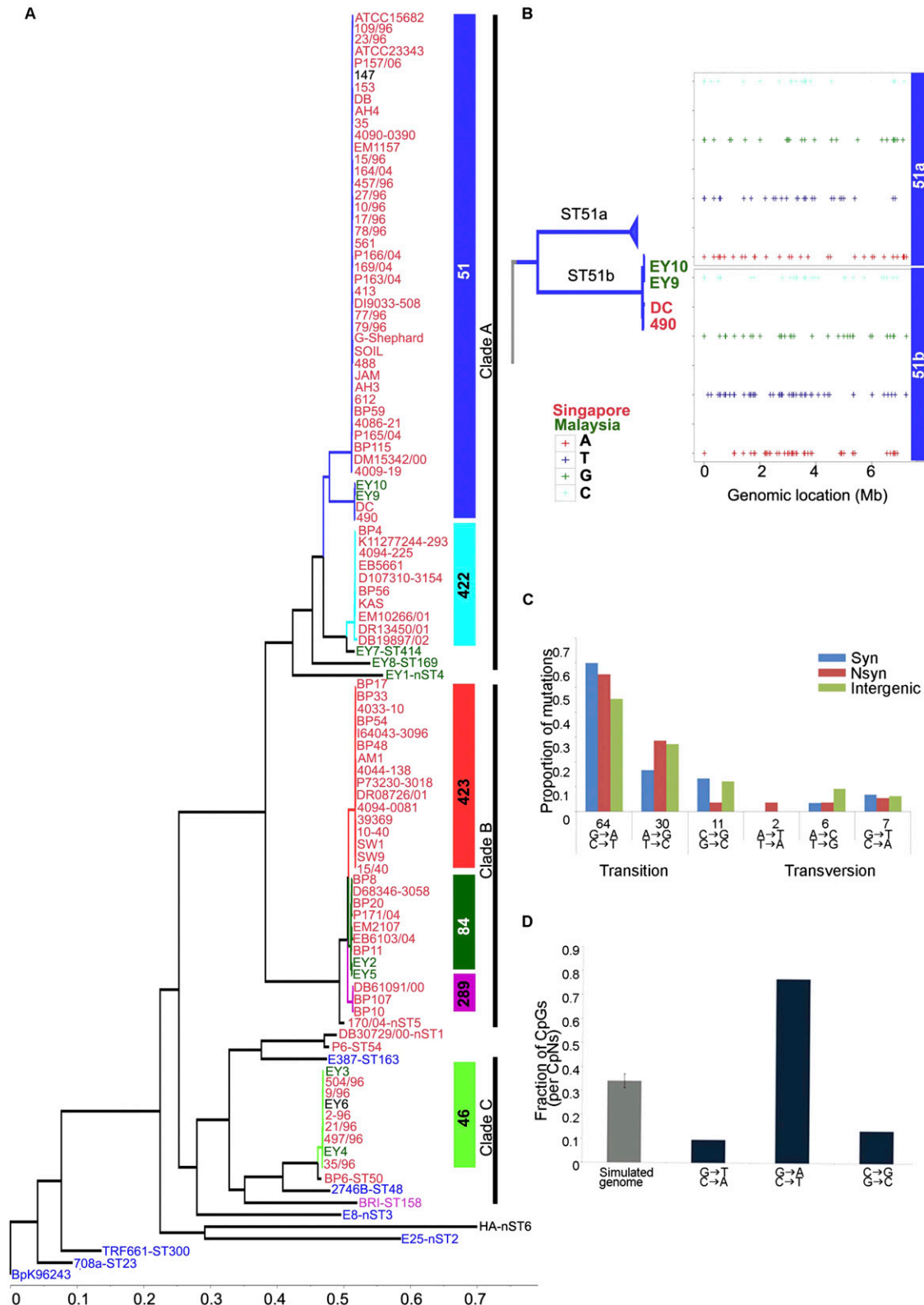


Figure 1. Whole-genome phylogeny and sequence variation of Bp strains. (A) Global phylogeny of Bp strains. The maximum likelihood tree was constructed using SNPs not associated with recombination events (see Results). Tip labels are colored according to the geographic locations of isolation ([red] Singapore; [green] Malaysia; [blue] Thailand; [black] unknown; [pink] imported to UK). *Inset bars at right* indicate the MLST scheme ([blue] ST51; [cyan] ST422; [red] ST423; [dark green] ST84; [pink] ST289; [light green] ST46). Three major genomic clades are identified: Clade A (ST51, ST422, ST414, ST169, nST4); Clade B (ST423, ST84, ST289, nST5); and Clade C (ST46, ST50). (B) Intra-ST subgroups resolved by WGS. ST51 strains cluster into two groups: ST51a and ST51b. Genomic locations of 342 L-SNPs (including both intra- and intergenic SNPs) distinguishing ST51a and ST51b are shown. The *top* and *bottom* panels with four rows show SNPs exclusively present in the two groups ST51a⁺ST51b⁻ and ST51a⁻ST51b⁺, respectively. (C) Mutation spectra of ST51a: relative rates of six possible mutation categories. The most common mutations are C/G → T/A transitions. (D) Fraction of the three classes of cytosine mutations occurring at CG dinucleotides in the Bp genome compared with the expected fraction based on the average of 100 simulated genomes of the same size and composition (gray).

with the Bp “core” genome (regions common to all Bp strains, estimated size 5.64 Mb) (Supplemental Table S5).

Widespread recombination among Bp isolates

Bp strains that are genetically distinct may interact within environmental reservoirs such as soil or water (Chantratita et al. 2008; Mayo et al. 2011) or in co-infected animals and human hosts (Pitt et al. 2007), thereby providing opportunities for recombination. Hence, it is vital to understand pathways and processes that facilitate or constrain gene flow between strains. To identify genomic characteristics of Bp recombination, we proceeded to analyze SNPs associated with recombination (R-SNPs). From 74,532 R-SNPs, we identified 2373 recombination events across the three genomic clades, with recombination tract lengths ranging from 3 bp to 71 kb (median ~5 kb). We computed recombination/mutation (r/m) values, corresponding to the ratio of rates at which substitutions are introduced by recombination and mutation, across the entire population. The overall per site r/m ratio was 7.2. Based upon these data, we estimate that at least 78% of the BpK96243 reference genome (~5.67 Mb) has undergone recombination, a level comparable to *S. pneumoniae*, a highly recombinogenic species (74K/85K R-SNPs for Bp; 50K/57K R-SNPs for *S. pneumoniae*) (Croucher et al. 2011). Similar to L-SNPs, higher recombination levels were observed for Chr II than Chr I ($P < 2.2 \times 10^{-16}$, Mann-Whitney U test),

Besides estimating whole-population metrics, we also computed clade-specific recombination and mutation rates for the three major Bp genomic clades, using a previously described Bayesian approach (ClonalFrame) (Didelot and Falush 2007). To minimize mapping artifacts, we excluded mobile genetic elements (e.g., phages, transposons, and genomic islands) and based our analysis on a reduced core genome of 5.6 Mb (see Methods). For all three clades, the ratio of mutation rate (theta) to recombination rate (rho) was close to one, suggesting that recombination and mutation both happen at approximately the same rates. Recombination was also found to introduce more substitutions than mutation (r/m = 4.5 in Clade A, r/m = 8.5 in Clade B, and r/m = 6 in Clade C) with the highest impact observed in Clade B (Supplemental Table S6). These values are in general agreement with the values obtained from the total population.

The high levels of recombination in the Bp clades motivated us to also analyze potential sources of recombination imports. We used previously established methods to assess intra- and interclade recombination flux (Didelot et al. 2009, 2011). Briefly, recombined fragments were compared with homologous sequences from other Bp genomes across the three clades, and a “match” was found if the sequence was identical or contained a single nucleotide difference. If a match was found to members of a single clade, the origin of the recombination event was attributed to this clade (matches to strains from multiple clades were categorized as ambiguous). If no matches were found, the origin was categorized as unknown. To estimate their relative impact on genomic diversification, the flux of genomic content between clades was summarized as the proportion of each genome originating from different origins. Of 2481, 821, and 334 recombination events detected within genomic Clades A, B, and C, respectively, we could assign sources (“matches”) for ~60% of recombination events (1112 matches to single clades and 1059 matches to multiple clades). On average, ~5% of each genome from a given clade was found to have originated from another clade and approximately another 7% from a source not present in our data set (Supplemental Table S7).

Several of the interclade recombination events were found on recent branches of the clonal genealogy, suggesting that the isolation is not complete between the clades.

Genome-wide median recombination frequencies (RFs) were computed to identify genomic regions exhibiting elevated recombination rates and multiple recombination events (Fig. 2A). We identified 1630 protein-coding genes (Chr I: 897 genes; Chr II: 733) associated with regions of high recombination (RF > RF_{median} + 3MAD, median absolute deviation). Genes experiencing high recombination frequencies were significantly enriched in intracellular trafficking and secretion pathways (corrected $P = 0.0006$, binomial test), whereas genes involved in protein translation were under-represented (corrected $P = 0.012$, binomial test) (Supplemental Table S8). Examples of genomic regions exhibiting elevated recombination included a Type III secretion cluster (*TTSS3*; *BPSS1520–BPSS1537*) previously linked to mammalian virulence (Stevens et al. 2002), and a Type IVB pilus cluster (*TFP8*, Chr II: *BPSS2185–BPSS2198*) (Fig. 2B). Type IV pili (*TFP*), including those of the sub-type IVB, encode surface-associated protein complexes involved in multiple cellular processes (Craig et al. 2004). To evaluate if *TFP8* might modulate Bp virulence, we generated an isogenic Bp deletion strain lacking ~12.9 kb of the *TFP8* locus and assessed the virulence of the *TFP8* mutant in a BALB/c mouse intranasal infection assay. The *TFP8* deletion mutant exhibited significantly reduced virulence compared to parental Bp K96243 wild-type controls ($P = 0.026$, Mantel-Haenszel log-rank test, Fig. 2C). These results support a role for Type IVB pili in Bp murine virulence, and more generally that a subset of recombination hotspots in Bp may influence mammalian virulence.

Genome-wide recombination haplotype map of Bp

Extended genomic stretches with high recombination rates often displayed specific combinations of local independent recombination events in individual strains, resulting in the creation of distinct haplotypes. Using the *TFP8* gene cluster as an example, some strains displayed recombination events R2, R7, and R8 (Haplotype 1 [H1]), whereas other strains displayed events R3, R7, and R8 (Haplotype 2 [H2]). In total, we identified five haplotypes (H1–H5) in the *TFP8* gene cluster (Supplemental Table S9). We found that these five *TFP8* haplotypes were tightly associated with specific clades; for example, haplotype H1 was associated with ST51 strains, whereas haplotype H4 (corresponding to recombination event R4) was associated with ST84 strains (Fig. 2D). To evaluate this association at the whole-genome level, we generated a whole-genome haplotype map of Bp, identifying 85 genomic regions exhibiting multiple (five or more) haplotypes (Supplemental Table S10). Similar to *TFP8*, the vast majority of haplotypes occurred in a genomic-clade specific pattern (Supplemental Fig. S2). Many of the multi-haplotype genomic regions were involved in specialized functions such as iron and cofactor metabolism, detoxification, and virulence (Supplemental Table S10). Almost half (48%) of the multiple-haplotype genomic regions exhibited at least one haplotype with an excess of nonsynonymous to synonymous SNPs, consistent with these regions having altered phenotypic properties. For instance, one haplotype over-represented in nonsynonymous SNPs occurred in the virulence-associated *TFP1* locus (*BPSL0782–BPSL0783*), within the *pilA* gene in ST51a strains. Notably, *pilA* plays a role in virulence yet its role in adherence and microcolony formation varies considerably in different Bp strains (Essex-Lopresti et al. 2005; Boddey et al. 2006). These findings suggest that haplotype variation may

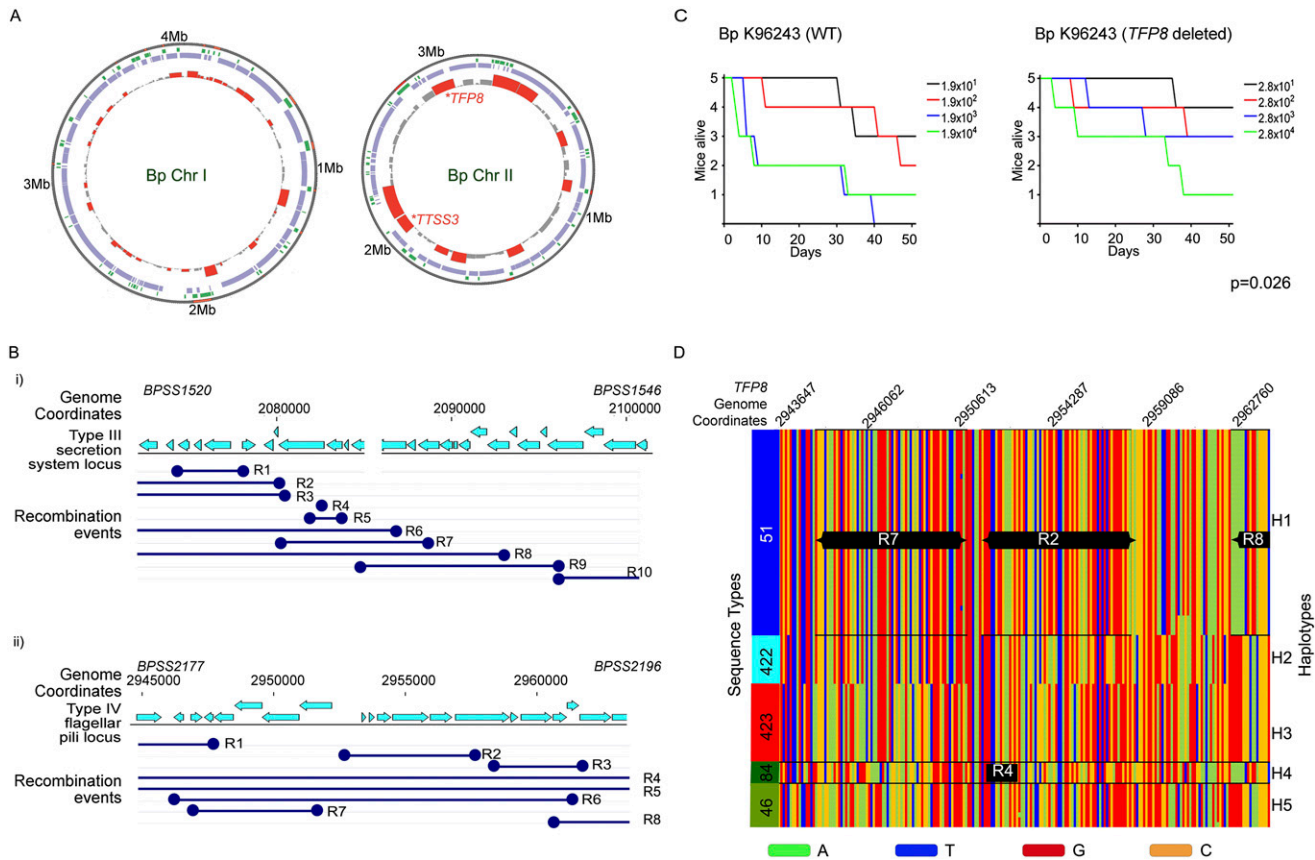


Figure 2. Recombination landscape of Bp. (A) Recombination hotspots in Bp. Circles: (outside) genome coordinates; (middle) compositionally biased regions identified by Alien hunter (Vernikos and Parkhill 2006) (green) and Bp core genome (violet); (innermost) regions of elevated recombination (height of red bars). Note that recombination levels are higher on Chr II than Chr I. Location of the *TFP8* and *TTSS3* clusters are indicated. (B) Local recombination events in the Type III secretion system and Type IV pilus cluster. (Top) Genomic coordinates and location of protein-coding genes; (dark blue) predicted recombination events (R1 to Rn, n = number of recombination events) observed in Bp strains belonging to genomic clades (ST group 46, 51, 84, 289, 422, and 423). The recombination boundaries are indicated by the dark blue circles and the boundaries that fall beyond the depicted locus are shown as open ended. (C) Relative virulence of *TFP8* deletion mutant. Graphs show survival curves of BALB/c mice following intranasal challenge with varying dosages of Bp (left: K96243 wild-type; right: *TFP8* deletion mutant, units are colony forming units, CFU). See Methods for infection assay details. The *TFP8* deletion mutant is significantly less virulent compared to Bp K96243 parental controls ($P = 0.026$, Mantel-Haenszel log rank test). (D) Distinct haplotypes at the *TFP8* genomic locus. Each row represents an individual Bp strain arranged according to genomic clade/ST (shown on left with color bars indicating ST51 [blue]; ST422 [cyan]; ST423 [red]; ST84 [dark green]; and ST46 [light green]). Across each row (strain), SNP positions are ordered by genomic coordinate (top numbers, Bp Chr II, genomic locus 2,935,860–2,976,718), and color-coded according to nucleotide identity (A → green; T → blue; C → orange; and G → red). The right y-axis "Haplotypes" refers to the specific linear combination of SNPs exhibited by individual strains. In some cases, haplotypes can be composed of a specific combination of smaller recombination regions (R). For example, Haplotype H1 is composed of recombination regions R2, R7, and R8. Haplotype alignments were generated using Clustal X (Larkin et al. 2007).

contribute to differences in Bp pathogenicity or survival in different strains.

Bp accessory genome elements exhibit clade restriction

The availability of WGS data for a large Bp panel also provided the opportunity to quantitatively assess the Bp accessory genome. Using the Velvet and NUCmer algorithms, we generated de novo sequence assemblies of genomic sequences not found in the K96243 reference genome. On average, ~183 kb of novel accessory regions (N_{AE}) were identified for each Bp strain (minimum region length 1 kb). We found that the Bp genome is "open," with at least 2897 new non-K96243 genes associated with the accessory regions (Fig. 3A). The Bp pan genome (Bp core + Bp accessory) is thus at least 8802 genes, which is 2× the size of the Bp core genome. Accessory genes were characterized by a lower percentage GC content (median value: $\sim 59\% \pm 5.6$) than the core genome

(~68%), consistent with their horizontally acquired nature. Accessory genes were also significantly enriched in pathways related to defense mechanisms (corrected $P < 0.0005$ relative to Bp core genes) (Fig. 3B).

Using pairwise similarity metrics, we evaluated the extent to which accessory elements found in one Bp strain might be shared with other strains. Similar to the recombination haplotypes, strains belonging to the same genomic clade had a tendency to share many of the same accessory elements (Fig. 3C). For example, strains in genomic Clade A shared a 15-kb gene cluster of metabolic genes, including biotin carboxylase, NAD-dependent malic enzymes, mandelate racemase, and 5-enolpyruvylshikimate-3-phosphate synthase (Priestman et al. 2005; Tang et al. 2005; Li et al. 2009). Similarly, strains from genomic Clade B (ST423/ST84/ST289) shared accessory genes such as filamentous hemagglutinin, *fhaC*, which plays a crucial role in mediating adherence to eukaryotic cells (Relman et al. 1989).

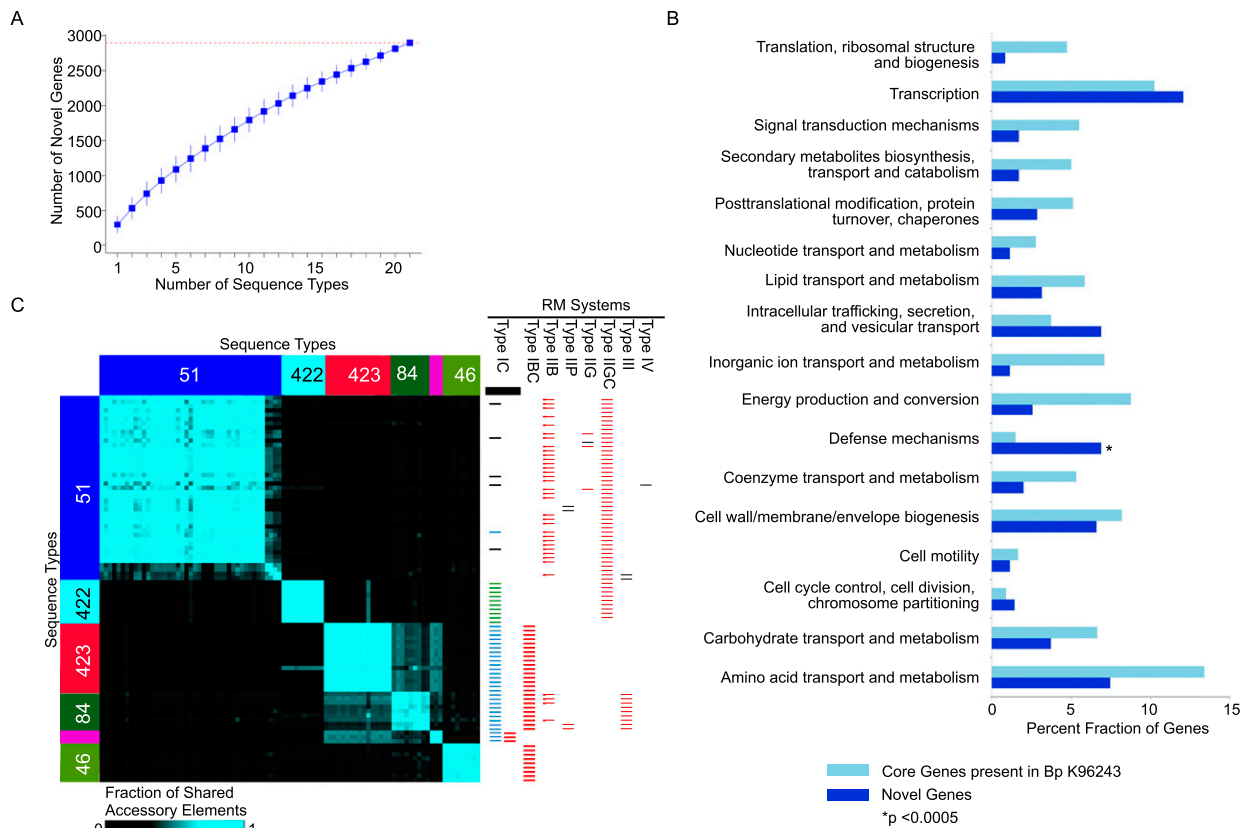


Figure 3. Accessory genome landscape of Bp. (A) Accumulation curves for Bp novel accessory genes (blue). Vertical bars represent standard deviation values based upon 100 randomized input orders of different Bp STs. The total number of accessory genes is indicated by the red dotted line. (B) Functional enrichment of Bp accessory genes. COG functional categories are indicated on the y-axis, and the percentage of genes in each COG category is shown on the x-axis. Dark blue columns represent novel accessory genes, and light blue columns indicate all Bp core genes with COG annotations. COG categories exhibiting a significant enrichment among the Bp accessory genes are highlighted by asterisks (* $P < 0.0005$, binomial test; after Bonferroni correction). The COG category “DNA replication, recombination and repair” was excluded as it was represented mainly by mobility genes, particularly transposases and integrases. (C) Distribution of accessory elements across Bp clades. The heatmap represents an all-pairwise strain comparison showing the degree of accessory element overlap between pairs of strains. Strains are arranged on the x- and y-axis according to their genomic clades and sequence types (ST51 [blue]; ST289 [pink]; ST422 [cyan]; ST423 [red]; ST84 [dark green]; and ST46 [light green]). The color scale bar at the bottom indicates the degree of accessory element sharing (more blue equates to increased sharing). The right-hand chart depicts the different types of restriction-modification (RM) systems associated with different clades. In each column, the RM systems are color-coded based on their encoded protein-coding sequences. In the first column, the bars in green and blue refer to two distinct sets of RM genes that belong to Type IC RM systems. Strain-specific RM systems are in black.

Evidence of ongoing recombination and gene exchange within clades

The analyses described above revealed a strong correlation between Bp clades, core genome recombination haplotypes, and complements of accessory elements. We hypothesized that these correlations could be explained by two alternative models—“ongoing recombination” or “vertical descent” (Fig. 4A). In the first model, active recombination is ongoing in Bp, but preferentially restricted to exchange of DNA within a clade. To test for ongoing recombination, we computed within-clade nucleotide divergence levels in DNA sequences predicted to have undergone recombination and compared these to divergence levels within regions of non-recombined DNA in strains exhibiting the recombination event. If recombination were ongoing among strains within a clade, then this would serve to homogenize the recombining sequences across strains, whereas nonrecombining regions would accumulate mutations independently in different strains (Fig. 4A, “Ongoing Recombination”). Thus, the within-clade nucleotide divergence of recombined regions would be predicted to be lower relative to

nonrecombined regions. In the alternative model, recombination is not commonly taking place. Here, recombined regions would have entered a clade in its founder, and then would be found throughout the clade due mostly to strict vertical descent (Fig. 4A, “Vertical Descent”). In this case, recombined regions would be predicted to accumulate mutations at the same rate as non-recombined regions.

For each recombined region in a clade, we calculated the average sequence divergence level in that region, using strains exhibiting the recombination event (see Supplemental Fig. S3 and Supplemental Text for a detailed description of this analysis). To obtain a conservative set of nonrecombined sequences for comparison, we then took only those sections of the Bp genome predicted not to have undergone any recombination in any of the Bp strains; and for the same strains we calculated the sequence divergence levels in these nonrecombined sequences. We found that recombined regions in the core genome had uniformly lower sequence divergence than nonrecombined regions (Fig. 4B), suggesting that recombination is active and ongoing within clades.

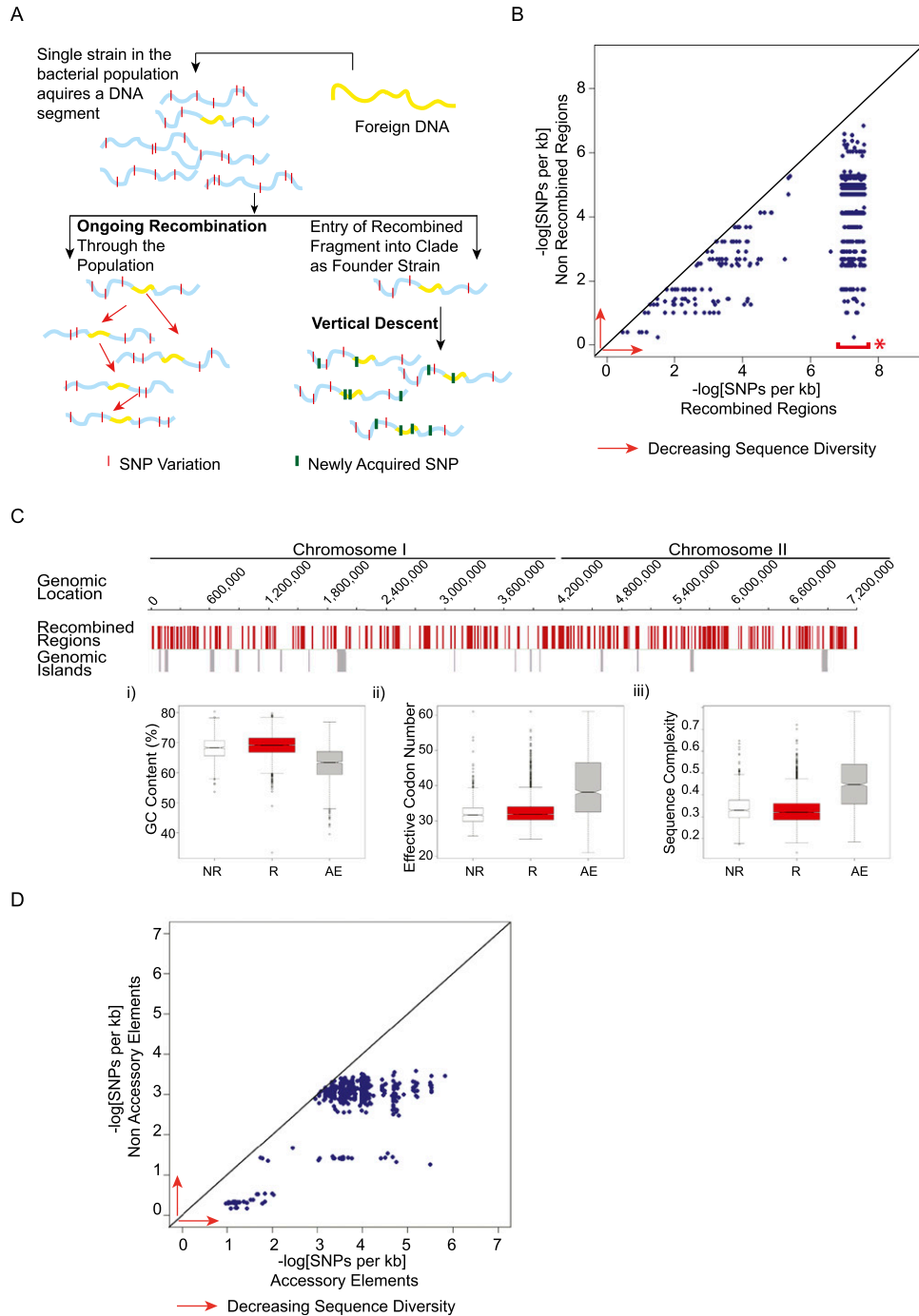


Figure 4. Distinguishing between ongoing recombination and vertical descent and in Bp. (A) Alternative models for clade-specific recombination haplotypes. (Left) In the “Ongoing Recombination” model, an imported fragment sweeps through the population via recombination, resulting in homogenization of the recombining fragment across strains. The recombining fragment should show lower levels of sequence diversity compared to nonimported regions. (Right) In the “Vertical Descent” model, an ancestral strain acquires a genomic fragment (yellow) from an external strain and subsequently transmits that fragment to all daughter strains in a clonal fashion. In this model, the imported fragment should accumulate new point mutations (green bars) at a similar rate to nonimported regions. (B) Within-clade sequence diversity of recombined regions compared to nonrecombined regions. Scatter plots comparing within-clade sequence diversity values of individual recombined regions (x-axis) to nonrecombined regions (y-axis) for the same strains in a given clade. Sequence diversity decreases in the direction of the red arrows (to right and upward). (*) Data points highlighted by the red bar correspond to recombined regions exhibiting 100% sequence identity. To visualize these points in a manner that captures both their density and extremely low sequence diversity, these were plotted within the x-axis range of 6.9–7.6 on a negative log scale. Sequence diversity is defined as the number of SNPs per kb. (C) Sequence features of nonrecombined regions (NR), recombined regions (R), and accessory elements (AE). (Top) Bp K96243 genomic tracks of Chr I and Chr II. Row 1: Genomic locations of recombined regions (red). Row 2: Genomic locations of 16 known Bp genomic islands (gray). (Bottom) Sequence feature comparison of genes in nonrecombined (white; NR), recombined (red; R), and accessory elements (gray; AE): (i) GC content (Puigbò et al. 2008); (ii) effective codon number (Puigbò et al. 2008); and (iii) sequence complexity (Petrokovski et al. 1990). Each hourglass plot spans the 25th to 75th percentile (interquartile range [IQR]) of all genes in that category, with the bottleneck at the median. Horizontal tick marks show data ranges within $1.5 \times \text{IQR}$ of the 25th and 75th percentiles. Open circles represent outliers outside this range. The width of the bottleneck (i.e., the length of the V-shaped notch) depicts the 95% confidence interval for the median. (D) Within-clade sequence diversity of accessory elements compared to nonaccessory elements. Accessory elements are defined as regions not present in the BpK96243 reference strain (see Methods). Scatter plots compare average sequence diversity values for individual accessory elements (x-axis) to corresponding nonaccessory elements (y-axis) for the same strain pairs in a given clade. Sequence diversity is defined as the number of SNPs per kb.

Extending this analysis to the gene level, we found that recombination has opposite effects on within- and between-clade divergence; genes in recombined regions had higher between-clade diversity compared with genes in nonrecombined regions, as expected, but much lower within-clade diversity (Supplemental Fig. S4). To rule out the possibility that the lower sequence divergence in recombined regions might be due to recombined regions possessing different sequence features or gene functions than nonrecombined regions (despite both regions being part of the same core genome), we also compared GC content, effective codon number, sequence complexity, and COG functions between the recombined regions, nonrecombined regions, and accessory elements. The latter was included because accessory elements are known to be distinct in gene function and sequence characteristics from the core genome (Kung et al. 2010). We found recombined and nonrecombined regions to be highly similar and distinct from accessory elements (Fig. 4C). For example, nucleotide frequencies between recombined and nonrecombined regions were similar (Chr I: $\chi^2 = 0.0012$, P -value = 1; Chr II: $\chi^2 = 2 \times 10^{-4}$, P -value = 1, χ^2 test) (Supplemental Table S11), and COG analysis of all genes associated with recombination regions failed to reveal any significantly enriched biological pathways compared to the whole genome (Supplemental Table S12), indicating that general baseline recombination in Bp is not functionally selected.

Besides recombination in the core genome, the correlation between Bp clades and complements of accessory elements suggested a further test for whether recombination is ongoing in Bp. Similar to the logic for the core recombined sequences above, accessory elements that are pervasive throughout a given clade could be undergoing active, ongoing exchange (which would homogenize their sequence within the clade) or be inherited through vertical descent (and thus accumulate mutations similar to adjacent nonrecombining regions) (Fig. 4A). Both these possibilities are consistent with clade-specific complements of accessory elements (Fig. 3C). We found that within each clade, accessory elements also showed lower sequence divergence levels compared to nonaccessory elements in the same strains (Fig. 4D). Thus, these results suggest that in both the core and accessory genome, there is a strong signal for ongoing, active recombination within Bp clades.

Identification of clade-specific RM systems

The clade-specific pattern of haplotypes and accessory elements in Bp, coupled with evidence of ongoing gene flow within strains of the same clade, suggests that reciprocal barriers to gene exchange may exist between strains belonging to different clades. We hypothesized that these barriers might be due, at least in part, to the use of distinct restriction-modification (RM) systems in each clade. RM systems comprise different combinations of endonuclease, methylase, and DNA specificity domains that use specific methylation patterns to label endogenous “self” genomic DNA, whereas unmodified exogenous DNA is recognized as “non-self” and subsequently cleaved and destroyed (Ershova et al. 2012; Makarova et al. 2013). Studies have proposed that RM systems can act as barriers to horizontal gene transfer (Waldron and Lindsay 2006; Hoskisson and Smith 2007; Dwivedi et al. 2013). However, a role for RM systems in restricting intraspecies recombination is less well described (Waldron and Lindsay 2006).

By interrogating genes in the Bp accessory genome and mobile genetic elements, we identified four different Bp RM systems (I, II, III, and IV) (Roberts et al. 2007). Notably, specific sets of RM systems were found in association with genomic clades bearing

distinct haplotypes and accessory genome features. For example, although clearly related by lineage (Fig. 1), clades ST51 and ST422 exhibit different recombination patterns and sets of accessory elements: We found that ST51 strains contained RM Type IIGC genes, whereas ST422 strains harbored both RM Type IIGC and RM Type IC systems. Similarly, strains from genomic Clade B (ST423/ST84/ST289) were largely dominated by RM Type IC and Type IBC systems, with type III RM genes additionally present in ST84 strains (Fig. 3C). The presence of these clade and ST-specific RM systems, which are predicted to result in clade-specific patterns of DNA methylation, may provide a molecular barrier to interclade gene sharing.

Methylome sequencing reveals clade-specific epigenetic profiles

To provide direct experimental data that Bp strains from different clades have distinct methylation profiles, we subjected one representative strain from Clade A (Bp35) and one strain from Clade B (Bp33) to whole-genome methylome analysis using SMRT sequencing technology (Murray et al. 2012). Because SMRT sequencing has the ability to measure DNA polymerase activity in real time, base modifications such as methylation can be detected as a change in the kinetics of base pair incorporation (Flusberg et al. 2010; Schadt et al. 2010).

SMRT sequencing followed by de novo assembly for Bp33 and Bp35 was performed to obtain two circular contigs of 4.0 and 3.1 Mb (average GC content is 68%) with 240 \times and 147 \times post-filter base coverage, 21 \times and 22 \times preassembled read coverage, respectively (Chin et al. 2013). We identified a 12.4-kb plasmid in Bp35 called pBp35 that to our knowledge represents the first plasmid described for Bp (Supplemental Fig. S5; Supplemental Data: plasmid sequence and annotation in GenBank format). We analyzed the local sequence contexts for all the methylated bases in both the strains and identified sequence motifs associated with these methylated bases. Both Bp strains showed methylation throughout their entire genomes. In total, six unique methylated motifs were identified in the two Bp strains (Table 1). Of these, one motif (5'-CACAG-3') was shared by the two strains, whereas the other five were strain- or clade-specific. For example, the type II motif (5'-GTAWAC-3') is unique to Bp35 (Clade A representative), whereas the type I motif (5'-GTCATN₅TGG-3') is present only in Bp33 (Clade B representative).

We proceeded to match the different motifs to specific RM systems found in the two genomes. Reassuringly, the shared CACAG motif was found to be associated with a conserved Type III RM system found in both strains. However, Bp 35 exhibited three strain-specific methylated motifs, and these could be associated with Type I and II systems found specifically in the Bp35 clade. Similarly, Bp 33 exhibited two strain-specific methylated motifs, and these could be associated with Type I RM systems found specifically in the Bp33 clade. For both strains, the fraction of strain-specific methylated motifs was close to 100%, consistent with their predicted methylation and restriction functions operating at high efficiency (Table 1). The demonstration that strains belonging to different Bp clades indeed have distinct methylation patterns is consistent with our hypothesis that clade-specific RM activity may represent a barrier to Bp interclade recombination and accessory element transfer.

Bp RM systems impede foreign DNA uptake in *E. coli*

To functionally test if clade-specific RM systems of Bp can impede the transfer of non-self DNA, we cloned and tested the Type I RM

Table 1. DNA methylation sequence motifs in Bp35 (Clade A) and Bp33 (Clade B)

Type of RM system	Methyltransferase activity ^a	Type of methylation	Total number of sites ^b	Number of methylated sites	Sites methylated (%)	Assignment	Locus	Reference
Type I	5'- GATC N ₅ GATG-3' 3'-CTAGN ₅ CT AC -5'	m ⁶ A	3086	Bp35 strain 3082	99.87	—	—	This study
Type II	5'-GTAW AC -3' 3'-CATWTG-5' 5'- CAGN ₆ CTG-3' 3'-GTCN ₆ G AC -5'	m ⁶ A	1152	1141	99.05	—	—	This study
Type III	5'-CAC AG -3'	m ⁶ A	5214	5197	99.67	—	—	This study
Type I	5'-CC ATN ₇ CTTC-3' 3'-GGTAN ₇ GA AG -5' 5'-GTC ATN ₅ TGG-3' 3'-CAGTAN ₅ ACC-5'	m ⁶ A	86	Bp33 strain 86	100	—	—	This study
Type III	5'-CAC AG -3'	m ⁶ A	211	210	99.53	—	—	This study
Type III	5'-CAC AG -3'	m ⁶ A	5214	5197	99.67	BceII	BURCENBC7_AP5195	REBASE

^aThe methylated position within the motif is highlighted in bold. Pairs of reverse-complementary motifs belonging to the same recognition sequence were grouped together.

^bThe total number includes motifs occurring on both the “+” and “-” strands.

system associated with Genomic Clade A (Bp33) in *E. coli* (Janscak et al. 1999; Kasarjian et al. 2003). We engineered one plasmid to carry the Type I “restriction” endonuclease (R⁺), and a second separate plasmid to carry the “specificity” and “methylase” proteins (M⁺) (Fig. 5A). M⁺R⁺ and M⁺R⁻ *E. coli* strains were then secondarily transformed with reporter plasmids carrying zero, one, and two copies of the RM recognition site predicted from SMRT sequencing (5'-GTCATN₅TGG-3'; see Table 1), and efficiencies of transformation (EOT) were calculated (Fig. 5B). We found that when transformed into M⁺R⁺ strains, unmethylated reporter plasmids carrying one or two recognition sites exhibited a > 100-fold decrease in EOT compared to reporter plasmids with no recognition sites ($P < 0.01$; Student's *t*-test) (Fig. 5C). Importantly, the Type I restriction endonuclease is required for this decrease, as no EOT differences were observed when the plasmids (zero, one, and two sites) were transformed into M⁺R⁻ strains which only express the methylase (Fig. 5C). This result indicates that the restriction endonuclease of the Clade A-specific Type I RM system is indeed active and capable of impeding the uptake of non-self DNA harboring an unmethylated Type I recognition site.

Next, we isolated plasmids with methylated recognition sites by passing them through M⁺R⁻ *E. coli* strains and transformed them into M⁺R⁺ strains. In contrast to the results using unmethylated plasmids, all three plasmids (zero, one, or two recognition sites) exhibited no significant EOT differences (Fig. 5D). This result indicates that, at least for one Bp clade-specific RM system, methylation of the recognition sites by RM methylases is sufficient to facilitate uptake of non-self DNA, even in the presence of its cognate restriction endonuclease.

Discussion

In this study, we performed WGS on a panel of Bp strains drawn from a restricted geographic locale to explore the contribution of gene mutation, recombination, and horizontal gene transfer to the molecular diversity of closely related Bp isolates. We found that Bp strains can be partitioned into distinct genomic clades and that a major proportion of the Bp core genome variation is strongly influenced by both mutation and recombination. Bp diversity is further enhanced by an accessory genome component that is at

least double the Bp core genome. Moreover, using diverse approaches, including (1) sequence diversity comparisons in both recombination and accessory regions supporting active gene flow within but not across clades; (2) genome-wide methylome sequencing demonstrating clade-specific epigenetic profiles associated with distinct RM-systems; and (3) experimental demonstration in an *E. coli* system that Bp RM systems are functionally active and sufficient to mediate the methylation and restriction of non-self DNA, our results point toward a model in which Bp RM systems may function as a barrier to gene exchange between different Bp clades.

Phylogenetic analysis of the Bp clades revealed that they comprised mixtures of Bp isolates from animal, clinical, and environmental sources, arguing against the existence of a genetically distinct population of Bp capable of infecting humans. Supporting this model, in a separate analysis, we were unable to confidently identify a consistent set of signature genetic changes in strains associated with human disease (T Nandi and P Tan, unpubl.). The genetic similarity between clinical, animal, and environmental Bp strains raises the possibility that additional genetic changes may not be required for an environmental Bp strain to successfully cause human disease. This model is consistent with previous proposals that Bp is an “accidental pathogen,” in which adaptations incurred by Bp to survive in its natural reservoir (soil and potentially single-celled organisms located therein, e.g., amoebae) must have indirectly contributed to its ability to colonize a mammalian host (Casadevall and Pirofski 2007; Nandi et al. 2010). This “accidental virulence” hypothesis is further supported by epidemiological data in which patients with clinical melioidosis often possess pre-infection morbidities such as diabetes, which may contribute to a weakened host immune response (Currie et al. 2010).

Our data revealed several genome-wide features of the Bp recombination landscape. We found that recombination in Bp is pervasive, approaching levels previously reported for *S. pneumoniae* (Croucher et al. 2011), and frequently involved defined sets of haplotypes. Importantly, analysis of genes in regions associated with high recombination suggests that haplotypes and recombination hotspots in Bp are not randomly distributed, but biased toward genomic regions associated with niche adaptation, survival, and virulence. This included a TTSS cluster involved in

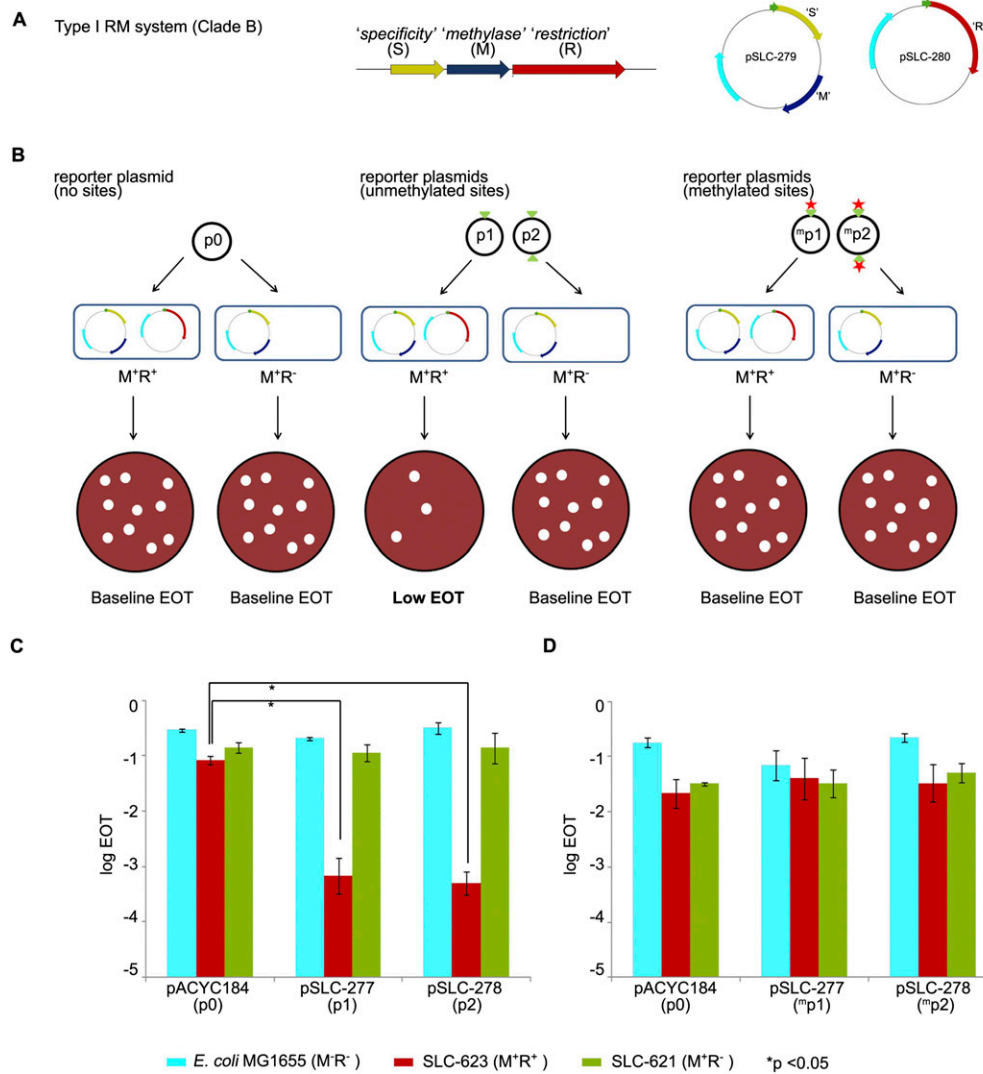


Figure 5. Restriction of non-self DNA by clade-specific Bp RM systems. (A) Molecular cloning of a Type I RM system specific to Bp genomic Clade A. The RM system comprises three genes: (R) “restriction,” (M) “methylase,” (S) “specificity.” Genes S (yellow) and M (blue) were cloned in plasmid pSLC-279 with kanamycin resistance (Km^R) to give the M⁺ plasmid. Gene R (red) was cloned in plasmid pSLC-280 with ampicillin resistance (Ap^R) to give the R⁺ plasmid. Resistance genes are depicted in cyan. Green arrows represent the T5 promoter used to induce expression of the cloned genes. Plasmids are not drawn to scale. (B) Efficiency of transformation (EOT) assay. Reporter plasmids p0, p1, and p2 harbor zero, one, and two copies of the predicted Type I recognition site (5'-GTCAAT₅TGG-3'; indicated by green triangles). Plasmid p0 should not show any EOT changes because it does not contain Type I recognition sequences. Unmethylated plasmids p1 and p2, when transformed into M⁺R⁺ strains, should be recognized via their Type I sites and cleaved by the Type I restriction enzyme (registered as a drop in number of transformants). However, when transformed into M⁺R⁻ strains that express the methyltransferase alone, no EOT differences should be observed. In contrast, methylated p1 and p2 plasmids (obtained by passage through M⁺R⁻ strains; methylated sites indicated by red stars and superscript ^m), when transformed into M⁺R⁺ strains, should be recognized as “self” DNA by the Type I system and resist cleavage, resulting in minimal EOT changes. (C) EOT assay results using unmethylated plasmids. Host strains are MG1655 (M⁻R⁻, no RM system, cyan); SLC-623 (M⁺R⁺, complete RM system, red); and SLC-621 (M⁺R⁻, methyltransferase only, green). Reporter plasmids are pACYC184 (p0, control plasmid); pSLC-277 (p1, 1 recognition site); and pSLC-278 (p2, 2 recognition sites). Significant differences in EOT are observed between control plasmid p0 and plasmids p1 and p2 when transformed into M⁺R⁺ strains ($P < 0.01$) but not in host *E. coli* or M⁺R⁻ strains. EOT in this study is the normalized number of Cm^R transformants obtained per unit amount of plasmid DNA. (D) EOT assay using methylated plasmids. Reporter plasmids were passed through M⁺R⁻ strains prior to transformation, which is predicted to cause recognition site methylation. No significant EOT differences are observed across the strains. All experiments were performed in triplicate, and data are presented as mean and standard deviations. Data are presented as log₁₀ values of EOT. Student's *t*-test was used to test for significant differences.

intracellular survival of Bp and virulence (Stevens et al. 2002) and a type IVB pilus cluster (*TFP8*) that we have validated in this study as required for maximal virulence in murine infection assays. It is thus possible that the other regions of high recombination identified in this study may contain additional genes involved in Bp adaptation, survival, and virulence.

Previous studies have shown that strains of different STs can be frequently co-isolated in the wild (Wuthiekanun et al. 2009). However, it is not known if such co-isolated strains are able to engage in an unrestricted exchange and transfer of genetic material. We found that strains associated with different Bp genomic clades tended to exhibit distinct sets of recombination haplotypes

and accessory elements. These findings suggest that Bp clade/ST subgroups may represent a functional and potentially limiting unit of within-species genetic diversity. However, it is important to note that although our findings suggest a general scenario of recombination events being largely clade-specific, exceptions do exist. One example of a possible recombination across separate clades involved a heme/porphyrin locus (*BPSS1245–BPSS1247*), where a haplotype present in strain P171/04 from ST84 (but not other ST84 strains) was highly similar to the haplotypes of ST51 strains (Supplemental Fig. S6). One explanation for these exceptions is that although barriers to interclade gene exchange do exist, they may be incomplete or perhaps only recently established.

Two broad models of Bp evolution could explain this clade-restriction pattern. First, clade restriction might result from effective physical or niche separation, in which strains from different clades are adapted or restricted to distinct niches in the environment, and therefore do not share DNA. However, as mentioned above, strains of different STs can be co-isolated; and while this does not rule out the presence of microscale niche differences between strains, such “micro-niches” remain to be experimentally proven. Alternatively, the discovery of clade-specific RM systems provides an epigenetic explanation for Bp clade restriction. In this model, acquisition of a diversity of RM systems to combat invading DNA may thus have occurred as a primary event, and the resulting epigenetic differences may in turn have established barriers to intraclade DNA exchange. A synthesis of these two models is also possible, where early subspeciation is initially driven by epigenetic barriers and followed subsequently by traditional niche selection. Consistent with this model, we observed that among ST84 strains, the two Malaysian Bp strains, EY2 and EY5, clustered separately from the remaining seven Singapore strains, being separated by 13 L-SNPs mapping to 12 protein-coding genes. This scenario could be explained by initial RM-driven epigenetic isolation of the ST84 clade, followed by geographical separation between the Singapore and Malaysian ST84 strains. The presence of localized concentrations of nonsynonymous changes suggests the possible existence of specific selective pressures driving further divergence, and it is also possible that the driver for speciation in Bp is currently shifting from divergence due to genetic/epigenetic separation to divergence due to selection in different niches. As such, our results provide a potential snapshot of early incipient speciation in a microorganism associated with diverse genomes and habitats.

Methods

Ethics statement

This research was approved by the GIS Institutional Review Board. Animal studies were performed in accordance with the UK Scientific Procedures Act (Animals) (1986) and UK Codes of Practice for the Housing and Care of Animals Used in Scientific Procedures (1989).

Bacterial strains, plasmids, and primers

Strains used were obtained from DMERI, DSO National Laboratories. These include (1) 56 clinical isolates from melioidosis patients between 1996 and 2004; (2) 34 animal isolates from various species (e.g., monkeys, pigs, birds, and dogs) diagnosed with melioidosis between 1996 and 2005; and (3) 16 soil isolates from 1996 to 2000 (Supplemental Table S1). The isolates were sampled from a diversity of locations and not a single site (AL Tin, pers. comm.). For *E. coli* experiments, plasmids bearing predicted

recognition sequences for the Clade A-specific Type I RM system (5'-GTCATN₃TGG-3') were generated by PCR-mediated insertion (see Supplemental Methods). Gene sequences encoding the Bp33 Type I restriction modification system proteins (“specificity,” “methylase,” and “restriction endonuclease”) were cloned into expression vectors driven by a *T5* inducible promoter (DNA2.0, Singapore). All plasmids were propagated in *E. coli* K12 strain MG1655. Bacterial strains, plasmids, and primers are listed in Supplemental Table S13.

Genomic DNA extraction and multiplex sequencing

Live bacteria were grown in a BioSafety Level 3 facility in DSO National Laboratories. Genomic DNA was extracted using the Qiagen Genomic Tip 500/G kit (Qiagen). Unique index-tagged libraries for each sample were created, and up to 33 separate libraries were sequenced per lane on an Illumina HiSeq instrument with 100 base paired-end reads. Libraries were constructed using an amplification-free method (Kozarewa et al. 2009). Raw Illumina data were split to generate paired-end reads, and assembled using a de novo genome-assembly program, Velvet v0.7.03 (Zerbino and Birney 2008), to generate a multicontig draft genome for each Bp isolate.

Gene annotation, SNP, and phylogenetic analysis

Paired-end reads were mapped against the chromosomes of *B. pseudomallei* K96243 (accession numbers BX571965 and BX571965) (Holden et al. 2004). Bp genes were predicted using FGENESB (<http://www.softberry.com>). Gene orthologs were determined using OrthoMCL (Chen et al. 2006). RM systems were inferred based on the specificity sequences of homologs in REBASE (Roberts et al. 2007) and categorized into subtypes—IC, IBC, IIG, IIGC, IIP, and IIB—on the basis of their genetic organization, mode of action, recognition sites, and cleavage loci (Roberts et al. 2003). SNPs predicted to have arisen by homologous recombination were identified using Gubbins and excluded from phylogenetic reconstruction (Croucher et al. 2011). Indels were identified using Dindel (Albers et al. 2011). Maximum likelihood phylogenies were constructed using RAxML v0.7.4 (Stamatakis et al. 2005). SNPs ancestral and derived alleles (polarization) were determined according to the outgroup reference strain sequence.

Recombination analysis

The general time-reversible model with gamma correction was used for among-site rate variation for 10 initial random trees. To measure clade-specific recombination rates, ClonalFrame (Didelot and Falush 2007) was applied separately to each Bp clade. To reduce mapping artifacts, we focused on the 5.6-Mb portion of the core genome that excludes mobile genetic elements and other potentially biased regions such as surface polysaccharides, secretion systems, and tandem repeats. Recombination events were extracted from ClonalFrame as genomic fragments, where the probability of recombination for a given branch of the tree was consistently > 50% and reached 95% in at least one location. Potential origins of recombination imports were investigated as previously described (Didelot et al. 2009, 2011; Sheppard et al. 2013). To determine haplotypes, SNP alleles at the recombination loci were concatenated to give a single haplotype string for each strain. The aligned strings were then subjected to hierarchical clustering as implemented in R package “hclust.” The resulting dendrogram was used to assign strains to distinct haplotype groups using the “cutree” function in R. Within-clade sequence diversity comparisons between recombined and nonrecombined regions

were performed as described in Supplemental Figure S3 and the Supplemental Text. Potential differences in sequence composition between recombined and nonrecombined regions were assessed using Artemis release 13.0 (Carver et al. 2012) or the K2 algorithm in CLC Main workbench 6.5 (<http://www.clcbio.com>) (Wootton and Federhen 1993).

Bp mutagenesis and mouse virulence studies

Isogenic Bp mutants carrying a 12.9-kb deletion *TFP8* were generated in a two-step process as previously described (see Supplemental Methods; Essex-Lopresti et al. 2005; Boddey et al. 2006). Virulence of wild-type and mutant Bp strains was assessed using an intranasal BALB/c mouse model (Essex-Lopresti et al. 2005). Briefly, groups of six age-matched BALB/c female mice were anesthetized and infected intranasally with 10-fold dilutions (10^1 – 10^6) of either wild-type Bp K96243 or *TFP8* deletion strains grown overnight at 37°C with shaking. Mice were recovered, and survival was recorded for up to 51 d. Analysis was performed using the Mantel-Haenszel log rank test in GraphPad Prism 4 or by Regression with Life Data in MiniTAB v13.0, using a significance threshold of $P = 0.05$.

Accessory genome analysis

Nucmer (Kurtz et al. 2004) was used to generate alignments of Velvet contigs against the reference strain Bp K96243 to identify novel accessory regions (N_{AE}). N_{AE} values for individual Bp strains were defined as blocks with a minimal 1000-bp length that was absent in Bp K96243 (median N_{AE} per strain = 183,482 bp). Sequence diversity comparisons between accessory and nonaccessory regions utilized accessory regions > 1000 bp and performed using MUMmer 3.20 under DNAdiff default settings (Kurtz et al. 2004).

SMRT sequencing and data analysis

Twenty micrograms of gDNA was processed to create SMRTbell sequencing templates > 10 kb (average insert size 17 kb) and sequenced using a PacBio RS II System in which polymerase-MagBead-bound templates were loaded at an on-plate concentration of 150 pM. Templates were subsequently sequenced using DNA Sequencing Kit 2.0, with data collection of 180 mins (Pacific Biosciences). Genomes were assembled using HGAP (Chin et al. 2013) with default parameters in SMRT Analysis Suite version 2.1 (Pacific Biosciences). Additional manual assembly of contigs was carried out in cases of unique overlapping sequence. Consensus sequence polishing was done using the Quiver algorithm in Genomic Consensus version 0.7.0. Base modification analysis was performed by mapping SMRT sequencing reads to the respective assemblies using the BLASR mapper (Chaisson and Tesler 2012) and SMRT Analysis Suite version 2.1 using standard mapping protocols. Clustering of sequence motifs was performed using Motif Finder (<https://github.com/PacificBiosciences/DevNet/wiki/Motiffinder>). See Supplemental Methods for further details.

Restriction-modification assay

Plasmids containing methylated and unmethylated Type I restriction sites were transformed into *E. coli* strains engineered to express all three proteins of the Type I RM system or only the specificity and methylase units. Efficiency of transformation (EOT) values were computed by comparing bacterial titers (colony forming units per mL, cfu/mL) on antibiotic selection plates divided by the corresponding titers from LB plates. EOT values were log transformed and plotted for analysis of RM system restriction activity. EOT values from triplicate experiments were compared

using a two-tailed Student's *t*-test. See Supplemental Methods for further details.

Statistical analysis

All statistical analyses were performed using R-2.15.1 (Ihaka and Gentleman 1996).

Data access

The data from this study have been submitted to the European Nucleotide Archive (ENA; <http://www.ebi.ac.uk/ena>) under accession number ERP000251. Methylation data have been submitted to the NCBI Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE55168.

Acknowledgments

This study was supported by a core grant to P.T. from the GIS, an A-STAR research institute. The sequencing of the *Burkholderia pseudomallei* strains was supported by Wellcome Trust grant 098051 to J.P.

References

- Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. 2011. Dindel: accurate indel calls from short-read data. *Genome Res* **21**: 961–973.
- Boddey JA, Flegg CP, Day CJ, Beacham IR, Peak IR. 2006. Temperature-regulated microcolony formation by *Burkholderia pseudomallei* requires *pilA* and enhances association with cultured human cells. *Infect Immun* **74**: 5374–5381.
- Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28**: 464–469.
- Casadevall A, Pirofski LA. 2007. Accidental virulence, cryptic pathogenesis, martians, lost hosts, and the pathogenicity of environmental microbes. *Eukaryot Cell* **6**: 2169–2174.
- Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**: 238.
- Chantratita N, Wuthiekanun V, Limmathurotsakul D, Vesaratchavest M, Thanwisai A, Amornchai P, Tumapa S, Feil EJ, Day NP, Peacock SJ. 2008. Genetic diversity and microevolution of *Burkholderia pseudomallei* in the environment. *PLoS Negl Trop Dis* **2**: e182.
- Chantratita N, Rhol DA, Sim B, Wuthiekanun V, Limmathurotsakul D, Amornchai P, Thanwisai A, Chua HH, Ooi WF, Holden MT, et al. 2011. Antimicrobial resistance to ceftazidime involving loss of penicillin-binding protein 3 in *Burkholderia pseudomallei*. *Proc Natl Acad Sci* **108**: 17165–17170.
- Chen F, Mackey AJ, Stoekert CJ Jr, Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **34**: D363–D368.
- Cheng AC, Ward L, Godoy D, Norton R, Mayo M, Gal D, Spratt BG, Currie BJ. 2008. Genetic diversity of *Burkholderia pseudomallei* isolates in Australia. *J Clin Microbiol* **46**: 249–254.
- Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**: 563–569.
- Craig L, Pique ME, Tainer JA. 2004. Type IV pilus structure and bacterial pathogenicity. *Nat Rev Microbiol* **2**: 363–378.
- Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, et al. 2011. Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**: 430–434.
- Cruz-Migoni A, Hautbergue GM, Artymiuk PJ, Baker PJ, Bokori-Brown M, Chang CT, Dickman MJ, Essex-Lopresti A, Harding SV, Mahadi NM, et al. 2011. A *Burkholderia pseudomallei* toxin inhibits helicase activity of translation factor eIF4A. *Science* **334**: 821–824.
- Currie BJ, Fisher DA, Howard DM, Burrow JN. 2000. Neurological melioidosis. *Acta Trop* **74**: 145–151.
- Currie BJ, Ward L, Cheng AC. 2010. The epidemiology and clinical spectrum of melioidosis: 540 cases from the 20 year Darwin prospective study. *PLoS Negl Trop Dis* **4**: e900.

- Didelot X, Falush D. 2007. Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**: 1251–1266.
- Didelot X, Barker M, Falush D, Priest FG. 2009. Evolution of pathogenicity in the *Bacillus cereus* group. *Syst Appl Microbiol* **32**: 81–90.
- Didelot X, Bowden R, Street T, Golubchik T, Spencer C, McVean G, Sangal V, Anjum MF, Achtman M, Falush D, et al. 2011. Recombination and population structure in *Salmonella enterica*. *PLoS Genet* **7**: e1002191.
- Dwivedi GR, Sharma E, Rao DN. 2013. *Helicobacter pylori* DprA alleviates restriction barrier for incoming DNA. *Nucleic Acids Res* **41**: 3274–3288.
- Ershova AS, Karyagina AS, Vasiliev MO, Lyashchuk AM, Lunin VG, Spirin SA, Alexeevski AV. 2012. Solitary restriction endonucleases in prokaryotic genomes. *Nucleic Acids Res* **40**: 10107–10115.
- Essex-Lopresti AE, Boddey JA, Thomas R, Smith MP, Hartley MG, Atkins T, Brown NE, Tsang CH, Peak IR, Hill J, et al. 2005. A type IV pilin, PilA, contributes to adherence of *Burkholderia pseudomallei* and virulence in vivo. *Infect Immun* **73**: 1260–1264.
- Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**: 461–465.
- Hayden HS, Lim R, Brittnacher MJ, Sims EH, Ramage ER, Fong C, Wu Z, Crist E, Chang J, Zhou Y, et al. 2012. Evolution of *Burkholderia pseudomallei* in recurrent melioidosis. *PLoS ONE* **7**: e36507.
- Holden MT, Titball RW, Peacock SJ, Cerdeño-Tarraga AM, Atkins T, Crossman LC, Pitt T, Churcher C, Mungall K, Bentley SD, et al. 2004. Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. *Proc Natl Acad Sci* **101**: 14240–14245.
- Hoskisson PA, Smith MC. 2007. Hypervariation and phase variation in the bacteriophage 'resistome'. *Curr Opin Microbiol* **10**: 396–400.
- Howard K, Inglis TJ. 2003. Novel selective medium for isolation of *Burkholderia pseudomallei*. *J Clin Microbiol* **41**: 3312–3316.
- Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. *J Comput Graph Stat* **5**: 299–314.
- Janscak P, MacWilliams MP, Sandmeier U, Nagaraja V, Bickle TA. 1999. DNA translocation blockage, a general mechanism of cleavage site selection by type I restriction enzymes. *EMBO J* **18**: 2638–2647.
- Kahramanoglou C, Prieto AI, Khedkar S, Haase B, Gupta A, Benes V, Fraser GM, Luscombe NM, Seshasayee AS. 2012. Genomics of DNA cytosine methylation in *Escherichia coli* reveals its role in stationary phase transcription. *Nat Commun* **3**: 886.
- Kasarjian JK, Iida M, Ryu J. 2003. New restriction enzymes discovered from *Escherichia coli* clinical strains using a plasmid transformation method. *Nucleic Acids Res* **31**: e22.
- Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* **6**: 291–295.
- Kung VL, Ozer EA, Hauser AR. 2010. The accessory genome of *Pseudomonas aeruginosa*. *Microbiol Mol Biol Rev* **74**: 621–641.
- Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12.
- Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.
- Larsen E, Smith JJ, Norton R, Corkeron M. 2013. Survival, sublethal injury, and recovery of environmental *Burkholderia pseudomallei* in soil subjected to desiccation. *Appl Environ Microbiol* **79**: 2424–2427.
- Li L, Lu W, Han Y, Ping S, Zhang W, Chen M, Zhao Z, Yan Y, Jiang Y, Lin M. 2009. A novel RPMXR motif among class II 5-enolpyruvylshikimate-3-phosphate synthases is required for enzymatic activity and glyphosate resistance. *J Biotechnol* **144**: 330–336.
- Makarova KS, Wolf YI, Koonin EV. 2013. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res* **41**: 4360–4377.
- Mayo M, Kaesti M, Harrington G, Cheng AC, Ward L, Karp D, Jolly P, Godoy D, Spratt BG, Currie BJ. 2011. *Burkholderia pseudomallei* in unchlorinated domestic bore water, Tropical Northern Australia. *Emerg Infect Dis* **17**: 1283–1285.
- Murray IA, Clark TA, Morgan RD, Boitano M, Anton BP, Luong K, Fomenkov A, Turner SW, Korlach J, Roberts RJ. 2012. The methylomes of six bacteria. *Nucleic Acids Res* **40**: 11450–11462.
- Nandi T, Ong C, Singh AP, Boddey J, Atkins T, Sarkar-Tyson M, Essex-Lopresti AE, Chua HH, Pearson T, Kreisberg JF, et al. 2010. A genomic survey of positive selection in *Burkholderia pseudomallei* provides insights into the evolution of accidental virulence. *PLoS Pathog* **6**: e1000845.
- Petrokovski S, Hirshon J, Trifonov EN. 1990. Linguistic measure of taxonomic and functional relatedness of nucleotide sequences. *J Biomol Struct Dyn* **7**: 1251–1268.
- Pitt TL, Trakulsomboon S, Dance DA. 2007. Recurrent melioidosis: possible role of infection with multiple strains of *Burkholderia pseudomallei*. *J Clin Microbiol* **45**: 680–681.
- Priestman MA, Funke T, Singh IM, Crupper SS, Schönbrunn E. 2005. 5-Enolpyruvylshikimate-3-phosphate synthase from *Staphylococcus aureus* is insensitive to glyphosate. *FEBS Lett* **579**: 728–732.
- Puigbò P, Bravo IG, Garcia-Vallve S. 2008. CAIcal: a combined set of tools to assess codon usage adaptation. *Biol Direct* **3**: 38.
- Relman DA, Domenighini M, Tuomanen E, Rappuoli R, Falkow S. 1989. Filamentous hemagglutinin of *Bordetella pertussis*: nucleotide sequence and crucial role in adherence. *Proc Natl Acad Sci* **86**: 2637–2641.
- Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, Blumenthal RM, Degtyarev S, Dryden DT, Dybvig K, et al. 2003. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res* **31**: 1805–1812.
- Roberts RJ, Vincze T, Posfai J, Macelis D. 2007. REBASE—enzymes and genes for DNA restriction and modification. *Nucleic Acids Res* **35**: D269–D270.
- Schadt EE, Turner S, Kasarskis A. 2010. A window into third-generation sequencing. *Hum Mol Genet* **19**: R227–R240.
- Sheppard SK, Didelot X, Jolley KA, Darling AE, Pascoe B, Meric G, Kelly DJ, Cody A, Colles FM, Strachan NJ, et al. 2013. Progressive genome-wide introgression in agricultural *Campylobacter coli*. *Mol Ecol* **22**: 1051–1064.
- Sim SH, Yu Y, Lin CH, Karuturi RK, Wuthiekanun V, Tuanyok A, Chua HH, Ong C, Paramalingam SS, Tan G, et al. 2008. The core and accessory genomes of *Burkholderia pseudomallei*: implications for human melioidosis. *PLoS Pathog* **4**: e1000178.
- Sprague LD, Neubauer H. 2004. Melioidosis in animals: a review on epizootiology, diagnosis and clinical presentation. *J Vet Med B Infect Dis Vet Public Health* **51**: 305–320.
- Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**: 456–463.
- Stevens MP, Wood MW, Taylor LA, Monaghan P, Hawes P, Jones PW, Wallis TS, Galyov EE. 2002. An Inv/Mxi-Spa-like type III protein secretion system in *Burkholderia pseudomallei* modulates intracellular behaviour of the pathogen. *Mol Microbiol* **46**: 649–659.
- Tang DJ, He YQ, Feng JX, He BR, Jiang BL, Lu GT, Chen B, Tang JL. 2005. *Xanthomonas campestris* pv. *campestris* possesses a single gluconeogenic pathway that is required for virulence. *J Bacteriol* **187**: 6231–6237.
- Vernikos GS, Parkhill J. 2006. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics* **22**: 2196–2203.
- Waldron DE, Lindsay JA. 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**: 5578–5585.
- Wiersinga WJ, Currie BJ, Peacock SJ. 2012. Melioidosis. *N Engl J Med* **367**: 1035–1044.
- Wootton JC, Federhen S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* **17**: 149–163.
- Wuthiekanun V, Smith MD, Dance DA, White NJ. 1995. Isolation of *Pseudomonas pseudomallei* from soil in north-eastern Thailand. *Trans R Soc Trop Med Hyg* **89**: 41–43.
- Wuthiekanun V, Limmathurotsakul D, Chantratita N, Feil EJ, Day NP, Peacock SJ. 2009. *Burkholderia pseudomallei* is genetically diverse in agricultural land in Northeast Thailand. *PLoS Negl Trop Dis* **3**: e496.
- Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**: 821–829.

Received April 22, 2014; accepted in revised form September 15, 2014.

A set of powerful negative selection systems for unmodified Enterobacteriaceae

Varnica Khetrpal¹, Kurosh Mehershahi¹, Shazmina Rafee¹, Siyi Chen¹, Chiew Ling Lim¹ and Swaine L. Chen^{1,2,*}

¹National University of Singapore, Department of Medicine, Yong Loo Lin School of Medicine, 1E Kent Ridge Road, NUHS Tower Block, Level 10, Singapore 119074 and ²Genome Institute of Singapore, Infectious Diseases Group, 60 Biopolis Street, Genome, #02-01, Singapore 138672

Received September 23, 2014; Revised March 06, 2015; Accepted March 10, 2015

ABSTRACT

Creation of defined genetic mutations is a powerful method for dissecting mechanisms of bacterial disease; however, many genetic tools are only developed for laboratory strains. We have designed a modular and general negative selection strategy based on inducible toxins that provides high selection stringency in clinical *Escherichia coli* and *Salmonella* isolates. No strain- or species-specific optimization is needed, yet this system achieves better selection stringency than all previously reported negative selection systems usable in unmodified *E. coli* strains. The high stringency enables use of negative instead of positive selection in phage-mediated generalized transduction and also allows transfer of alleles between arbitrary strains of *E. coli* without requiring phage. The modular design should also allow further extension to other bacteria. This negative selection system thus overcomes disadvantages of existing systems, enabling definitive genetic experiments in both lab and clinical isolates of *E. coli* and other Enterobacteriaceae.

INTRODUCTION

Elucidation of the molecular basis for a given phenotype in bacteria (such as the ability to cause an infection) is heavily reliant on the ability to create defined genetic mutations. Numerous systems exist for manipulation of bacterial chromosomes; in general, the most powerful of these use selection to isolate the desired mutant. Positive selection markers based on antibiotic resistance genes are thus a mainstay of the genetic toolbox in nearly every genetically tractable bacterium.

However, cloning strategies relying solely on positive selection markers (which enable isolation of bacteria carrying the marker) result in 'marked' strains, where the marker it-

self or a residual scar remains in the genome, potentially causing unanticipated effects. Creation of definitive genetic constructs (e.g. a single point mutation in a gene of interest) is therefore greatly facilitated by negative selection markers (which allow selection of bacteria without the marker). Having negative selection enables a two-step strategy of (i) positive selection-mediated deletion/replacement of a gene of interest by a selection cassette followed by (ii) negative selection-mediated, seamless replacement of the same cassette by a mutated allele (Figure 1A). Comparison with another strain in which the selection cassette is similarly replaced with a wild-type allele thus allows rigorous assignment of phenotypic differences to the engineered point mutation.

Negative selection markers are in general less efficient than positive selection markers, even for well-studied bacteria such as *Escherichia coli* (1–8). Particularly for disease-causing clinical isolates, negative selection systems that require host genotype modification (e.g. *ccdB*, *tolC*, *thyA*, *rpsL* and thymidine kinases (4–8)) are impractical, as the required host modifications often involve conserved metabolic functions that may impact virulence (9), thereby confounding the analysis or requiring an additional step to restore the original genotype. Among negative selection systems that are usable in unmodified host strains (and are therefore candidates for direct use in clinical isolates), additional drawbacks include low selection stringency (as for *sacB* (2)) and the need for strain-specific optimization of selection conditions (as for the *tetA* and *tetA-sacB* systems (1,3)).

To overcome these disadvantages in unmodified hosts, we designed a general and modular negative selection system based on inducible toxins. We optimized selection conditions in one clinical strain of *E. coli*, UT189, then verified that this design was usable without further modification or optimization in lab and other clinical isolates of *E. coli* as well as in *Salmonella enterica*, enabling convenient two-step creation of markerless and scarless chromosomal point mutations. The negative selection module achieved high selec-

*To whom correspondence should be addressed. Tel: +6568088074; Fax: +6568088036; Email: slchen@gis.a-star.edu.sg

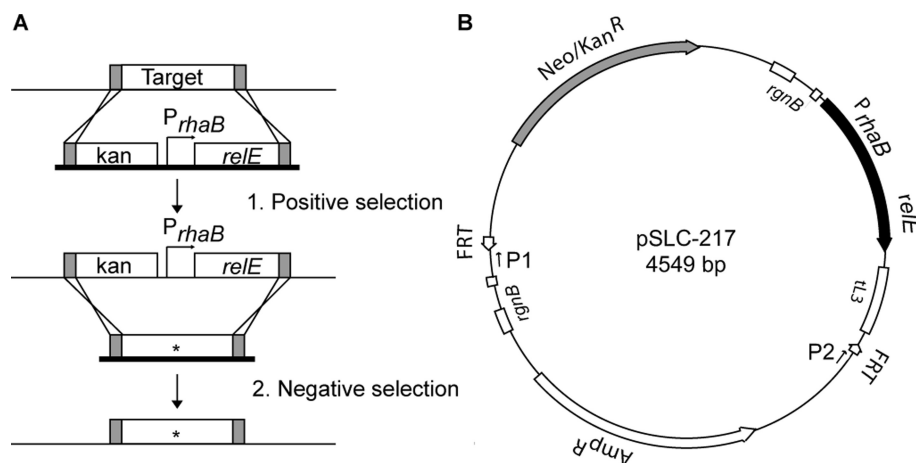


Figure 1. Design and characterization of a negative selection module. (A) Schematic of a two-step cloning strategy for allelic replacement using positive and negative selection. *Step 1.* A PCR product (depicted with a thick black baseline) containing a positive–negative selection cassette (kan- P_{rhaB} - $relE$ here) flanked by sequence (shaded gray) homologous to the targeted gene is subjected to double crossover recombination (indicated by crossing lines) to replace the gene. Selection for this replacement is done using the positive selection marker (kanamycin here). *Step 2.* Another PCR product (thick black baseline) containing the same or a different allele (indicated by *) of the target gene, also flanked by sequence homologous to the targeted locus is subjected to double crossover recombination to reintegrate the original locus. The resulting genome carries no residual markers or DNA scar. Selection for this reintegration is done using the negative selection marker (plating on rhamnose in this case). (B) Schematic diagram of template plasmid pSLC-217. The negative selection module (P_{rhaB} driving $relE$) is shaded in black. The positive selection module (neo , mediating kanamycin resistance) is shaded gray. Transcriptional terminators are indicated by open boxes. Universal priming sites P1 and P2 are shown as arrows. Other features (FRT sites, Amp^R gene) are indicated by open box arrows.

tion stringency in all tested strains, exceeding all other reported negative selection systems usable in unmodified lab isolates of *E. coli* by up to 60-fold, and approaching theoretical predictions of maximum stringency based on mutation rates. The high negative selection stringency enabled its use in traditional phage-mediated generalized transduction, in which an additional benefit is the generation of unmarked transductants using a wild-type donor. Furthermore, we were able to use the negative selection system to perform a modified phage-free ‘transduction’ (which we term generalized allelic exchange) using transformed whole genomic DNA between two uropathogenic isolates of *E. coli*, an otherwise difficult problem due to the lack of transducing phages for clinical strains (10). Therefore, this negative selection system is a convenient and powerful addition to the genetic toolbox for all *E. coli* strains that, in combination with traditional antibiotic markers, enables the creation of arbitrary unmarked mutations in all tested strains of *E. coli* and other Enterobacteriaceae and allows for definitive genetic experiments even in clinical strains. It further enables new applications for generalized allelic exchange between different strains.

MATERIALS AND METHODS

Media and culture conditions

LB and M9 agar were used as complex and minimal media, respectively. For plasmid maintenance and selection of positive selection antibiotic markers, media was supplemented with antibiotics (Sigma, Singapore) at the following concentrations: ampicillin (100 μ g/ml), kanamycin (50 μ g/ml), chloramphenicol (20 μ g/ml) and tetracycline (10 μ g/ml). M9 agar was supplemented with 0.2% glucose to repress toxin gene expression or with 5% sucrose (for *cat-sacB* se-

lection), 0.2% rhamnose (for all P_{rhaB} -driven toxins) or 0.2% arabinose (for P_{araB} -driven toxins). Maintenance of plasmids carrying the toxin genes was carried out in LB supplemented with 2% glucose.

Bacterial strains and plasmids

Relevant characteristics of bacterial strains and plasmids used in this study are listed in Supplementary Table S1. The template plasmids pKD3 and pKD4 were obtained from AddGene and were modified in accordance with the AddGene Material Transfer Agreement for non-commercial purposes.

tetA negative selection

Overnight cultures of UTI89 carrying *tetA* on plasmids or integrated into the chromosome were diluted 1:100 into fresh media and grown to an optical density at 600nm (OD_{600}) of 1. Serial 10-fold dilutions were plated to quantify the number of surviving bacteria (CFU) on selective and non-selective (LB) plates. Selective plates were supplemented with tetracycline (for positive selection) or 2–4 mM $NiCl_2$ (for negative selection).

Plasmid construction

All plasmids used in this study are listed in Supplementary Table S1. To facilitate cloning, we first inserted several restriction sites (AatII, SmaI and SacII) into the AfeI site of pKD4, generating pSLC-243. Toxic genes were amplified from *E. coli* MG1655 (*relE*, *chpBK*, *mqsR*, *higB*, *yafQ* and *yhaV*), *Pseudomonas aeruginosa* (*tse2* (11)) or *E. coli* UTI89 (phage λ holin (12)) genomic DNA by PCR then cloned using NdeI and BamHI sites (in pAH120) or NdeI and NheI

(in pAH150 (13)) to place them under the control of the *rhaB* or *araB* promoter, respectively. Each promoter-toxin module, including the flanking tL3 and *rgnB* transcription terminators from pAH120 or pAH150, was cloned by blunt ligation into the SmaI site of pSLC-243 to generate template plasmids (pSLC-17 containing kan-*P_{rhaB}-relE* is shown in Figure 1B). All plasmids were propagated in *E. coli* strain BW23473 (*pir+*, low copy number) or BW23474 (*pir116*, high copy number). The sequences of all negative selection cassettes are available upon request.

Recombineering

All genomic manipulations were carried out using previously described Red recombinase recombineering protocols optimized for lab or clinical isolates of *E. coli* (14), with minor modifications. The positive–negative selection cassette was amplified by PCR using the P1 and P2 universal primers for pKD4. Primers used for cloning are listed in Supplementary Table S8. Primers used for *fimH* constructs were as previously described (15). Each primer was synthesized with an extra 50 bp of sequence at the 5' end that was homologous to the targeted genomic locus. Red recombinase was expressed from the pKM208 (in *E. coli*) or pKD46 (in *S. enterica* Serovar Typhimurium 14028S) helper plasmids. Overnight cultures (LB-ampicillin, 30°C with agitation) of the target strain were diluted 1:100 into fresh media, grown at 30°C with agitation to OD₆₀₀ = 0.2–0.3, and induced with 1 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG) (pKM208) or 0.2% arabinose (pKD46) for 30–45 min. The induced cells were then heat shocked for 15 min at 42°C, followed by swirling in an ice water bath for 15 min. Cells were harvested by centrifugation (5000 rpm for 10 min), washed two times with cold (4°C) water, re-suspended in 1/100 of the original culture volume of cold 10% glycerol, and frozen in 50 μl aliquots as competent cells. To perform double crossover homologous recombination, 1 μg of a PCR product was transformed by electroporation into thawed competent cells using 1mm electroporation cuvettes in a GenePulser XCEL set to an output voltage of 1500 V with 25 μF capacitance and 400 Ω resistance (Bio-Rad, Singapore). Cells were then recovered in LB at 37°C with shaking for 2 h followed by static incubation (no shaking) for 2 h, plated on selective plates, and incubated overnight (for kanamycin selection) or for 24–48 h (for rhamnose selection) at 37°C.

Diagnostic PCR and Sanger sequencing

Colony PCR reactions were used to check for insertion/replacement of the selection cassette. The first primer pair was used to specifically detect the locus of desired homologous recombination and the second primer pair was to detect the presence or absence of the specific toxin gene at the desired locus. Primers used are listed in Supplementary Table S8. Primers for *fimH* constructs were *uti8* + 4913333 and *uti8* – 4915222 as described previously (15). The strains were confirmed using Sanger sequencing of the diagnostic PCR products with the same primers used for amplification (1st Base, Singapore).

Immunoblot

Bacteria from 1 ml of culture at OD₆₀₀ = 1.0 was harvested by centrifugation (14 000 rpm, 1 min). The bacterial pellet was resuspended in 300 μl of 4× sodium dodecyl sulfate (SDS) loading buffer (50% glycerol, 0.5% SDS, 0.25 M Tris, mercaptoethanol and bromophenol blue). Ten microliters of this sample was analyzed by sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) (15% polyacrylamide gel). Proteins were transferred to a nitrocellulose membrane using a Trans-Blot® SD Semi-Dry transfer cell (Bio-Rad, Singapore). Blots were probed with anti-GroEL antibody (Sigma, Singapore) at a 1:6000 dilution in Tris-Borate-EDTA (TBE) with 5% skimmed milk to assess loading quantities. Blots were probed with custom anti-OmpC monoclonal antibodies (Abmart Inc., Shanghai, China). Antibody 4H15 (1:1000 dilution) recognizes both the wild type and mutant OmpC protein (mutated in loop 5 residues) while 4K3 (1:1000) was raised against the OmpC loop 5 residues and thus does not detect the mutant OmpC protein. Secondary antibodies used were ECL™ anti-rabbit IgG Horseradish peroxidase (HRP)-linked whole antibody (1:10 000) for anti-GroEL antibody and ECL™ anti-mouse IgG HRP-linked whole antibody (1:10 000). Blots were visualized using Amersham HRP substrate in a Chemidoc machine (Bio-Rad, Singapore).

Whole genome sequencing

Genomic DNA from strain SLC-561 (UTI89 *fimH*::FRT; *m1aA*::*m1aA*^{UTI89}) was prepared from 1mL of an OD₆₀₀ of 1 culture using the Wizard® Genomic DNA Extraction Kit (Promega) and prepared for Illumina sequencing using the TruSeq DNA library preparation kit (Illumina) according to the manufacturer's recommendations. This library was sequenced on a HiSeq 2000 (2 × 76 bp reads), yielding 5 680 917 read pairs (~170× coverage). After quality filtering, fastq files were analyzed with cortex_var (version 1.0.5.20) from the Cortex package (16) using a joint analysis with UTI89 (NCBI accession number NC_007946) as the reference genome; the resulting VCF files indicated the known deletion at the *fimH* locus but no other variants (insertions, deletions, or single nucleotide polymorphisms) that passed quality filters. The sample was further tested for large (>100 bp) structural variations relative to the UTI89 reference genome using the SVRE program (Sukumaran and Chen, unpublished data) ; no large structural variations besides the *fimH* deletion were found that differed from wt UTI89.

Hemagglutination titers

These were performed as previously described (15). Briefly, 1 ml of an OD₆₀₀ = 1.0 suspension of bacterial cells in PBS was gently pelleted (4000 × g, 2 min) and resuspended in 100 μl phosphate buffered saline (PBS). Twenty-five microliters of this was serially diluted in a row of a 96-well V-bottom plate (Corning #3897) where each well contained 25 μl of PBS (with or without mannose) (dilution range 1:2 to 1:4096). Twenty-five microliters of guinea pig red blood cells in PBS were added to each well. The plate was gently agitated and incubated overnight at 4°C. The HA titer

reported was the greatest dilution of cells that resulted in visible clumping of erythrocytes.

SDS–EDTA sensitivity

Overnight cultures were diluted 1:100 into fresh media and grown to $OD_{600} = 1$. Serial 10-fold dilutions were plated on different media to quantify the number of surviving bacteria (CFU). Selective plates were supplemented with 0.5% SDS and between 0.75–2.25 mM ethylenediaminetetraacetic acid (EDTA) and non-selective plates were supplemented with 0.5% SDS, with or without kanamycin.

Selection stringency

Overnight grown cultures were diluted 1:100 into fresh media and grown to $OD_{600} = 1$. Serial 10-fold dilutions were carried out and plated on non-restrictive (plain LB for all strains; also including antibiotic plates for antibiotic resistant strains) media to quantify the number of surviving bacteria (CFU) in the original suspension. 10^{10} cells (10 ml culture of $OD_{600} = 1$, centrifuged and re-suspended with 200 μ l PBS) were spread on appropriate restrictive plates to check for negative selection (on M9-rhamnose or M9-arabinose for strains carrying the negative selection cassette) and positive selection (LB-kanamycin for strains not carrying the kanamycin positive selection marker) stringency. Selection stringency was determined by dividing the calculated CFU of the original inoculum on restrictive plates by the average number of CFU that grew on non-restrictive plates.

Luria–Delbruck fluctuation test

Strains were grown overnight on a non-restrictive plates (LB for MG1655 and LB-kanamycin for SLC-568) and one single colony was resuspended in 50 μ l of PBS. Twenty microliters of this suspension was diluted 200 \times and aliquoted into the central 60 wells of each of 2 \times 96-well plates per strain. Each well received 20 μ l of the diluted suspension (approximately 10^2 or 10^4 CFU/well) and 180 μ l of non-restrictive media (LB for MG1655 and M9 with 0.2% glucose for SLC-568). The plates were then incubated at 37°C with shaking (220 rpm) for 6 h (MG1655, starter cultures with 10^4 CFU/well) or 18 h (SLC-568, starter cultures with 10^2 CFU/well). (As a control, we also carried out 8-h incubations with SLC-568; the resulting data was similar to that from experiments using an 18-h incubation). Ten wells from each plate were used to calculate total CFU per well by serial dilutions. The remaining 50 wells from each of the two plates were pelleted (100 total) and resuspended in 200 μ l of selective medium (LB-kanamycin for MG1655 and M9 with 0.2% rhamnose for SLC-568). The plates were then incubated overnight at 37°C with shaking (220rpm) after which the cells were again pelleted and resuspended in 10 μ l of restrictive medium and spotted on restrictive agar plates. The agar plates were incubated at 37°C for 24–48 h. The number of wells which gave rise to colonies growing on restrictive agar plates was counted and the mutation rate of the

strains was determined using the following formula (17);

$$\text{Mutation rate} = -\ln \frac{(\text{no. of negative samples}/\text{total no. of samples})}{\text{total no. of cell divisions}}$$

Generalized transduction

Transduction using P1_{vir} was performed in MG1655 using standard protocols (18). A transducing lysate generated from wild type MG1655 was used to infect SLC-568 (MG1655 *hsdS::kan-P_{rhaB-tse2}*) followed by plating on M9-rhamnose as described above in the ‘Recombineering’ section.

Generalized allelic exchange

Isolated genomic DNA from *E. coli* CFT073 was sheared to an average length of 10 kb fragments by sonication (Branson Digital Sonifier Disruptor 450) (2 \times 5 s pulses at 10% power on ice). Ten micrograms of sheared DNA was transformed into competent SLC-557 (UTI89 *ompC::kan-P_{rhaB-reLE}*) expressing red recombinase. Cells were recovered in LB at 37°C with shaking for 2 h followed by static incubation (no shaking) for 2 h then plated on M9-rhamnose agar plates, and screened via PCR and sequencing for allelic exchange at *ompC* locus using P21 and P22 primer set.

BLAST analysis of type II toxins

Type II toxin genes present in MG1655 were taken from (19) and from the NCBI RefSeq annotation of the MG1655 genome (NC.000913). Finished *E. coli* genome sequences were downloaded from NCBI RefSeq (Supplementary Figure S2). Each non-K12 genome was used as the database to search for each toxin gene using the blastn (using the toxin gene sequences) and tblastn (using the toxin protein sequences) programs (BLAST 2.2.28+) with default parameters. Examination of the blastn and tblastn results showed that a cutoff of a blastn alignment over 80% of the toxin gene length captured all genes with simultaneously >88% DNA identity and >84% protein identity over at least 80% of the toxin gene (or protein) sequence. This 80% blastn alignment cutoff also included four additional gene/genome combinations which had potential frameshift mutations (*yafQ* and *yafO* in SMS_3_5; *cbtA* in O55_H7 RM12579 and P12b) but for which the entire gene sequence was otherwise present. We thus used 80% alignment over the toxin gene length by blastn as the cutoff to call toxins as present or absent in a given genome.

RESULTS

Existing negative selection systems are not practical for a clinical isolate of *E. coli*

We first tested those existing negative selection systems that do not require host genotype modification for their ability to function in single copy in the chromosome of an unmodified clinical uropathogenic isolate of *E. coli*, UTI89. Only two systems, *tetA* (1) and *sacB* (20), could potentially be used in UTI89. Selection against Tet resistance by fusaric

acid is known to require optimization in different strains of *E. coli* and may not be usable in others (21); we were unable to demonstrate negative selection mediated by the *tetA* gene found on pBR322 in UTI89 after optimization of fusaric acid and salt concentrations (data not shown). It has also been reported that Ni²⁺ salts can be used to select against Tet^R strains and that the selectivity can in some cases exceed 1 in 10⁶ cells (1), but the selection power and optimal Ni²⁺ concentration varied between different *E. coli* strains, typically requiring above 6mM NiCl₂. We found that, indeed, growth of UTI89/pBR322 (Tet^R) relative to UTI89 was inhibited by 4 orders of magnitude at an optimal concentration of 3 mM NiCl₂. However, pBR322 has an estimated copy number of 20/cell. In order to use this system for chromosomal manipulation, it must work at a copy number of 1/cell. We found that the sensitivity of Tet^R UTI89 to NiCl₂ decreased with copy number, to the point that no NiCl₂ sensitivity was found with UTI89::tet (containing a single copy chromosomal *tetA* gene) relative to UTI89 (Supplementary Figure S1), as was also seen with DH10B (a lab adapted strain of *E. coli*) (1).

Gram negative bacteria carrying the *B. subtilis sacB* gene accumulate toxic metabolites when grown in the presence of sucrose. The *sacB* gene is widely used for negative selection (20,22,23), but it suffers from relatively high false positive rates for chromosomal manipulations due to spontaneous suppressor mutations that inactivate *sacB* (24). This has the overall effect of reducing its stringency to $<2.3 \times 10^{-5}$ (20). However, in UTI89, a single chromosomal copy of *cat-sacB* mediated negative selection with a relatively weak selection stringency of 3.40×10^{-2} . Recently, a system combining both the *tetA* and *sacB* genes has been reported with improved stringency (3) but is very sensitive to selection conditions and therefore requires strain-specific optimization. Therefore, a negative selection system immediately usable in clinical strains of *E. coli* is lacking.

Design of a modular negative selection cassette based on inducible toxins

We sought to design a new negative selection system that would both provide high stringency and not require strain-specific optimization. Negative selection using the *mazF* toxin gene has been demonstrated in single strains of *Bacillus subtilis* (25) and *Clostridium acetobutylicum* (26), though no data on stringency was reported. These reports suggested that isolated toxins from toxin-antitoxin (TA) systems could be generally useful as negative selection markers, even when used in heterologous species (*mazF* is an *E. coli* gene, yet seems to function in Gram positives). We therefore asked whether a general system providing tightly controlled induction of a toxin gene could be used in *E. coli* and other Enterobacteriaceae. The *E. coli* MG1655 (a lab-adapted fecal isolate) chromosome is known to carry numerous TA systems (19); we focused on Type II TA systems (in which both toxin and antitoxin are proteins) because they are more easily identified and therefore likely to be more completely annotated. To develop a system for use in clinical isolates (in particular, UTI89), we reasoned that the TA system we used should be absent from the UTI89 genome. Using BLAST analysis, we identified 12 candidate

toxins present in MG1655 but not in the UTI89 genome (Supplementary Table S2). Further analysis using 55 non-K12 full genome sequences available at NCBI RefSeq highlighted 7 MG1655 toxins (*rnlA*, *ypjF*, *ykfI*, *ydaS*, *yjhX*, *relE* and *mqsR*) that were not found in more than half of non-K12 genomes (Supplementary Figure S2 and Supplementary Table S2).

We created a negative selection cassette consisting of the *relE* toxin (chosen for its small size, to minimize the PCR length in subsequent steps and its potential for inactivation by mutation) driven by the *rhaB* promoter (chosen for its 'tight' control; it is induced by growth on rhamnose and repressed by growth on glucose). To further ensure deliberate control of expression, we included the *tL3* and *rgnB* terminators flanking the P_{*rhaB-relE*} to guard against read-through transcription from other promoters; and we retained the native *rhaB* ribosome binding site. For use in making unmarked chromosomal mutations, we combined this negative selection cassette with a kanamycin resistance gene, a traditional positively selectable antibiotic marker. We built the construct as a derivative of the well-designed cloning system described by Datsenko and Wanner (14) based on Lambda Red recombinase and antibiotic selection using the pKD4 template plasmid. We refer to the derived template plasmid as pSLC-217 and to the combined positive-negative selection cassette as kan-P_{*rhaB-relE*} (Figure 1B).

Efficient, scarless and markerless manipulation of the UTI89 chromosome

To validate our system, we first used the kan-P_{*rhaB-relE*} cassette to perform a two-step allelic replacement of the *ompC* gene at its native locus in UTI89. We used a Red recombinase protocol optimized for clinical isolates of *E. coli* (22) to replace the native *ompC* gene by homologous recombination with a linear PCR product (containing kan-P_{*rhaB-relE*} flanked by homology arms targeting the *ompC* locus), using positive selection on kanamycin (Figure 1A, step 1). We then used PCR products for the wild type UTI89 *ompC* and a mutated *ompC* (carrying mutations in extracellular loop L5) allele in a subsequent Red recombinase-mediated double-crossover replacement of the integrated positive-negative selection cassette, using negative selection on rhamnose (Figure 1A, step 2). Clones were screened using PCR for the entire *ompC* locus as well as for specific insertion of the *relE* toxin gene at the *ompC* locus (Supplementary Figure S3A). Positive clones identified by PCR were verified by testing for loss of kanamycin resistance, restored ability to grow on rhamnose (Supplementary Figure S3B), and Sanger sequencing of the entire *ompC* locus. The efficiency for isolating the correct recombinant during the counterselection step was 87–100% as calculated by testing 30 colonies from each counterselection step by PCR, growth on rhamnose, and loss of the positive selection marker. To further confirm that the restored *ompC* locus was functional, we tested for OmpC protein expression using immunoblotting. We saw OmpC protein expression in whole cell extracts for the parental wild type strain as well as both recombined strains (carrying a wt and L5 mutant *ompC*, respectively), but no OmpC band in the intermediate knockout strain. Furthermore, using an antibody specific to

the L5 loop of OmpC, we saw detection only in the parental wild type (wt) and the strain with the restored wt *ompC* allele but not in the knockout or the strain carrying the L5 mutant allele (Supplementary Figure S3C). Therefore, the combination of our negative selection cassette with a traditional positive selectable marker has enabled us to create a functional, scarless mutation directly in the *ompC* locus of UTI89.

Creation of markerless mutations at arbitrary loci in UTI89

We verified that the kan- P_{rhaB} -*relE* positive–negative selection cassette could be used to manipulate arbitrary genes in UTI89 by performing gene deletion and complementation at three other loci (Supplementary Table S3). We used the two-step integration/knockout followed by replacement of the integrated dual selection cassette, using negative selection on rhamnose to create isogenic mutant strains or unmarked mutations with defined deletion junctions (Supplementary Tables S1 and S3). We successfully deleted and reintegrated the *miaA* and *hsdS* genes, verifying each by both diagnostic PCR and Sanger sequencing of the locus. Functional confirmation of *miaA* knockout and reintegrated constructs was also done using an SDS–EDTA sensitivity assay (Supplementary Table S4). We also made unmarked versions of previously reported strains carrying point mutations in the *fimH* gene (15), eliminating the downstream linked kanamycin gene previously required for the second step of cloning. All clones were confirmed to contain the predicted mutations by diagnostic PCR and sequencing of the new junctions created by recombination. Furthermore, all of the unmarked *fimH* mutants were verified to have similar type 1 pilus function to their marked counterparts based on hemagglutination of guinea pig red blood cells (Supplementary Table S5).

No off-target mutations are created

Expression of Red recombinase can induce a 10-fold increase in spontaneous mutations (22). As the goal is to perform well-defined chromosomal manipulations, the two rounds of Red recombinase mediated recombination are a potentially large problem for off-target, unintended mutations. To mitigate this, we followed the recommendations of Campellone and Murphy to minimize Red recombinase induction to 30–45 min prior to generation of competent cells in both recombination steps (22). To further address whether off-target mutations were being generated, however, we performed whole genome sequencing on a strain that had undergone a chromosomal knockout then reintegration of the *miaA* gene. Using Cortex (which detects SNPs and small indels) (16), we found no mutations compared with the parental strain (prior to the initial knockout of *miaA*). We also used a new sensitive structural variation caller (which detects larger inversions, duplications, insertions, deletions and other more complex rearrangements; manuscript in preparation) and again found no differences in genomic rearrangements compared with the parental strain. While we only examined one strain, the complete absence of any detectable off-target mutations in over 5 Mbp of sequence in this strain indicates that unintended muta-

tions are not generally being created at a high rate during these engineering steps.

Generalization of negative selection for other host strains by module replacement

The positive–negative selection cassette we created was built from several well known genetic elements (promoters, terminators, RBS sequences and antibiotic markers). In the interest of generality, potentially enabling use in other *E. coli* strains as well as non-*E. coli* bacteria, we first verified that we could simply replace the promoter and antibiotic marker without having to perform any optimization of sequences or selection conditions. We created P_{araB} -*relE*, where the negative selection would be carried out with growth on arabinose. We also switched the parental template plasmid to pKD3, encoding a chloramphenicol positive selection cassette (Supplementary Table S3). Indeed, switching between P_{rhaB} and P_{araB} or chloramphenicol and kanamycin had little effect on the efficiency of the negative selection step (based on limited testing of 7–8 colonies by PCR and loss of the positive selection marker) when used in a similar two-step allelic replacement at the *ompC* locus of UTI89 using the previously defined selection conditions we used for UTI89 (although the antibiotic and sugar were altered as necessary).

Given that several toxins encoded by MG1655 besides *relE* were also potential candidates for use in UTI89, but that may be variably present in other *E. coli* strains, we next tested whether we could also simply replace the toxin gene in a modular fashion. We tested several other toxins (*mqsR*, *higB*, *yhaV*, *yafQ* and *chpBK*) with P_{rhaB} , using the same cloning sites we used for *relE* to make identical fusions of the start codon of each toxin gene with the promoter and ribosome binding site of our template plasmid. We further reasoned that alternative toxins, such as those from phage or antibacterial secretion systems, might function effectively both in *E. coli* (thus enabling use in MG1655 and other lab strains) as well as in other bacteria. We therefore also included the holin gene (324 bp) from phage λ (12) in UTI89 and the *tse2* type 6 secretion system toxin (477 bp) from *Pseudomonas aeruginosa* (11). These positive–negative selection cassette variants were again tested using the same two-step knockout/replacement protocol without modification of transformation or growth conditions. We found that, except for variants containing *higB* and *yafQ* (see below), all variants were usable for creating markerless mutations in UTI89.

Finally, we tested our system in different bacterial strains. We replaced the *ompC* gene in CFT073 (another uropathogenic *E. coli* clinical isolate (27)), EDL933 (an enterohemorrhagic *E. coli* isolate (28)), TOP2515 (a cystitis isolate that has never been reported to be manipulated on its chromosome (29)), and *S. enterica* serovar Typhimurium 14028S (30) with the *ompC* allele from UTI89 (Supplementary Table S3). In addition, we replaced the *hsdS* gene in *E. coli* MG1655 with the *hsdS* gene from UTI89 (Supplementary Table S3) using the *P. aeruginosa* *tse2* gene as the negative selection system toxin. All replacements were successful as verified by PCR, growth phenotypes, and Sanger sequencing. Furthermore, for all variant

cassettes and all strains tested, we used the same amplification primers, transformation conditions, and negative selection conditions (M9 with 0.2% rhamnose) that we had originally used for UTI89. Thus, our positive–negative selection cassette is indeed general and needs no optimization when used in different clinical isolates of *E. coli* and other Enteric bacteria.

Inducible toxins mediate high negative selection stringency

We noted in these manipulations that, in general, very few background/false positive colonies grew on rhamnose plates (Supplementary Figure S3B). Furthermore, two toxins, *higB* and phage λ holin, could suppress growth on rhamnose when present on a high copy plasmid but not when present in single copy on the chromosome (Supplementary Figure S4). These data suggested that different toxins vary in selection stringency, which might be useful for creating negative selection cassettes in organisms lacking tightly regulated promoters. We first tested selection stringency by calculating the fraction of viable colony forming units (CFUs) that were able to grow under restrictive conditions (smaller numbers indicate higher stringency). This test measures the frequency of colonies growing under restrictive conditions (as opposed to the actual mutation rate – see below), is naturally related to the background colonies one would observe during a typical cloning procedure, and has been used previously in rapid screens of mutability (31–33). We therefore refer to these values as ‘stringency frequencies’. We plated bacteria carrying a positive–negative selection cassette in single copy in the chromosome onto each of LB, LB-antibiotic and M9-rhamnose or M9-arabinose plates. On non-restrictive plates, we quantified bacteria by plating serial dilutions of the original culture, while on restrictive plates, we concentrated 10^{10} CFU from 10 ml of $OD_{600} = 1$ culture and plated them all. For reference, the parental strains and the strains with the replaced alleles (both sensitive to kanamycin or chloramphenicol) were also plated to the same plates to test the stringency of positive selection on antibiotics (Figure 2A). We found that positive selection markers mediated very strong selection (stringency frequencies of 10^{-9} – 10^{-10}) in UTI89 (Figure 2A). Among negative selection markers, different toxins, when driven by the P_{rhaB} promoter, did indeed vary in their strength of selection, but the toxins that were usable for negative selection in single copy on the UTI89 chromosome had stringency frequencies of 10^{-7} – 10^{-8} (Figure 2A and Supplementary Table S6). Because the insertion site of a negative selection cassette may affect its relative copy number due to the timing of DNA replication, and copy number influences expression levels and thereby potentially stringency, we repeated these tests with the kan- P_{rhaB} -*relE* selection cassette integrated into different loci in UTI89. In general, stringency did not vary much across different loci within UTI89 (Figure 2B, Supplementary Figure S5A), across different *E. coli* strains, or in *Salmonella* (Figure 2C, Supplementary Figure S5B). Notably, the *tse2* toxin (11) from the *Pseudomonas* Type VI secretion system had the highest stringency frequency (1.96×10^{-8}) at the *hdsS* locus in MG1655 (Figure 2C). In comparison, the highest stringency for any negative selection system reported to date for

unmodified *E. coli* strains (MG1655 or its close relatives, all lab-adapted) is 6×10^{-7} (presumably also a stringency frequency) using a combined *tetA*–*sacB* cassette (3).

We verified that CFU quantification on rhamnose plates was accurate by comparing CFU calculated by serial dilutions plated to both plain LB and M9-rhamnose; no differences were seen for any of the wild-type strains (Supplementary Table S7). Furthermore, we also tested a control in which the kanamycin gene and P_{rhaB} were integrated into the chromosome without an accompanying toxin gene (SLC-658); this also showed no toxicity from growth on rhamnose compared with plain LB (Supplementary Table S7).

To further validate the high stringency of our system, we explored several additional lines of reasoning for the strain carrying the *tse2* toxin in MG1655 (SLC-568). First, we repeated our initial stringency frequency test by plating a higher number (10^{11} CFU; total culture volume 100 ml, $OD_{600} = 1$) of cells onto M9-rhamnose plates; we found a similar stringency frequency (1.11×10^{-8} ; Supplementary Figure S6), representing growth of at least several hundred colonies in each of three replicates of this experiment.

Second, we estimated the theoretical maximum possible stringency (i.e. lowest mutation rate) based on chromosomal replication error rates for MG1655, assuming that only mutations within the negative selection cassette itself could confer breakthrough growth. The range of mutation rates for MG1655 has been reported to be between 1×10^{-3} and 3.3×10^{-3} /genome/generation (34,35); this translates into 2.2×10^{-10} to 7.1×10^{-10} mutations/nucleotide/generation. The P_{rhaB} -*tse2* cassette is 811 bp; therefore the expected mutation rate is 1.8×10^{-7} to 5.8×10^{-7} mutations/cassette/generation. Not all mutations would inactivate the toxin or disrupt its expression; this fraction is difficult to quantify, but we take as limits the expected fraction of coding mutations that would be expected to be nonsense (5%) (36) and the fraction of *lacI* mutations that are dominant (40%) (37) (this range encompasses the estimated fraction (20%) of missense mutations (~70% of all *de novo* mutations) that are strongly deleterious (overall 14% of all *de novo* mutations)) (36). This gives an expected inactivation rate between 0.9×10^{-8} and 2.3×10^{-7} /cassette/generation. In addition the toxin could be inactivated by a spontaneous large deletion, which occurs at a rate of between 1.79×10^{-9} and 2.5×10^{-7} per generation in *E. coli* (38–40). These calculated mutation rate values are now *rates* at which colonies growing on rhamnose could be generated, and should be considered distinct from the stringency frequencies measured and discussed above.

Measurement of experimental *rates* of mutation (as opposed to stringency *frequencies*) is more involved; the relevant issues have been previously elucidated by Luria and Delbruck (17). We carried out a Luria–Delbruck fluctuation test to directly measure experimental *rates* of mutation to breakthrough growth on rhamnose. Using two different starting small culture sizes (see ‘Materials and Methods’ section), we calculated a mutation rate toward growth on rhamnose of 1.42×10^{-8} , within the theoretical range based on per-nucleotide mutation rates calculated above. As a control, we also performed a Luria–Delbruck fluctuation test for mutations conferring resistance to kanamycin. The

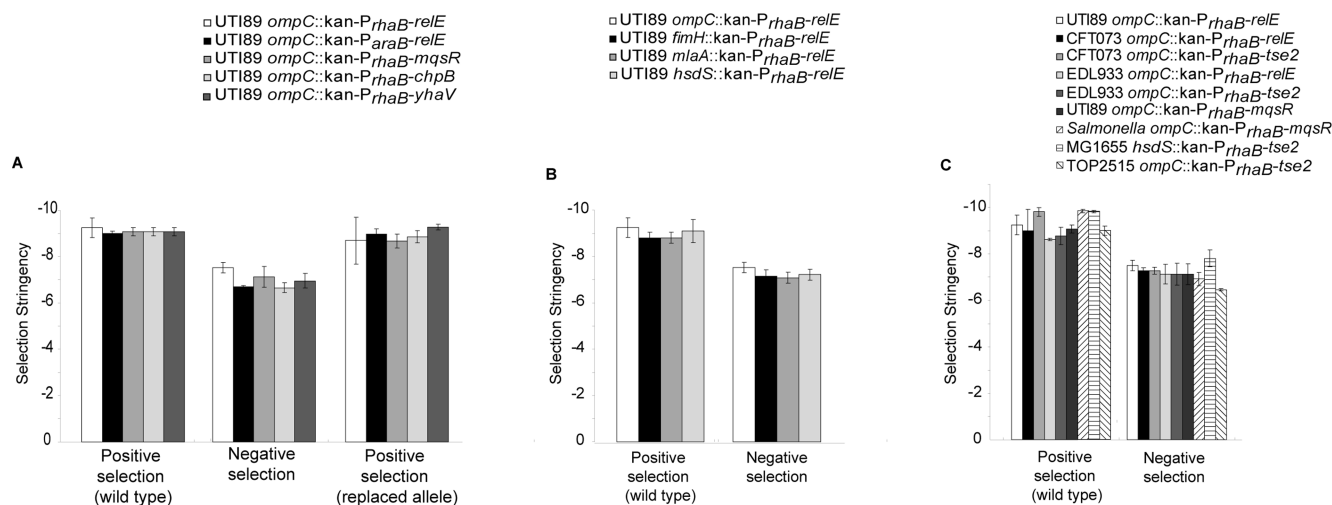


Figure 2. Selection stringency tests. (A) Selection stringency tests for different toxins. Data shows the selection stringency, which is calculated as: CFU/ml on restrictive agar divided by CFU/ml on non-restrictive agar. First set of bars represents the positive selection stringency for wild type (UTI89) strains tested in each experiment. The middle set of bars represents the negative selection stringency for each strain carrying the dual selection cassette integrated into the *ompC* locus in UTI89. The last set of bars represents positive selection stringency for the strains in which the selection cassette has been replaced with the wild type *ompC* allele in a subsequent negative selection step. (B) Stringency does not vary with chromosomal locus. Data shows the positive selection stringency for wild-type strains (on the left) followed by the negative selection stringency for strains carrying the *kan-P_{rhaB}-relE* cassette at different loci in UTI89 (on the right). (C) The positive-negative selection cassette is usable in different strains of *E. coli* and in *Salmonella*. Toxins and strains are as indicated. For graphs in (A–C), data represents the average of the log-transformed values of selection stringencies with error bars indicating standard deviation of the log-transformed stringency values calculated from three independent biological replicates.

fluctuation test gave a mutation rate to resistance (1.11×10^{-10}) that was lower than our rate for mutation of our negative selection cassette, in keeping with the better stringency frequency of kanamycin (2.30×10^{-10}) as measured above.

As a final verification of the reproducibility of the high stringency of our negative selection cassettes, we performed the same cloning to create a separate but isogenic strain to SLC-568 (designated SLC-657) and assayed its frequency of breakthrough growth on rhamnose (by plating 10^{10} CFU), again arriving at 2.38×10^{-8} (Supplementary Figure S6). We therefore conclude that the negative selection stringency (mutation frequencies) of the *tse2* system in MG1655 is indeed better than previously published negative selection systems and that its mutation rate is very close to the theoretical minimum mutation rate expected for a system of this size.

Negative selection by inducible toxins enables unmarked allelic exchange

We then asked whether high negative selection stringency could be leveraged to enable ‘difficult’ applications such as generalized transduction, where positive selection markers are traditionally required (though recently P1 transduction has been performed using a *tetA-sacB* negative selection cassette (3)). We performed generalized transduction with a phage P1 lysate (generated from unmodified wild-type MG1655) to complement an *hsdS* knockout (replaced with *kan-P_{rhaB}-tse2*; host strain MG1655 (SLC-568)); selection on rhamnose resulted in growth of >2000 transductants (out of $\sim 10^9$ plated CFU), of which 16/16 (100%) were positive for restoration of the wild-type *hsdS* locus by PCR. In contrast, no colonies were observed when the P1 phage lysate itself was plated, and 200 colonies were

observed on rhamnose when only SLC-568 (the recipient strain) cells were plated. The estimated efficiency of >90% compares favorably with the 56–63% efficiency reported using *tetA-sacB* (3), consistent with the higher stringency of our system. Transduction using negative selection improves on traditional generalized transduction because the resulting transductants contain no residual antibiotic marker or scar and because unmarked, unmodified strains can be used as the donor.

However, clinical strains are often difficult to manipulate using generalized transduction, either due to phage specificity (10) or low efficiency. We were unable to perform a restoration of the *ompC* locus in UTI89 using P1-mediated generalized transduction from unmodified MG1655, and we were unable to create a P1 transducing lysate from either UTI89 or CFT073. We speculated that we could circumvent these difficulties by mimicking transduction using transformation of sheared genomic DNA (Figure 3A). Such a strategy would remove limitations imposed by the need to create a transducing lysate from the donor strain that could successfully infect the recipient strain. We refer to this phage-free ‘transduction’ as generalized allelic exchange. Reimplementation of generalized transduction in this way would not only expand its potential use to strains without transducing phage, but could also potentially lead to a novel application for strain hybridization we term ‘mass allelic exchange’ (see ‘Discussion’ section). We directly transformed sheared genomic DNA from CFT073 by electroporation into SLC-557 (UTI89 *ompC::kan-P_{rhaB}-relE*) expressing Red recombinase. Selection on rhamnose followed by PCR screening resulted in identification of the desired recombinant, where CFT073 genomic DNA had replaced the *ompC* locus in UTI89, in 2/55 colonies (3.6%). We sequenced the DNA flanking *ompC* in these two clones and

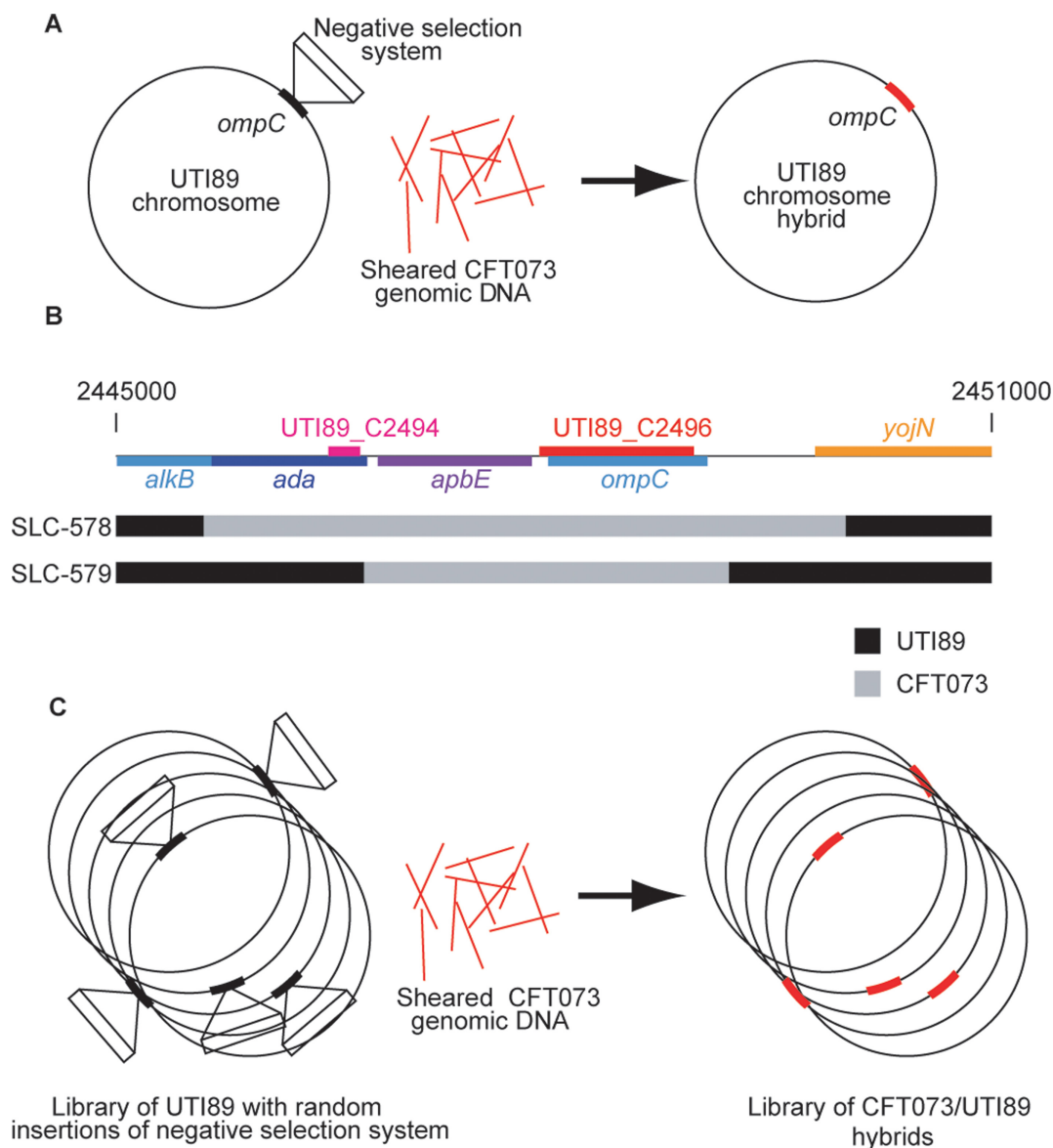


Figure 3. Generalized allelic exchange. (A) Schematic of the generalized allelic exchange strategy. A negative selection marker (open rectangle) is inserted into the recipient strain (UTI89 here) at a defined locus (thick black arc, *ompC* in this case). Sheared whole genomic DNA from another strain (CFT073 here, red lines) is transformed into UTI89, and growth on restrictive conditions allows selection for replacement of the targeted locus (thick red arc) with no residual marker or scar. (B) Recombination breakpoints for *ompC* allelic exchange from CFT073 to UTI89. A schematic of the gene organization of 6kb surrounding the *ompC* gene in UTI89 is shown at the top. The rectangles below depict recombination breakpoints of two clones (SLC-578 and SLC-579) as determined by Sanger sequencing of the *ompC* gene and flanking sequence. In each, a central region of CFT073 sequence encompassing the *ompC* gene is indicated in gray, while the surrounding UTI89 sequence is indicated by black. (C) Schematic for strain hybridization by mass generalized allelic exchange. A library of UTI89 with random insertions of the negative selection system (thick black arcs) throughout the chromosome is subjected to the process in (A) *en masse*. Growth on restrictive conditions would result in a library of hybrids in which each locus originally carrying the negative selection system insertion has been replaced by the homologous locus from CFT073 (thick red arcs).

found that the precise recombination breakpoints were different (Figure 3B), as expected from the random shearing of the CFT073 genomic DNA. Of note, selection stringency is paramount in this application, as only 0.2% of the input DNA (the CFT073 genome is ~5 Mbp, while DNA was sheared to 10 kb) is 'on-target' for replacement of the negative selection module at the *ompC* locus. Therefore, our negative selection system enables phage-free generalized allelic exchange between different strains of *E. coli*, substituting for generalized transduction.

DISCUSSION

There is a surprising lack of stringent and general negative selection tools, even for cloning strains of *E. coli*. Indeed, several recent reports (3,4,41) have highlighted the need for better negative selection systems in *E. coli*. In general, previous systems have suffered from low stringency, a requirement for modified hosts, or a need to optimize selection conditions separately for each host strain. In the context of understanding the genetic basis of pathogenesis in pri-

mary clinical isolates, these drawbacks have made negative selection largely impractical. Particularly for loss of function studies, high frequency site-specific recombinase systems such as Cre-lox (42) and Flp-FRT (14) can be leveraged to manipulate chromosomal loci with only a small scar left behind, largely alleviating the need for negative selection. However, allelic replacement experiments in clinical isolates to date have been relatively rare compared with loss of function experiments; when done, they are limited either due to laborious construction methods or due to use of a linked positive selection marker.

General and stringent negative selection systems could remove these limitations on direct genetic studies of unmodified clinical pathogenic isolates. Recent reports of new negative selection systems have continued to focus on cloning strains of *E. coli* (3,4,41). These reported systems may be applicable to clinical strains, though the requirement for or complexity of strain-specific optimization is still unclear. Our system, due to its modular design, can accommodate essentially any toxin gene, whether from a toxin-antitoxin system, phage system, secretion system or potentially other as yet unknown genetic modules. We have demonstrated that the majority of tested toxins can be converted without optimization into effective negative selection systems that vary little in their efficiency and stringency. Furthermore, we have shown that no optimization is needed to use these systems in multiple clinical strains of *E. coli*, including a previously uncharacterized clinical isolate and a well-studied lab strain of *E. coli*, as well as in *S. enterica* Serovar Typhimurium. Therefore, we anticipate that little or no optimization will be needed for the vast majority of *E. coli* strains.

The stringency and generality of our design certainly depends on the availability of tightly regulated promoters in *E. coli* and other enteric bacteria. We have shown that both the rhamnose and arabinose promoters are sufficiently well controlled in single copy to enable bacterial growth when not specifically induced and are very stringent in preventing growth when induced. Based on similar promoter dynamics when present in single copy on the chromosome, we also expect the widely used *lac* promoter to be a viable promoter for driving toxin genes (43,44). Additional promoters native to *E. coli* and other Enterobacteria that respond to pH (45), temperature (46) or oxygenation conditions (47) could also be used. This intentional modularity of transcriptional control, enabling use of different promoters, is particularly valuable for generalization to other bacterial systems.

We noted that, while four of six tested TA system genes could be immediately used in our strategy, *higB* was only viable for negative selection when present on a plasmid in multiple copies, while *yafQ* could not be used for negative selection at all. Interestingly, this implies that the induced expression level from a single copy is insufficient to prohibit growth. In all cases, the 5' untranslated sequence and ribosome binding sequence were identical among our constructs. Therefore, we suspect that the observed variation in toxicity at the same levels of promoter induction could be due to additional transcriptional (such as pause or termination sites), translational (such as rare codons or mRNA stability), or post-translational effects (such as protein folding or intrinsic toxicity thresholds) unique to individual tox-

ins. Regardless of the mechanism, however, our results imply that toxins differ in their 'effective toxicity' when we controlled copy number and promoter induction; this concept of a threshold for toxicity has been introduced previously in relation to levels of the antitoxin within a cell (48). Because the toxins themselves are known to function in a broad range of bacteria (25,26) as well as eukaryotes (yeast) (49), toxin genes that are permissive for cell growth at low transcription levels may enable the use of less optimal (less tight) inducible promoters. Furthermore, due to the large number of toxin-antitoxin systems (and other bacterial toxins from phages and secretion systems) present in public databases from genome sequencing efforts (cataloged, for example, in TADB (50), RASTA (51) and the PanDaTox (52) databases), it seems likely that effective toxicity will vary greatly, especially if codon usage or transcriptional controls within the coding sequences are not removed. Indeed, one could imagine collecting a database of toxin genes that are characterized based on the threshold transcriptional or induction level required for toxicity; this could then be queried to identify reasonable candidates for use with existing inducible promoters to generate additional negative selection systems depending on the available genetic tools in a given system.

We have performed numerous chromosomal manipulations in UTI89 and other lab and clinical *E. coli* strains using all variants of these negative selection systems. The construction of definitive strains for genetic tests in clinical isolates is no longer less convenient than in lab strains. However, we have relied on the widely used Red recombinase system from phage λ . Expression of this system in *E. coli* is known to be mutagenic, and we use it in each of two steps to manipulate the chromosome. While following published recommendations for limiting the length of recombinase expression (22), the possibility still remains that unselected mutations may be induced that could confound interpretation of downstream experiments. To definitively address this possibility, we used full genome sequencing of one of our allelic replacement strains and found no unexpected mutations of any class, including SNPs, small indels, or large scale rearrangements. Furthermore, phenotypic testing of a panel of FimH mutants, all constructed by allelic replacement, showed perfect concordance with previously published results (in which the strains carried a linked positive selection marker).

Our system exceeds the reported stringency of selection (measured using stringency frequencies) of the next best reported system (*tetA-sacB* in MG1655, W3110 and/or DH10B; 6×10^{-7} (3)) by up to $60\times$ in lab strains of *E. coli* (kan- $P_{rhaB-tse2}$ in MG1655; 1.11×10^{-8} as measured by plating 10^{11} CFU); this begins to approach (within $5\times$) the selection stringency frequency of positively selectable antibiotic markers in some strains we have tested (kanamycin in *E. coli* EDL933; 2.40×10^{-9}). The stringency frequency in clinical strains (kan- $P_{rhaB-reIE}$ in UTI89; 3.31×10^{-8}) is 17-fold better than the *tetA-sacB* system as measured in lab strains, and it remains high without the need for strain- or species-specific optimization of transformation or growth conditions. We have further measured mutation rates for growth on restrictive conditions and find that these also correspond well with expected minimum theoretical rates for

inactivating mutations as calculated by previously reported genomic mutation rates in *E. coli* (34,35). We have used the more easily measured stringency frequencies to compare the different systems. This is similar to previous screens of mutation rates in *E. coli* (31–33), though perhaps not fully rigorous. We note, however, that our stringency frequencies indicate that kanamycin is more stringent than our negative selection system (namely, using *P_{rhaB-tse2}*), and we obtain the same result when making a more careful measurement of mutation rate. Furthermore, measurement of mutation frequencies is able to detect mutator strains as those with higher mutation frequencies (32,33). Thus our and previous data both argue that measurement of stringency (mutation) frequencies is reasonable for a relative comparison of mutation rates, a notion further supported by additional theoretical analysis of fluctuation tests for mutation rates (53).

Although mutation rates and frequencies are not directly comparable, we note here two further observations inspired by a discussion with one of the reviewers. First, our expected minimum mutation rate (0.9×10^{-8}) and calculated mutation rate (2.45×10^{-8}) are both similar to the mutation rates towards resistance to phage measured by Luria and Delbruck themselves (1.1×10^{-8} – 4.1×10^{-8}) (17). In looking closely at their original data, the culture sizes (0.2–10 ml with 1×10^8 – 4×10^{10} CFU in each) are also similar to ours. The ‘average’ number of resistant bacteria measured in each of their cultures, due to the jackpot effect, is higher than the number of resistant bacteria measured in the majority of their cultures (50–95% of cultures in each experiment have fewer resistant bacteria than the average), which forms part of the motivation for developing the fluctuation test instead of relying on mutation frequencies. Furthermore, using a ‘simple expectation’ calculation of (# of bacteria) \times (mutation rate) as an arbitrary cutoff, Luria and Delbruck observe an average frequency of mutant CFU above this number, while we see frequencies of background growth near or below this number. This again highlights the distinction between mutation frequencies and rates, arguing that direct measurements of rates are more reliable. This observation may also indicate that we have not sampled enough independent cultures to capture the high variance in the number of mutations present per culture; unfortunately, due to the way we performed the fluctuation tests, we did not collect data on the number of resistant CFU per culture.

However, as a second point, we do note another interesting set of data that provides an informative comparison with our work. Recently, Luria–Delbruck fluctuation tests were performed on *E. coli* for resistance to rifampicin, in the context of examining the relationship of cell density with mutation rate (54); the measured range of mutation rates to rifampicin resistance spanned $\sim 2 \times 10^{-9}$ to 1×10^{-8} . Several earlier studies have examined mutation to rifampicin resistance in *E. coli* using simple frequency tests; two large studies of *E. coli* isolates obtained mutation frequencies of 2.5×10^{-9} to 1×10^{-8} (31) and 5×10^{-9} to 5×10^{-8} (32), and a third study of *E. coli* and *Salmonella* obtained $< 1 \times 10^{-8}$ to 5×10^{-8} (33). The mutation rate and frequency studies differ in the strains and media used, but overall the results match our data in that the (more simply measured) mutation frequencies are quantitatively sim-

ilar to the mutation rates measured using the fluctuation tests. Furthermore, we also note a quantitative similarity between the stringency frequency (2.30×10^{-10}) and mutation rate (1.11×10^{-10}) that we measured for kanamycin resistance. Again, we note only that the quantitative similarity between stringency (mutation) frequencies and mutation rates that we observe in our data is mirrored in published data and present in multiple experiments with different selection cassettes; we therefore suggest that this likely indicates our measurements are accurate. However, the conceptual separation of rates and frequencies must still be maintained. For most other negative selection cassettes, however, only stringency frequencies are available, and the following discussion of stringencies should be considered in the context of the foregoing discussion.

Unlike several commonly used systems including *sacB* or *tetA*, our negative selection cassettes appear to come very close to the theoretical maximal stringency (based on mutation rate) for genes of their size, which may account for their superiority over these systems. Although the theoretical maximal stringency for *sacB* is between 1.67×10^{-8} and 4.25×10^{-7} , the reported stringencies (usually using stringency frequency) measured for this system are much lower (*E. coli* HS996; 2.3×10^{-5} (20)). Similarly, although the theoretical maximal stringency for *tetA* is between 1.33×10^{-8} and 3.40×10^{-7} , the reported stringencies in different host strains and at different nickel concentrations vary between 10^{-4} and 10^{-6} (1). These systems fall short of expectations in their behavior by 100- to 1000-fold (assuming the correspondence between mutation frequencies and rates should also hold for these systems). Why is our system more robust to selection conditions and changing host strains than previous systems? We speculate that several aspects of our design contribute to higher stringency. First, we are using toxins that have presumably evolved for the purpose of efficiently stopping cell growth; in contrast, tetracycline resistance and the *sacB* gene were evolved for other purposes, and their use for negative selection is based on leveraging other properties they confer on cells. Second, we have included several features that enable us to have ‘tight’ control over toxin induction, including transcriptional terminators and the well characterized rhamnose and arabinose promoters to reduce basal uninduced expression; and the use of the native *rhaB* RBS sequence and minimal media with a single carbon source to ensure good ‘induced’ expression. The combination of an inducible toxin, therefore, for arresting cell growth becomes reliably stringent in different strains and species so long as we can maintain low or no basal expression. A further potential improvement may also be gained by using genomic templates for amplification of the positive–negative selection cassette, thereby eliminating the possibility of the template plasmids conferring resistance during recombination, despite their reliance on conditional replication origins. In contrast, the variable selection mediated by *tetA* or *sacB* may be due to their overall expression levels in and membrane properties of different strains. We finally note that generality and high stringency is common for negative selection systems that require special host genotypes (such as *tolC* (41), *thyA* (6) and *neo-rpsL* (7)); we have now provided this benefit to unmodified lab and clin-

ical isolates of *E. coli* and *Salmonella* and potentially other Enterobacteriaceae.

In practice, stringency does seem to vary for our negative selection system in different experimental settings, but no more than it does for positive selection markers. For example, in the reimplementation of P1 transduction using negative selection with MG1655 carrying the *tse2* toxin, we saw a higher number of background colonies (200 out of 10^9 CFU, giving a background frequency of 5×10^{-6}). However, we have observed this result with positive selection markers as well. Kanamycin stringency frequency is 2.3×10^{-10} , but use of the Datsenko and Wanner protocol with Red recombinase leads to background colonies (not the desired recombinants and not able to regrow on kanamycin upon passage) growing on kanamycin plates at a frequency higher than this. This discrepancy in kanamycin stringency frequency may be due to the mutagenic effect of Red recombinase expression, although we and others (22) see no evidence for higher mutation when Red recombinase expression is limited in time. Similarly, we have no good explanation for the higher background rate when performing P1 transduction compared to simple titering on rhamnose plates. However, the variation in stringency seems to apply to both negative and positive selection markers.

The generally high stringency of our negative selection system is sufficient to enable previously inaccessible large scale genomic manipulations. We have carried out an initial proof of concept experiment which shows that the system is powerful enough to select for clones of UTI89 that have been transformed with the correct 10 kb fragment of the CFT073 genome (0.2% of the total genome) and then recombined it into the correct *ompC* locus. Potentially, additional development of this technology to reduce background colonies and improve genomic DNA transfer length and efficiency, such as by using Hfr strains for conjugation or transducing lysates, would enable the transfer of entire swaths of genomic DNA from unrelated strains, subject to flanking homology. Combined with transposon engineering, a pool of strains containing randomly inserted negative selectable markers could then be used to create libraries of wholesale allelic exchange mutants in a procedure we term 'mass allelic exchange' (Figure 3C). This would finally enable a practical genetic technology for assigning phenotypic consequences to intra-species polymorphisms as well as close inter-species differences. This enabling of sexual genetics in non-naturally competent bacteria could be especially powerful for dissecting the differences between strains of *E. coli* (or other bacteria) that cause different diseases in humans, a longstanding and perplexing problem.

One drawback of *relE* and several of the other toxins used is that they tend to be bacteriostatic instead of bacteriocidal. This necessitates an extra step of colony purification during standard cloning procedures and must be accounted for in large-scale allelic exchange experiments. However, switching to a bacteriocidal toxin should alleviate this drawback and increase the efficiency and convenience of this system, while possibly simultaneously further broadening the host range.

In summary, we have designed a modular and general negative selection system that is both extremely powerful and highly versatile, potentially enabling generalization to

other bacteria. The high selection stringency enables novel and improved applications including generalized allelic exchange between arbitrary *E. coli* strains (or even between different species). This negative selection module should therefore be particularly useful in the dissection of bacterial virulence as well as regulation, physiology, speciation, and evolution. With further optimization, libraries of strain hybrids could be created that would surmount drawbacks inherent to positive selection based hybrids (55), enabling sexual genetic strategies to systematically evaluate the thousands of DNA polymorphisms that typically separate unrelated clinical isolates of *E. coli* and other bacteria that are not naturally competent.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Linda Kenney, Bill Burkholder, Kimberly Kline and members of the Chen lab for helpful discussions and suggestions on the method and the manuscript. We are also indebted to an anonymous reviewer for the insights and discussions regarding the measurement and interpretation of selection stringency. *Salmonella enterica* Serovar Typhimurium 14028S was kindly provided by Linda Kenney. P1 vir phage lysate was kindly provided by Shu Sin Chng.

FUNDING

National Research Foundation, Prime Minister's Office, Singapore under its NRF Research Fellowship Scheme [NRF-RF2010-10 to S.L.C.]; Genome Institute of Singapore (GIS)/Agency for Science, Technology and Research (A*STAR). Funding for open access charge: National Research Foundation, Prime Minister's Office, Singapore under its NRF Research Fellowship Scheme [NRF-RF2010-10 to S.L.C.].

Conflict of interest statement. None declared.

REFERENCES

- Podolsky, T., Fong, S.T. and Lee, B.T. (1996) Direct selection of tetracycline-sensitive *Escherichia coli* cells using nickel salts. *Plasmid*, **36**, 112–115.
- Gay, P., Le Coq, D., Steinmetz, M., Ferrari, E. and Hoch, J.A. (1983) Cloning structural gene *sacB*, which codes for exoenzyme levansucrase of *Bacillus subtilis*: expression of the gene in *Escherichia coli*. *J. Bacteriol.*, **153**, 1424–1431.
- Li, X.T., Thomason, L.C., Sawitzke, J.A., Costantino, N. and Court, D.L. (2013) Positive and negative selection using the tetA-*sacB* cassette: recombineering and P1 transduction in *Escherichia coli*. *Nucleic Acids Res.*, **41**, e204.
- Wang, H., Bian, X., Xia, L., Ding, X., Muller, R., Zhang, Y., Fu, J. and Stewart, A.F. (2014) Improved seamless mutagenesis by recombineering using *ccdB* for counterselection. *Nucleic Acids Res.*, **42**, e37.
- DeVito, J.A. (2008) Recombineering with *tolC* as a selectable/counter-selectable marker: remodeling the rRNA operons of *Escherichia coli*. *Nucleic Acids Res.*, **36**, e4.
- Wong, Q.N., Ng, V.C., Lin, M.C., Kung, H.F., Chan, D. and Huang, J.D. (2005) Efficient and seamless DNA recombineering using a thymidylate synthase A selection system in *Escherichia coli*. *Nucleic Acids Res.*, **33**, e59.

7. Heermann,R., Zeppenfeld,T. and Jung,K. (2008) Simple generation of site-directed point mutations in the Escherichia coli chromosome using Red(R)/ET(R) Recombination. *Microb. Cell Factor.*, **7**, 14.
8. Tashiro,Y., Fukutomi,H., Terakubo,K., Saito,K. and Umeno,D. (2011) A nucleoside kinase as a dual selector for genetic switches and circuits. *Nucleic Acids Res.*, **39**, e12.
9. Alteri,C.J. and Mobley,H.L. (2012) Escherichia coli physiology and metabolism dictates adaptation to diverse host microenvironments. *Curr. Opin. Microb.*, **15**, 3–9.
10. Battaglioli,E.J., Baisa,G.A., Weeks,A.E., Schroll,R.A., Hryckowian,A.J. and Welch,R.A. (2011) Isolation of generalized transducing bacteriophages for uropathogenic strains of Escherichia coli. *Appl. Environ. Microbiol.*, **77**, 6630–6635.
11. Hood,R.D., Singh,P., Hsu,F., Guvener,T., Carl,M.A., Trinidad,R.R., Silverman,J.M., Ohlson,B.B., Hicks,K.G., Plemel,R.L., Li,M. *et al.* (2010) A type VI secretion system of Pseudomonas aeruginosa targets a toxin to bacteria. *Cell Host Microbe*, **7**, 25–37.
12. Grundling,A., Manson,M.D. and Young,R. (2001) Holins kill without warning. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 9348–9352.
13. Haldimann,A. and Wanner,B.L. (2001) Conditional-replication, integration, excision, and retrieval plasmid-host systems for gene structure-function studies of bacteria. *J. Bacteriol.*, **183**, 6384–6393.
14. Datsenko,K.A. and Wanner,B.L. (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 6640–6645.
15. Chen,S.L., Hung,C.S., Pinkner,J.S., Walker,J.N., Cusumano,C.K., Li,Z., Bouckaert,J., Gordon,J.I. and Hultgren,S.J. (2009) Positive selection identifies an in vivo role for FimH during urinary tract infection in addition to mannose binding. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 22439–22444.
16. Iqbal,Z., Caccamo,M., Turner,I., Flicek,P. and McVean,G. (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.*, **44**, 226–232.
17. Luria,S.L.a.D., M. (1943) Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, **28**, 491–511.
18. Thomason,L.C., C.,N. and Court,D.L. (2007) *Current Protocols in Molecular Biology*. John Wiley & Sons, Inc, NY.
19. Yamaguchi,Y. and Inouye,M. (2011) Regulation of growth and death in Escherichia coli by toxin-antitoxin systems. *Nat. Rev. Microbiol.*, **9**, 779–790.
20. Stavropoulos,T.A. and Strathdee,C.A. (2001) Synergy between tetA and rpsL provides high-stringency positive and negative selection in bacterial artificial chromosome vectors. *Genomics*, **72**, 99–104.
21. Bochner,B.R., Huang,H.C., Schieven,G.L. and Ames,B.N. (1980) Positive selection for loss of tetracycline resistance. *J. Bacteriol.*, **143**, 926–933.
22. Murphy,K.C. and Campellone,K.G. (2003) Lambda Red-mediated recombinogenic engineering of enterohemorrhagic and enteropathogenic E. coli. *BMC Mol. Biol.*, **4**, 11.
23. Liang,R. and Liu,J. (2010) Scarless and sequential gene modification in Pseudomonas using PCR product flanked by short homology regions. *BMC Microbiol.*, **10**, 209.
24. Zhang,Y., Buchholz,F., Muyrers,J.P. and Stewart,A.F. (1998) A new logic for DNA engineering using recombination in Escherichia coli. *Nat. Genet.*, **20**, 123–128.
25. Zhang,X.Z., Yan,X., Cui,Z.L., Hong,Q. and Li,S.P. (2006) mazF, a novel counter-selectable marker for unmarked chromosomal manipulation in Bacillus subtilis. *Nucleic Acids Res.*, **34**, e71.
26. Al-Hinai,M.A., Fast,A.G. and Papoutsakis,E.T. (2012) Novel system for efficient isolation of Clostridium double-crossover allelic exchange mutants enabling markerless chromosomal gene deletions and DNA integration. *Appl. Environ. Microbiol.*, **78**, 8112–8121.
27. Welch,R.A., Burland,V., Plunkett,G. 3rd, Redford,P., Roesch,P., Rasko,D., Buckles,E.L., Liou,S.R., Boutin,A., Hackett,J. *et al.* (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 17020–17024.
28. Riley,L.W., Remis,R.S., Helgerson,S.D., McGee,H.B., Wells,J.G., Davis,B.R., Hebert,R.J., Olcott,E.S., Johnson,L.M., Hargrett,N.T. *et al.* (1983) Hemorrhagic colitis associated with a rare Escherichia coli serotype. *N. Engl. J. Med.*, **308**, 681–685.
29. Chen,S.L., Wu,M., Henderson,J.P., Hooton,T.M., Hibbing,M.E., Hultgren,S.J. and Gordon,J.I. (2013) Genomic diversity and fitness of E. coli strains recovered from the intestinal and urinary tracts of women with recurrent urinary tract infection. *Sci. Transl. Med.*, **5**, 13.
30. Jarvik,T., Smillie,C., Groisman,E.A. and Ochman,H. (2010) Short-term signatures of evolutionary change in the Salmonella enterica serovar typhimurium 14028 genome. *J. Bacteriol.*, **192**, 560–567.
31. Denamur,E., Bonacorsi,S., Giraud,A., Duriez,P., Hilali,F., Amarin,C., Bingen,E., Andremont,A., Picard,B., Taddei,F. *et al.* (2002) High frequency of mutator strains among human uropathogenic Escherichia coli isolates. *J. Bacteriol.*, **184**, 605–609.
32. Matic,I., Radman,M., Taddei,F., Picard,B., Doit,C., Bingen,E., Denamur,E. and Elion,J. (1997) Highly variable mutation rates in commensal and pathogenic Escherichia coli. *Science*, **277**, 1833–1834.
33. LeClerc,J.E., Li,B., Payne,W.L. and Cebula,T.A. (1996) High mutation frequencies among Escherichia coli and Salmonella pathogens. *Science*, **274**, 1208–1211.
34. Drake,J.W. (1991) A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl. Acad. Sci. U.S.A.*, **88**, 7160–7164.
35. Lee,H., Popodi,E., Tang,H. and Foster,P.L. (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, doi:10.1073/pnas.1210309109.
36. Kryukov,G.V., Pennacchio,L.A. and Sunyaev,S.R. (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.*, **80**, 727–739.
37. Schaaper,R.M. and Dunn,R.L. (1991) Spontaneous mutation in the Escherichia coli lacI gene. *Genetics*, **129**, 317–326.
38. Uematsu,N., Matsuoka,C., Agemizu,Y., Nagoshi,E. and Yamamoto,K. (1999) Asymmetric crossing over in the spontaneous formation of large deletions in the tonB-trp region of the Escherichia coli K-12 chromosome. *Mol. Gen. Genet.*, **261**, 523–529.
39. Albertini,A.M., Hofer,M., Calos,M.P. and Miller,J.H. (1982) On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. *Cell*, **29**, 319–328.
40. Schaaper,R.M., Danforth,B.N. and Glickman,B.W. (1986) Mechanisms of spontaneous mutagenesis: an analysis of the spectrum of spontaneous mutation in the Escherichia coli lacI gene. *J. Mol. Biol.*, **189**, 273–284.
41. Gregg,C.J., Lajoie,M.J., Napolitano,M.G., Mosberg,J.A., Goodman,D.B., Aach,J., Isaacs,F.J. and Church,G.M. (2014) Rational optimization of tolC as a powerful dual selectable marker for genome engineering. *Nucleic Acids Res.*, **42**, 4779–4790.
42. Tuntufye,H.N. and Goddeeris,B.M. (2011) Use of lambda Red-mediated recombineering and Cre/lox for generation of markerless chromosomal deletions in avian pathogenic Escherichia coli. *FEMS Microbiol. Lett.*, **325**, 140–147.
43. Novick,A. and Weiner,M. (1957) Enzyme induction as an all-or-none phenomenon. *Proc. Natl. Acad. Sci. U.S.A.*, **43**, 553–566.
44. Siegel,D. and Hu,J.C. (1997) Gene expression from plasmids containing the araBAD promoter at subsaturating inducer concentrations represents mixed populations. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 8168–8172.
45. Chou,C.H., Aristidpu,A.A., Meng,S.Y., Bennett,G.N. and San,K.Y. (1995) Characterization of a pH-inducible promoter system for high-level expression of recombinant proteins in Escherichia coli. *Biotechnol. Bioeng.*, **47**, 186–192.
46. Valdez-Cruz,N.A., Caspeta,L., Perez,N.O., Ramirez,O.T. and Trujillo-Roldan,M.A. (2010) Production of recombinant proteins in E. coli by the heat inducible expression system based on the phage lambda pL and/or pR promoters. *Microb. Cell Factor.*, **9**, 18.
47. Khosla,C., Curtis,J.E., Bydalek,P., Swartz,J.R. and Bailey,J.E. (1990) Expression of recombinant proteins in Escherichia coli using an oxygen-responsive promoter. *Nat. Biotechnol.*, **8**, 554–558.
48. Rotem,E., Loinger,A., Ronin,I., Levin-Reisman,I., Gabay,C., Shores,N., Biham,O. and Balaban,N.Q. (2010) Regulation of phenotypic variability by a threshold-based mechanism underlies bacterial persistence. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 12541–12546.
49. Yang,J., Jiang,W. and Yang,S. (2009) mazF as a counter-selectable marker for unmarked genetic modification of Pichia pastoris. *FEMS Yeast Res.*, **9**, 600–609.
50. Shao,Y., Harrison,E.M., Bi,D., Tai,C., He,X., Ou,H.Y., Rajakumar,K. and Deng,Z. (2011) TADB: a web-based resource for Type 2 toxin-antitoxin loci in bacteria and archaea. *Nucleic Acids Res.*, **39**, D606–D611.

51. Sevin, E.W. and Barloy-Hubler, F. (2007) RASTA-Bacteria: a web-based tool for identifying toxin-antitoxin loci in prokaryotes. *Genome Biol.*, **8**, R155.
52. Amitai, G. and Sorek, R. (2012) PanDaTox: a tool for accelerated metabolic engineering. *Bioengineered*, **3**, 218–221.
53. Stewart, F.M. (1994) Fluctuation tests: how reliable are the estimates of mutation rates? *Genetics*, **137**, 1139–1146.
54. Krasovec, R., Belavkin, R.V., Aston, J.A., Channon, A., Aston, E., Rash, B.M., Kadirvel, M., Forbes, S. and Knight, C.G. (2014) Mutation rate plasticity in rifampicin resistance depends on *Escherichia coli* cell-cell interactions. *Nat. Commun.*, **5**, 3742.
55. Freddolino, P.L., Goodarzi, H. and Tavazoie, S. (2014) Revealing the genetic basis of natural bacterial phenotypic divergence. *J. Bacteriol.*, **196**, 825–839.

Complete Genome Sequence of Uropathogenic *Escherichia coli* Strain CI5

Kurosh S. Mehershahi,^a Soman N. Abraham,^{c,d}  Swaine L. Chen^{a,b}

National University of Singapore, Singapore^a; Genome Institute of Singapore, Singapore^b; Duke–National University of Singapore Graduate Medical School, Singapore^c; Duke University Medical Center, Durham, North Carolina, USA^d

***Escherichia coli* represents the primary etiological agent responsible for urinary tract infections, one of the most common infections in humans. We report here the complete genome sequence of uropathogenic *Escherichia coli* strain CI5, a clinical pyelonephritis isolate used for studying pathogenesis.**

Received 24 April 2015 Accepted 29 April 2015 Published 28 May 2015

Citation Mehershahi KS, Abraham SN, Chen SL. 2015. Complete genome sequence of uropathogenic *Escherichia coli* strain CI5. *Genome Announc* 3(3):e00558-15. doi:10.1128/genomeA.00558-15.

Copyright © 2015 Mehershahi et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 3.0 Unported license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Swaine L. Chen, slchen@gis.a-star.edu.sg.

Urinary tract infections (UTIs) are one of the most frequently encountered bacterial infections, with uropathogenic *Escherichia coli* (UPEC) responsible for more than 80% of community acquired infections (1). Although often characterized as self-limiting and amenable to antibiotic therapy, UTIs often recur, causing significant morbidity to individual patients (2). Recurrent UTIs further lead to potential public health concerns due to high antibiotic usage (3, 4). Studies of UPEC pathogenesis have revealed that intracellular infection of bladder epithelial cells is a key feature leading to bacterial survival, antibiotic resistance, and recurrent UTI (5–20). UPEC strain CI5 is a clinical pyelonephritis isolate (21) that has been used in many of these studies using both *in vitro* cell culture models and *in vivo* murine infection models (22, 23). These have examined the role of Toll-like receptor 4 (TLR4) (13), cyclic AMP (cAMP), and Ca²⁺ signaling during UPEC invasion into bladder epithelial cells and the subsequent epithelial cell response (12, 16, 24). Additional studies have used CI5 to understand kidney infection and renal nephropathy during *in vivo* infection in mice (25). The genome sequence of CI5 will thus serve as a useful resource for future studies into the infection cycle of this important human pathogen.

CI5 genomic DNA was sheared to a size of approximately 10 kbp using a g-Tube (Covaris). An SMRTbell library was prepared according to the manufacturer's instructions, loaded with a Mag-Bead bound library protocol, and sequenced using the P4-C2 chemistry on the PacBio RS II instrument (Pacific Biosciences) with a 180-min movie time. *De novo* assembly was performed with the Hierarchical Genome Assembly Process (HGAP3) in the SMRT Analysis suite version 2.3 using default parameters (26). In total, there were 249,158 reads and 822,531,331 nucleotides that passed filtering, representing an approximate coverage of 80× (based on the final assembly) and a preassembly mean read length of 8,815 bp.

UPEC CI5 harbors a single chromosome of 4,885,378 bp with a G+C content of 50.8% and a previously unknown plasmid (pCI5) of 207,265 bp with a G+C content of 47.3%. Annotations of the CI5 genome and plasmid were performed using the NCBI

Prokaryotic Genome Annotation Pipeline (PGAAP) (27). The CI5 chromosome and plasmid together contain 4,879 protein coding sequences, as well as 22 rRNA and 88 tRNA genes. The finished genome sequence of UPEC CI5 and its newly discovered plasmid pCI5 will aid in precise genetic manipulation and thereby further improve the study of UPEC virulence.

Nucleotide accession numbers. The complete sequences of the uropathogenic *E. coli* CI5 chromosome and plasmid have been submitted to GenBank under the accession numbers [CP011018](https://ncbi.nlm.nih.gov/nucl/CP011018) and [CP011019](https://ncbi.nlm.nih.gov/nucl/CP011019), respectively.

ACKNOWLEDGMENTS


This work was supported by the National Research Foundation, Prime Minister's Office, Singapore, grant no. NRF-RF2010-10, and the National Medical Research Council, Ministry of Health, Singapore, grant no. NMRC/CIRG/1357/2013.

REFERENCES

1. Foxman B. 2002. Epidemiology of urinary tract infections: incidence, morbidity, and economic costs. *Am J Med* 113(Suppl 1A):5S–13S. [http://dx.doi.org/10.1016/S0002-9343\(02\)01054-9](https://doi.org/10.1016/S0002-9343(02)01054-9).
2. Barber AE, Norton JP, Spivak AM, Mulvey MA. 2013. Urinary tract infections: current and emerging management strategies. *Clin Infect Dis* 57:719–724. [http://dx.doi.org/10.1093/cid/cit284](https://doi.org/10.1093/cid/cit284).
3. Foxman B. 2010. The epidemiology of urinary tract infection. *Nat Rev Urol* 7:653–660. [http://dx.doi.org/10.1038/nrurol.2010.190](https://doi.org/10.1038/nrurol.2010.190).
4. Kodner CM, Thomas Gupton EK. 2010. Recurrent urinary tract infections in women: diagnosis and management. *Am Fam Physician* 82:638–643.
5. Mulvey MA, Lopez-Boado YS, Wilson CL, Roth R, Parks WC, Heuser J, Hultgren SJ. 1998. Induction and evasion of host defenses by type 1-piliated uropathogenic *Escherichia coli*. *Science* 282:1494–1497. [http://dx.doi.org/10.1126/science.282.5393.1494](https://doi.org/10.1126/science.282.5393.1494).
6. Martinez JJ, Mulvey MA, Schilling JD, Pinkner JS, Hultgren SJ. 2000. Type 1 pilus-mediated bacterial invasion of bladder epithelial cells. *EMBO J* 19:2803–2812. [http://dx.doi.org/10.1093/emboj/19.12.2803](https://doi.org/10.1093/emboj/19.12.2803).
7. Mulvey MA, Schilling JD, Hultgren SJ. 2001. Establishment of a persistent *Escherichia coli* reservoir during the acute phase of a bladder infection. *Infect Immun* 69:4572–4579. [http://dx.doi.org/10.1128/IAI.69.7.4572-4579.2001](https://doi.org/10.1128/IAI.69.7.4572-4579.2001).
8. Schilling JD, Lorenz RG, Hultgren SJ. 2002. Effect of trimethoprim-

- sulfamethoxazole on recurrent bacteriuria and bacterial persistence in mice infected with uropathogenic *Escherichia coli*. *Infect Immun* 70:7042–7049. <http://dx.doi.org/10.1128/IAI.70.12.7042-7049.2002>.
9. Anderson GG, Palermo JJ, Schilling JD, Roth R, Heuser J, Hultgren SJ. 2003. Intracellular bacterial biofilm-like pods in urinary tract infections. *Science* 301:105–107. <http://dx.doi.org/10.1126/science.1084550>.
 10. Justice SS, Hung C, Theriot JA, Fletcher DA, Anderson GG, Footer MJ, Hultgren SJ. 2004. Differentiation and developmental pathways of uropathogenic *Escherichia coli* in urinary tract pathogenesis. *Proc Natl Acad Sci U S A* 101:1333–1338. <http://dx.doi.org/10.1073/pnas.0308125100>.
 11. Mysorekar IU, Hultgren SJ. 2006. Mechanisms of uropathogenic *Escherichia coli* persistence and eradication from the urinary tract. *Proc Natl Acad Sci U S A* 103:14170–14175. <http://dx.doi.org/10.1073/pnas.0602136103>.
 12. Bishop BL, Duncan MJ, Song J, Li G, Zaas D, Abraham SN. 2007. Cyclic AMP-regulated exocytosis of *Escherichia coli* from infected bladder epithelial cells. *Nat Med* 13:625–630. <http://dx.doi.org/10.1038/nm1572>.
 13. Song J, Bishop BL, Li G, Duncan MJ, Abraham SN. 2007. TLR4-initiated and cAMP-mediated abrogation of bacterial invasion of the bladder. *Cell Host Microbe* 1:287–298. <http://dx.doi.org/10.1016/j.chom.2007.05.007>.
 14. Eto DS, Gordon HB, Dhakal BK, Jones TA, Mulvey MA. 2008. Clathrin, AP-2, and the NPXY-binding subset of alternate endocytic adaptors facilitate FimH-mediated bacterial invasion of host cells. *Cell Microbiol* 10:2553–2567. <http://dx.doi.org/10.1111/j.1462-5822.2008.01229.x>.
 15. Dhakal BK, Mulvey MA. 2009. Uropathogenic *Escherichia coli* invades host cells via an HDAC6-modulated microtubule-dependent pathway. *J Biol Chem* 284:446–454. <http://dx.doi.org/10.1074/jbc.M805010200>.
 16. Song J, Bishop BL, Li G, Grady R, Stapleton A, Abraham SN. 2009. TLR4-mediated expulsion of bacteria from infected bladder epithelial cells. *Proc Natl Acad Sci U S A* 106:14966–14971. <http://dx.doi.org/10.1073/pnas.0900527106>.
 17. Blango MG, Mulvey MA. 2010. Persistence of uropathogenic *Escherichia coli* in the face of multiple antibiotics. *Antimicrob Agents Chemother* 54:1855–1863. <http://dx.doi.org/10.1128/AAC.00014-10>.
 18. Schwartz DJ, Chen SL, Hultgren SJ, Seed PC. 2011. Population dynamics and niche distribution of uropathogenic *Escherichia coli* during acute and chronic urinary tract infection. *Infect Immun* 79:4250–4259. <http://dx.doi.org/10.1128/IAI.05339-11>.
 19. Totsika M, Kostakioti M, Hannan TJ, Upton M, Beatson SA, Janetka JW, Hultgren SJ, Schembri MA. 2013. A FimH inhibitor prevents acute bladder infection and treats chronic cystitis caused by multidrug-resistant uropathogenic *Escherichia coli* ST131. *J Infect Dis* 208:921–928. <http://dx.doi.org/10.1093/infdis/jit245>.
 20. Blango MG, Ott EM, Erman A, Veranic P, Mulvey MA. 2014. Forced resurgence and targeting of intracellular uropathogenic *Escherichia coli* reservoirs. *PLoS One* 9:e93327. <http://dx.doi.org/10.1371/journal.pone.0093327>.
 21. Abraham SN, Babu JP, Giampapa CS, Hasty DL, Simpson WA, Beachey EH. 1985. Protection against *Escherichia coli*-induced urinary tract infections with hybridoma antibodies directed against type 1 fimbriae or complementary D-mannose receptors. *Infect Immun* 48:625–628.
 22. Hagberg L, Engberg I, Freter R, Lam J, Olling S, Svanborg Edén C. 1983. Ascending, unobstructed urinary tract infection in mice caused by pyelonephritogenic *Escherichia coli* of human origin. *Infect Immun* 40:273–283.
 23. Hung CS, Dodson KW, Hultgren SJ. 2009. A murine model of urinary tract infection. *Nat Protoc* 4:1230–1243. <http://dx.doi.org/10.1038/nprot.2009.116>.
 24. Song J, Duncan MJ, Li G, Chan C, Grady R, Stapleton A, Abraham SN. 2007. A novel TLR4-mediated signaling pathway leading to IL-6 responses in human bladder epithelial cells. *PLoS Pathog* 3:e60. <http://dx.doi.org/10.1371/journal.ppat.0030060>.
 25. Bowen SE, Watt CL, Murawski IJ, Gupta IR, Abraham SN. 2013. Interplay between vesicoureteric reflux and kidney infection in the development of reflux nephropathy in mice. *Dis Model Mech* 6:934–941. <http://dx.doi.org/10.1242/dmm.011650>.
 26. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 10:563–569. <http://dx.doi.org/10.1038/nmeth.2474>.
 27. Angiuoli SV, Gussman A, Klimke W, Cochrane G, Field D, Garrity G, Kodira CD, Kyrpides N, Madupu R, Markowitz V, Tatusova T, Thomson N, White O. 2008. Toward an online repository of standard operating procedures (SOPs) for (meta)genomic annotation. *Omics* 12:137–141. <http://dx.doi.org/10.1089/omi.2008.0017>.

Complete Genome Sequence of *Streptococcus agalactiae* Serotype III, Multilocus Sequence Type 283 Strain SG-M1

Kurosh S. Meher Shahi,^a Li Yang Hsu,^{b,c} Tse Hsien Koh,^d  Swaine L. Chen,^{a,e} on behalf of the Singapore *Streptococcus agalactiae* Working Group

National University of Singapore, Singapore^a; Saw Swee Hock School of Public Health, Singapore^b; Singapore Infectious Diseases Initiative, Singapore^c; Singapore General Hospital, Singapore^d; Genome Institute of Singapore, Singapore^e

***Streptococcus agalactiae* (group B *Streptococcus*) is a common commensal strain in the human gastrointestinal tract that can also cause invasive disease in humans and other animals. We report here the complete genome sequence of *S. agalactiae* SG-M1, a serotype III, multilocus sequence type 283 strain, isolated from a Singaporean patient suffering from meningitis.**

Received 31 August 2015 Accepted 3 September 2015 Published 22 October 2015

Citation Meher Shahi KS, Hsu LY, Koh TH, Chen SL, Singapore *Streptococcus agalactiae* Working Group. 2015. Complete genome sequence of *Streptococcus agalactiae* serotype III, multilocus sequence type 283 strain SG-M1. *Genome Announc* 3(5):e01188-15. doi:10.1128/genomeA.01188-15.

Copyright © 2015 Meher Shahi et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 3.0 Unported license](https://creativecommons.org/licenses/by/3.0/).

Address correspondence to Swaine L. Chen, slchen@gis.a-star.edu.sg.

Streptococcus agalactiae (group B *Streptococcus*, or GBS) is a Gram-positive bacterium commonly found as a commensal in the gastrointestinal and genitourinary tract of up to 40% of humans (1). *S. agalactiae* can cause pregnancy-associated infections, which can lead to invasive neonatal disease (bacteremia, pneumonia, and meningitis) after delivery; invasive disease (bacteremia, meningitis, and soft tissue infections) in all age groups; and urinary tract infections (2, 3). Furthermore, GBS is an important veterinary pathogen in several animals, including cows (mastitis) (4) and fish (meningoencephalitis) (5). GBS has been classified by serotype (Ia, Ib, II to IX) (6). Strain SG-M1 is a serotype III clinical isolate of GBS, isolated from a patient in Singapore suffering from meningitis; this isolate was collected as part of an outbreak investigation in Singapore associated with the consumption of raw fish (Barkham et al., unpublished data), in accordance with Singapore ethics regulations for exemption from IRB review. By multilocus sequence typing (<http://pubmlst.org/sagalactiae>), SG-M1 was determined to be a multilocus sequence type 283 strain (ST283); previously, ST283 strains were isolated from a meningitis patient in Hong Kong (7) and aquatic animals, the latter representing a potential zoonotic source (8).

SG-M1 was grown overnight in brain heart infusion broth at 37°C with shaking. The bacterium was collected by centrifugation and lysed in a buffer containing 20 mM Tris-Cl (pH 8.0), 2-mM sodium EDTA, 1.2% Triton X-100, and 20 mg/mL lysozyme from chicken egg white (MP Biomedicals cat. no. 100831) at 37°C for 45 min. Genomic DNA was then extracted using a Qiagen QIAamp DNA minikit and sheared to a size of approximately 10 kbp using g-Tube (Covaris). A 10-kb SMRTBell library was prepared for sequencing according to the protocols recommended by Pacific Biosciences, loaded with a MagBead-bound library protocol, and sequenced using P5-C3 chemistry on the PacBio RS II instrument (Pacific Biosciences) with a 240-min movie time on two SMRTCells. *De novo* assembly was performed with the Hierarchical Genome Assembly Process (HGAP3) in the SMRT Analysis suite version 2.3 using default parameters (9). In total, there were 62,129 reads

and 740,417,316 nucleotides that passed filtering, representing an approximate coverage of 260× (based on the final assembly) and a preassembly mean read length of 11,917 bp.

S. agalactiae SG-M1 harbors a single chromosome of 2,116,810 bp with a G+C content of 35.5%; no plasmids were found. Annotations of the Cl5 genome and plasmid were performed using the NCBI Prokaryotic Genome Annotation Pipeline (PGAAP) (10). SG-M1 contains 2,037 protein coding sequences, as well as 7 rRNA operons and 81 tRNA genes. The finished genome sequence of *S. agalactiae* SG-M1 will aid in further understanding the causative basis for invasive disease caused by GBS.

Nucleotide sequence accession number. The complete sequence of the *S. agalactiae* strain SG-M1 chromosome has been submitted to GenBank under the accession number **CP012419**.

ACKNOWLEDGMENTS

Other members of the Singapore *Streptococcus agalactiae* Working Group include Timothy Barkham, Matthew Holden, Shirin Kalimuddin, October Sessions, and Tan Thean Yen.

The sequencing was supported by the Genome Institute of Singapore (GIS)/Agency for Science, Technology and Research (A*STAR).

REFERENCES

- Edwards MS, Nizet V, Baker CJ. 2006. Group B Streptococcal infections, p 403–464. In Remington JS, Klein JO, Wilson CB, Baker CJ (ed), *Infectious diseases of the fetus and newborn infant*, 6th ed. Elsevier Saunders, Philadelphia, PA.
- Winn HN. 2007. Group B streptococcus infection in pregnancy. *Clin Perinatol* 34:387–392. <http://dx.doi.org/10.1016/j.clp.2007.03.012>.
- Sendi P, Johansson L, Norrby-Teglund A. 2008. Invasive group B streptococcal disease in non-pregnant adults: a review with emphasis on skin and soft tissue infections. *Infection* 36:100–111. <http://dx.doi.org/10.1007/s15010-007-7251-0>.
- Keefe GP. 1997. *Streptococcus agalactiae* mastitis: a review. *Can Vet J* 38:429–437.
- Evans JJ, Klesius PH, Gilbert PM, Shoemaker CA, Al Sarawi MA, Landsberg J, Duremdz R, Al Marzouk A, Al Zenki S. 2002. Characterization of β-haemolytic Group B *Streptococcus agalactiae* in cultured seabream, *Sparus auratus* L., and wild mullet, *Liza klunzingeri* (day), in Ku-

- wait. *J Fish Dis* 25:505–513. <http://dx.doi.org/10.1046/j.1365-2761.2002.00392.x>.
6. Johri AK, Paoletti LC, Glaser P, Dua M, Sharma PK, Grandi G, Rappuoli R. 2006. Group B *Streptococcus*: global incidence and vaccine development. *Nat Rev Microbiol* 4:932–942. <http://dx.doi.org/10.1038/nrmicro1552>.
 7. Ip M, Cheuk ES, Tsui MH, Kong F, Leung TN, Gilbert GL. 2006. Identification of a *Streptococcus agalactiae* serotype III subtype 4 clone in association with adult invasive disease in Hong Kong. *J Clin Microbiol* 44:4252–4254. <http://dx.doi.org/10.1128/JCM.01533-06>.
 8. Delannoy CM, Crumlish M, Fontaine MC, Pollock J, Foster G, Dagleish MP, Turnbull JF, Zadoks RN. 2013. Human *Streptococcus agalactiae* strains in aquatic mammals and fish. *BMC Microbiol* 13:41. <http://dx.doi.org/10.1186/1471-2180-13-41>.
 9. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 10:563–569. <http://dx.doi.org/10.1038/nmeth.2474>.
 10. Angiuoli SV, Gussman A, Klimke W, Cochrane G, Field D, Garrity G, Kodira CD, Kyrpides N, Madupu R, Markowitz V, Tatusova T, Thomson N, White O. 2008. Toward an online repository of Standard operating procedures (SOPs) for (meta)genomic annotation. *Omics* 12:137–141. <http://dx.doi.org/10.1089/omi.2008.0017>.

Communication

Application and Optimization of *relE* as a Negative Selection Marker for Making Definitive Genetic Constructs in Uropathogenic *Escherichia coli*

Varnica Khetrpal^{1,†}, Kurosh S. Mehershahi^{1,†}, Siyi Chen¹ and Swaine L. Chen^{1,2,*}

Received: 11 September 2015; Accepted: 13 January 2016; Published: 18 January 2016

Academic Editor: Catharina Svanborg

¹ Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, 1E Kent Ridge Road, NUHS Tower Block, Level 10, Singapore 119074, Singapore; khetrpalv@gis.a-star.edu.sg (V.K.); mehershahiks@gis.a-star.edu.sg (K.S.M.); sychen@gis.a-star.edu.sg (S.C.)

² Genome Institute of Singapore, Infectious Diseases Group, 60 Biopolis street, Genome, #02-01, Singapore 138672, Singapore

* Correspondence: slchen@gis.a-star.edu.sg; Tel.: +65-6808-8074; Fax: +65-6808-8036

† These authors contributed equally to this work.

Abstract: Studies of Uropathogenic *Escherichia coli* (UPEC) pathogenesis have relied heavily on genetic manipulation to understand virulence factors. We applied a recently reported positive-negative selection system to create a series of unmarked, scarless FimH mutants that show identical phenotypes to previously reported marked FimH mutants; these are now improved versions useful for definitive assignment of phenotypes to FimH mutations. We also increased the efficiency of this system by designing new primer sites, which should further improve the efficiency and convenience of using negative selection in UTI89, other UPEC, and other Enterobacteriaceae.

Keywords: negative selection; UTI89; FimH

1. Introduction

Urinary tract infections (UTIs) are a major problem in women worldwide, with uropathogenic *Escherichia coli* (UPEC) causing 74% of community-acquired and 65% of nosocomial infections [1]. Several UPEC strains, including *Escherichia coli* (*E. coli*) UTI89, have been shown to form intracellular bacterial communities (IBCs) in mice and in humans [2–4]; IBCs may be responsible for recurrent UTIs in mice [5]. Type 1 pili, encoded by the *fim* operon in nearly all *E. coli*, is the major virulence factor for UPEC during UTI [6,7]. FimH is the mannose-binding adhesin found at the tip of type1 pili which mediates binding to mannosylated proteins on bladder epithelial cells [8–11], facilitating colonization and invasion of cells *in vitro* [12,13] and *in vivo* [6,14] and enabling IBC formation [14–16].

Previous FimH studies have used isogenic strains carrying mutant *fimH* alleles on plasmids [17] or on the chromosome [14,16]. Chromosomal strains, in particular, have the advantages of native copy number and chromosomal context, and therefore native transcriptional regulation. However, manipulating the chromosome of clinical *E. coli* isolates such as UTI89 is more challenging than in lab-adapted strains, for which most cloning systems have been developed. This results in strains that are either marked [14,16] or carry residual cloning scars [18]. Formally, from a genetic point of view, creation of unmarked, scarless mutations in genes such as *fimH* would enable the strongest possible assignment of phenotype changes to allelic differences.

We have recently published a novel negative selection system to facilitate chromosomal manipulation in UTI89 [19]. This system consists of a toxic gene (*relE*) expressed from a tightly controlled promoter, such as the rhamnose-inducible P_{rhaB} promoter. Under normal growth conditions,

the P_{rhaB} promoter is not active, no *relE* is transcribed, and cells grow normally. However, under rhamnose induction, the production of *relE* transcript results in the production of the RelE toxin (an mRNAse), which then stops cell growth. Thus, under restrictive conditions, only cells that do not contain a functional negative selection cassette are able to grow. This negative selection cassette is combined with the kanamycin resistance gene to create a dual positive-negative selection cassette, which was used in a simple two-step procedure to create unmarked, scarless mutations in *fimH* in its native chromosomal context. We verified that our new unmarked strains show no *in vitro* or *in vivo* differences from their corresponding marked strains, validating that the previously reported phenotypes are indeed due to FimH and that the negative selection system generates no artifactual phenotypes in a large set of independent cloning steps. Furthermore, we discovered that most undesired mutants during recombineering (which are subsequently screened out during routine strain verification) using this system were due to recombination at short homology sites in the selection cassette; we eliminated this with newly designed template priming sites. The strains we have created here will be useful in the future for the definitive assignment of phenotypes to FimH mutations, while our modified negative selection protocol will increase the efficiency and convenience of creating unmarked, scarless mutations in UTI89 and other clinical isolates of *E. coli*.

2. Results and Discussion

2.1. Creation of an Isogenic Series of Definitive (Unmarked, Scarless) FimH Mutants

We used the kan- P_{rhaB} -*relE* positive-negative selection cassette [19] to recreate an unmarked, isogenic series of previously characterized FimH mutants (Figure 1a and Table S1) [14]. These newly created strains no longer carry a linked antibiotic resistance cassette, and therefore are theoretically definitive genetic constructs for isolating phenotypes due to the corresponding FimH mutation. These new unmarked strains therefore also can be utilized in identical assays as the parental UTI89, including transformation with kanamycin-resistant plasmids or the use of kanamycin in subsequent chromosomal manipulations.

2.2. Definitive FimH Constructs Have No Artifactual Phenotypes

To validate that the new unmarked *fimH* mutant strains indeed recapitulate previously reported phenotypes, we tested them for *in vitro* type 1 pilus function using guinea pig red blood cell agglutination. We saw no difference between any of the eight pairs of marked and unmarked strains (Table S2). We then tested three alleles in an *in vivo* murine model of urinary tract infection. Marked versions of the wild-type, A62S, and A27V/V163A FimH alleles previously were shown to have no defect, a 1.5-log defect, or a 4–5 log defect, respectively, at 24 hpi in the bladder [14]. We saw the same phenotypes with the new unmarked strains (Figure 1b), and also verified that the reconstructed, unmarked wild-type FimH strain was not significantly different from the unmodified, parental UTI89. We also saw no difference among matched strains in kidney infection titers (Figure 1c). Therefore, use of a combined positive-negative selection cassette in UTI89 is indeed an effective method for creating definitive genetic constructs. This system is applicable in other Enterobacterial strains, including other UPEC, *E. coli*, and Salmonella, and should also be useful for definitive genetic studies in these organisms as well.

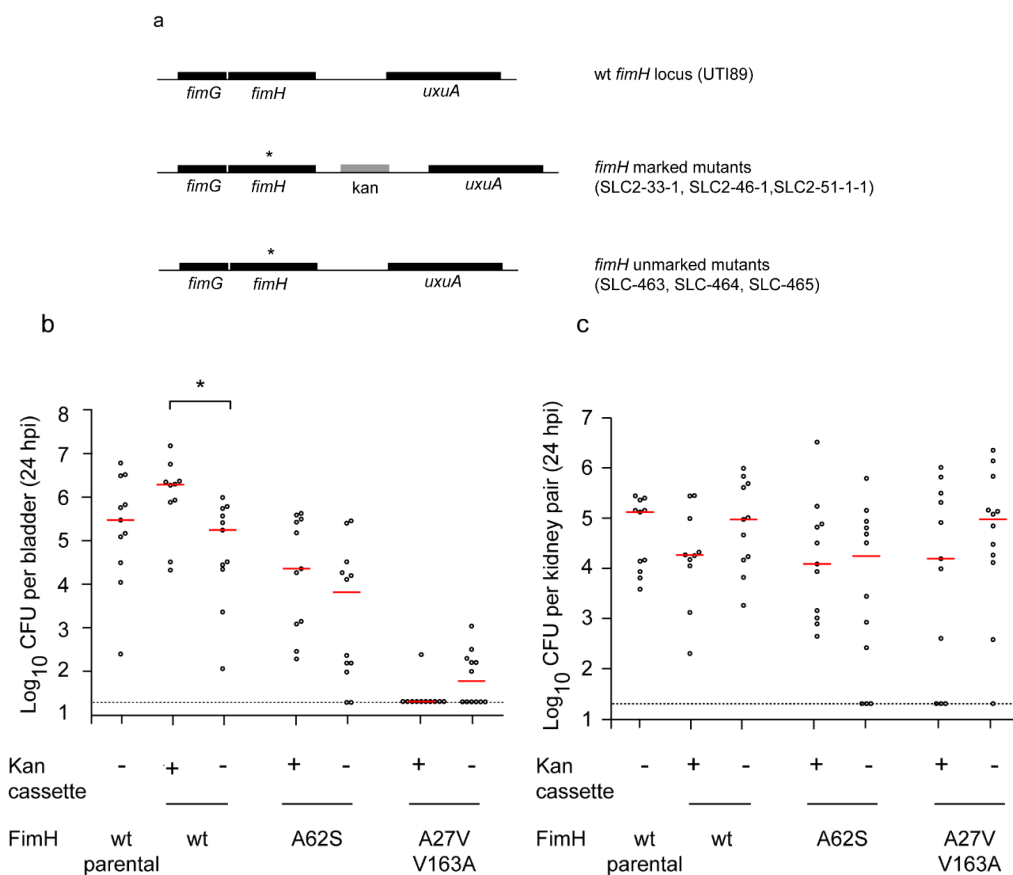


Figure 1. Creation and validation of unmarked FimH mutant strains. (a) Genetic organization of the *fimH* locus in UTI89 and in marked and unmarked FimH mutant strains. The top diagram depicts the native, wild-type locus of *fimH* in UTI89 with adjacent genes. The middle diagram depicts the genetic organization surrounding the mutated *fimH* gene (denoted by *) in marked strains reported in [14], indicating the location of the kanamycin selection marker. The bottom diagram depicts the genetic organization surrounding the mutated *fimH* gene (denoted by *) in the strains created in this study. (b,c) Bladder (b) and kidney (c) titers from *in vivo* murine infections at 24 hpi. FimH mutations are shown on x-axis. Strains carrying the marker (kanamycin cassette) are indicated. y-axis plots the logarithm (base 10) of the bacterial CFU measured in the designated organ at 24 hpi. Red horizontal bars indicate medians. Dotted line indicates limit of detection. * $p = 0.01$, two tailed Mann-Whitney test. Data from two independent experiments with five to seven mice in each experiment for each strain shown.

2.3. Redesigned Priming Sites Eliminate the Major Class of False Positive Colonies during Negative Selection-Mediated Recombineering

All *fimH* allelic replacements described above were made using the kan- P_{rhaB} -*relE* cassette amplified from template plasmid pSLC-217, which is derived from pKD4 [20]. We used priming sites 1 and 2, originally recommended for pKD4, which resulted in inclusion of Flp flippase recognition target (FRT) sites flanking the selection cassette [20]. While creating allelic variants in *fimH* and other loci (such as *ompC*) in UTI89, we observed variable numbers of false-positive background colonies during the final negative selection recombineering step (0%–83%). While these are screened out by routine verification of isolated colonies, further examination revealed that recombination at the flanking FRT sites (leading to elimination of the cassette equivalent to “Flp-ing” it out) was a common mechanism for these undesired background colonies (Figure 2a). We therefore designed alternative priming sites (primers; P1 forward, P2 reverse, Table S3) for the template plasmids that excluded the FRT sites (Figure 2b); as expected, this eliminated this class of background colonies in negative

selection-mediated allele replacements. This increased the efficiency of the negative selection step to nearly 100% (nearly all colonies growing have the intended recombination event), simplifying subsequent strain verification, although this eliminates the possibility of using FLP recombinase to eliminate the positive-negative selection step after the initial gene knockout.

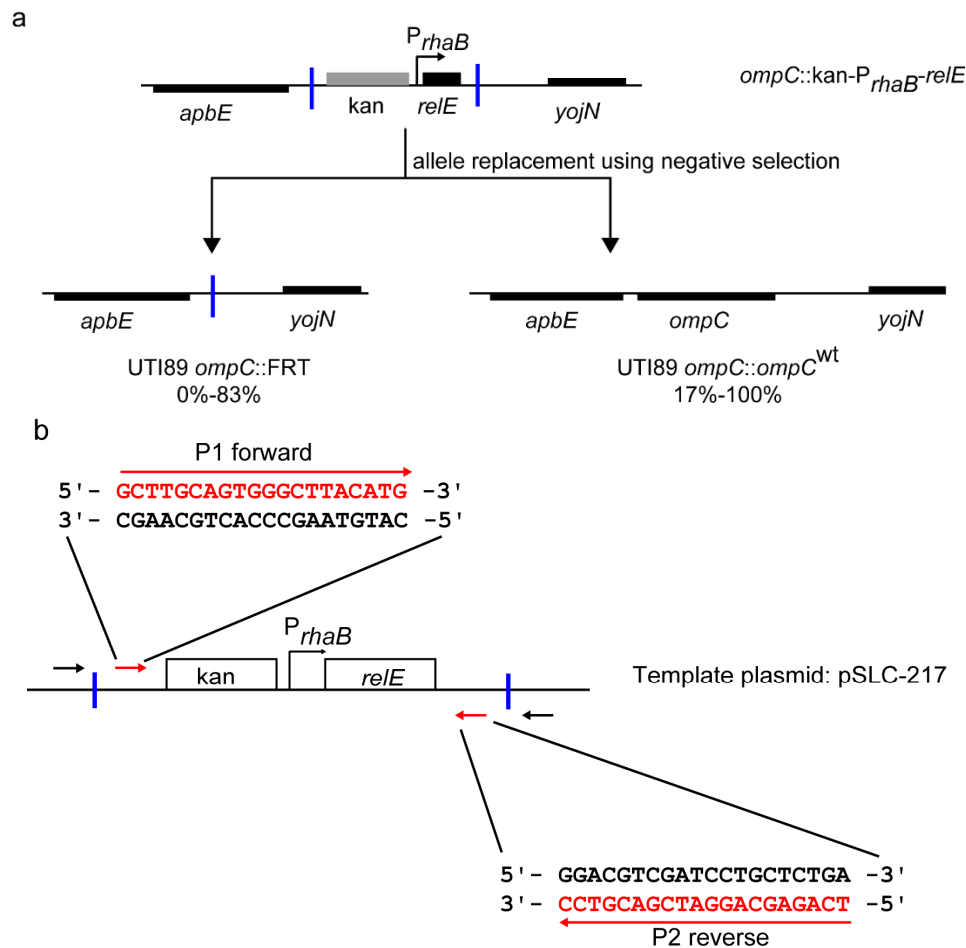


Figure 2. Elimination of undesired FRT recombinants during recombineering. (a) Genetic organization of the major recombinant products found during recombineering. The *ompC* locus and surrounding genes are shown as an example. Top, genetic organization of the initial *ompC* knockout strain (*ompC::kan-P_{rhaB}-relE*) created by recombineering using positive selection for kanamycin. Bottom left, genetic organization of the product of undesired recombination between FRT sites during subsequent allele replacement by recombineering using negative selection (*ompC::FRT*). Bottom right, the genetic organization of the desired recombination where a restored *wt ompC* allele has replaced the positive-negative selection cassette during recombineering using negative selection (*ompC::ompC^{wt}*). The percentage of colonies in each class, based on PCR screening after negative selection, is shown below each diagram (data taken from 30 recombineering steps using negative selection in 4 loci in UT189). Blue vertical bars indicate FRT sites. (b) Redesigned priming sites for template plasmid pSLC-217 (and other pKD4 derivatives). Red arrows indicate new priming sites which exclude flanking FRT sites in the resultant amplified positive-negative selection cassette. New primer sequences are indicated in red as P1 forward and P2 reverse. Black arrows indicate the original priming sites 1 and 2 (for pKD4). Blue vertical bars indicate FRT sites. New primer sequences for pKD3 derivatives are listed in Table S3.

3. Experimental Section

3.1. Bacterial Strains and Plasmids

Bacterial strains and plasmids used in this study are listed in Table S1.

3.2. Chromosomal *FimH* Mutations and Recombineering

SLC-502 (UTI89 *fimH*::kan-*P*_{rhaB}-*relE*) was used for all allelic replacements. Previously reported marked *fimH* mutants were used as templates to amplify *fimH* alleles [14] using primers UTI89+4913234 *fimH* and UTI89-4914539 *fimH* (Table S3). Recombineering was performed as previously described [19].

3.3. PCR and Sequencing

Colony PCRs were used to check for insertion/replacement of the selection cassette with locus-specific primers (Table S3). All mutations were confirmed by using Sanger sequencing (1st base, Singapore) on PCR products with the same primers used for amplification.

3.4. Hemagglutination Assay

These were performed as previously described [14].

3.5. In Vivo Mouse Infections

All bacterial strains were cultured in type 1 pili-inducing conditions and used to infect seven- to eight-week-old C3H/HeN female mice (InVivos, Singapore) as previously described [21].

4. Conclusions

Using a large series of scarless, unmarked *FimH* mutants, we have recapitulated previously reported *FimH* phenotypes, validating the assignment of them to *FimH* mutations. This further demonstrates that using a combined positive-negative selection cassette is an effective way to generate definitive genetic constructs in UTI89. We have further improved the efficiency of the negative selection step by eliminating one class of undesired recombinations. These protocols and cloning strategies can also be used to generate definitive genetic constructs in other UPEC, *E. coli*, and other Enterobacteriaceae.

Supplementary Materials: The following are available online at www.mdpi.com/2076-0817/5/1/9/s1, Table S1: Strains and plasmids used in this study, Table S2: *FimH* mutations and HA titers in UTI89, Table S3: Primers used in this study.

Acknowledgments: This project was funded by National Research Foundation, Prime Minister's Office, Singapore under its NRF Research Fellowship Scheme (NRF-RF2010-10 to S.L.C.), the Genome Institute of Singapore (GIS)/Agency for Science, Technology and Research (A*STAR), and Graduate Research Scholarships from the National University of Singapore.

Author Contributions: Varnica Khetrpal, Kurosh S. Mehershahi, Siyi Chen and Swaine L. Chen designed the experiments and analyzed data. Varnica Khetrpal and Swaine L. Chen wrote the paper. Varnica Khetrpal, Kurosh Mehershahi, and Siyi Chen performed experiments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Foxman, B. The epidemiology of urinary tract infection. *Nat. Rev. Urol.* **2010**, *7*, 653–660. [[CrossRef](#)] [[PubMed](#)]
2. Rosen, D.A.; Hooton, T.M.; Stamm, W.E.; Humphrey, P.A.; Hultgren, S.J. Detection of intracellular bacterial communities in human urinary tract infection. *PLoS Med.* **2007**, *4*, e329. [[CrossRef](#)] [[PubMed](#)]
3. Robino, L.; Scavone, P.; Araujo, L.; Algorta, G.; Zunino, P.; Vignoli, R. Detection of intracellular bacterial communities in a child with *Escherichia coli* recurrent urinary tract infections. *Pathog. Dis.* **2013**, *68*, 78–81. [[CrossRef](#)] [[PubMed](#)]

4. Robino, L.; Scavone, P.; Araujo, L.; Algorta, G.; Zunino, P.; Pirez, M.C.; Vignoli, R. Intracellular bacteria in the pathogenesis of *Escherichia coli* urinary tract infection in children. *Clin. Infect. Dis.* **2014**, *59*, e158–e164. [[CrossRef](#)] [[PubMed](#)]
5. Schilling, J.D.; Lorenz, R.G.; Hultgren, S.J. Effect of Trimethoprim-Sulfamethoxazole on Recurrent Bacteriuria and Bacterial Persistence in Mice Infected with Uropathogenic *Escherichia coli*. *Infect. Immun.* **2002**, *70*, 7042–7049. [[CrossRef](#)] [[PubMed](#)]
6. Connell, I.; Agace, W.; Klemm, P.; Schembri, M.; Mårild, S.; Svanborg, C. Type 1 fimbrial expression enhances *Escherichia coli* virulence for the urinary tract. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 9827–9832. [[CrossRef](#)] [[PubMed](#)]
7. Hultgren, S.J.; Porter, T.N.; Schaeffer, A.J.; Duncan, J.L. Role of type 1 pili and effects of phase variation on lower urinary tract infections produced by *Escherichia coli*. *Infect. Immun.* **1985**, *50*, 370–377. [[PubMed](#)]
8. Wu, X.R.; Sun, T.T.A.; Medina, J.J. *In vitro* binding of type 1-fimbriated *Escherichia coli* to uroplakins Ia and Ib: Relation to urinary tract infections. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 9630–9635. [[CrossRef](#)] [[PubMed](#)]
9. Eto, D.S.; Jones, T.A.; Sundsbak, J.L.; Mulvey, M.A. Integrin-mediated host cell invasion by type 1-piliated uropathogenic *Escherichia coli*. *PLoS Pathog.* **2007**, *3*, e100. [[CrossRef](#)] [[PubMed](#)]
10. Hung, C.S.; Bouckaert, J.; Hung, D.; Pinkner, J.; Widberg, C.; DeFusco, A.; Auguste, C.G.; Strouse, R.; Langermann, S.; Waksman, G.; *et al.* Structural basis of tropism of *Escherichia coli* to the bladder during urinary tract infection. *Mol. Microbiol.* **2002**, *44*, 903–915. [[CrossRef](#)] [[PubMed](#)]
11. Malaviya, R.; Gao, Z.; Thankavel, K.; van der Merwe, P.A.; Abraham, S.N. The mast cell tumor necrosis factor alpha response to FimH-expressing *Escherichia coli* is mediated by the glycosylphosphatidylinositol-anchored molecule CD48. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 8110–8115. [[CrossRef](#)] [[PubMed](#)]
12. Martinez, J.J.; Mulvey, M.A.; Schilling, J.D.; Pinkner, J.S.; Hultgren, S.J. Type 1 pilus-mediated bacterial invasion of bladder epithelial cells. *EMBO J.* **2000**, *19*, 2803–2812. [[CrossRef](#)] [[PubMed](#)]
13. Duncan, M.J.; Mann, E.L.; Cohen, M.S.; Ofek, I.; Sharon, N.; Abraham, S.N. The distinct binding specificities exhibited by enterobacterial type 1 fimbriae are determined by their fimbrial shafts. *J. Biol. Chem.* **2005**, *280*, 37707–37716. [[CrossRef](#)] [[PubMed](#)]
14. Chen, S.L.; Hung, C.S.; Pinkner, J.S.; Walker, J.N.; Cusumano, C.K.; Li, Z.; Bouckaert, J.; Gordon, J.I.; Hultgren, S.J. Positive selection identifies an *in vivo* role for FimH during urinary tract infection in addition to mannose binding. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 22439–22444. [[CrossRef](#)] [[PubMed](#)]
15. Wright, K.J.; Seed, P.C.; Hultgren, S.J. Development of intracellular bacterial communities of uropathogenic *Escherichia coli* depends on type 1 pili. *Cell. Microbiol.* **2007**, *9*, 2230–2241. [[CrossRef](#)] [[PubMed](#)]
16. Schwartz, D.J.; Kalas, V.; Pinkner, J.S.; Chen, S.L.; Spaulding, C.N.; Dodson, K.W.; Hultgren, S.J. Positively selected FimH residues enhance virulence during urinary tract infection by altering FimH conformation. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 15530–15537. [[CrossRef](#)] [[PubMed](#)]
17. Sokurenko, E.V.; Chesnokova, V.; Dykhuizen, D.E.; Ofek, I.; Wu, X.R.; Krogfelt, K.A.; Struve, C.; Schembri, M.A.; Hasty, D.L. Pathogenic adaptation of *Escherichia coli* by natural variation of the FimH adhesin. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 8922–8926. [[CrossRef](#)] [[PubMed](#)]
18. Hannan, T.J.; Mysorekar, I.U.; Chen, S.L.; Walker, J.N.; Jones, J.M.; Pinkner, J.S.; Hultgren, S.J.; Seed, P.C. LeuX tRNA-dependent and -independent mechanisms of *Escherichia coli* pathogenesis in acute cystitis. *Mol. Microbiol.* **2008**, *67*, 116–128. [[CrossRef](#)] [[PubMed](#)]
19. Khetrapal, V.; Mehershahi, K.; Rafee, S.; Chen, S.; Lim, C.L.; Chen, S.L. A set of powerful negative selection systems for unmodified Enterobacteriaceae. *Nucl. Acids Res.* **2015**. [[CrossRef](#)] [[PubMed](#)]
20. Datsenko, K.A.; Wanner, B.L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 6640–6645. [[CrossRef](#)] [[PubMed](#)]
21. Hung, C.S.; Dodson, K.W.; Hultgren, S.J. A murine model of urinary tract infection. *Nat. Protoc.* **2009**, *4*, 1230–1243. [[CrossRef](#)] [[PubMed](#)]



Conference Report

Brighter Fluorescent Derivatives of UTI89 Utilizing a Monomeric vGFP

Majid Eshaghi ¹, Kurosh S. Mehershahi ¹ and Swaine L. Chen ^{1,2,*}

Received: 11 September 2015; Accepted: 29 December 2015; Published: 5 January 2016

Academic Editor: Catharina Svanborg

¹ Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, 1E Kent Ridge Road, NUHS Tower Block, Level 10, Singapore 119074, Singapore; eshaghim@gis.a-star.edu.sg (M.E.); mehershahiks@gis.a-star.edu.sg (K.M.)

² Infectious Diseases Group, Genome Institute of Singapore, 60 Biopolis street, Genome, #02-01, Singapore 138672, Singapore

* Correspondence: slchen@gis.a-star.edu.sg; Tel.: +65-6808-8074; Fax: +65-6808-8036

Abstract: Fluorescent proteins, especially green fluorescent protein (GFP), have been instrumental in understanding urinary tract infection pathogenesis by uropathogenic *Escherichia coli* (UPEC). We have used a recently developed GFP variant, vsfGFP-9, to create new plasmid- and chromosome-based GFP derivatives of the UPEC strain UTI89. The vsfGFP-9 strains are nearly 10× brighter with no *in vitro* growth or *in vivo* virulence defects compared to previously reported GFP-expressing UTI89 strains. The chromosomal vsfGFP-9 strain is equivalent to the wild type UTI89 during *in vivo* UTI, while both plasmid GFP constructs have an equivalent virulence defect compared to non-plasmid carrying UTI89. These new vsfGFP-9 expressing strains should be useful for further studies of the pathogenesis of UTI89, and similar strategies can be used to create improved fluorescent derivatives of other UPEC strains.

Keywords: GFP; uropathogenic *Escherichia coli*; urinary tract infection

1. Introduction

Fluorescent proteins (FPs) have been instrumental to our understanding of bacterial pathogenesis [1]. FPs have been used widely in plasmid- or chromosome-based strategies in *in vitro* and *in vivo* studies [1,2]. Plasmid-based GFPs generally have higher copy number and expression, benefitting brightness at the cost of cell-to-cell variation (due to different copy numbers in different cells), plasmid instability, and fitness defects due to plasmid carriage or high GFP expression. Fitness defects in particular then complicate studies of pathogenesis, which more often manifest *in vivo*. In contrast, chromosomal GFPs tend to have lower expression and lower brightness, limiting utility for visualizing small (or individual) bacterial collections while mitigating these other problems with variability, stability, and fitness.

Urinary tract infections (UTIs) are one of the most common bacterial infections of humans, accounting for over \$2.3 billion in medical expenditures annually in the US [3]. Most UTIs are caused by strains of *E. coli*, thus the term uropathogenic *E. coli* (UPEC). As with other infectious diseases [1,4], fluorescent proteins have been instrumental for many discoveries of the pathogenic mechanisms utilized by UPEC, including the development of intracellular bacterial communities (IBCs) [5,6], quiescent intracellular reservoirs (QIRs) [7], and avoidance of neutrophil killing [6] by the cystitis strain UTI89 [8]. For these studies, two strains are commonly used, both of which express the GFPmut3 variant of GFP [1]: UTI89 carrying plasmid pANT4 [9] and UTI89 *att_{HK022}::COM-GFP* [10]. Both of these strains have been used to monitor formation of intracellular structures during UTI

by microscopy [7,10,11], but to date UTI89/pANT4 has not been further characterized for other infection phenotypes.

Since the identification of GFPmut3, new variants of GFP demonstrate various improved properties [12]. One of these in particular, superfolder GFP (sfGFP) [13], has higher brightness and faster folding kinetics than the currently used GFPmut3. We have further improved the brightness of sfGFP by fusing it to a GFP-specific single domain antibody [14] using the vGFP strategy to create a monomeric fluorophore with 30%–50% increased brightness and pH resistance [15]. We refer to this improved sfGFP as vsfGFP-9.

We here report the creation of new derivatives of UTI89 carrying vsfGFP-9 on the chromosome or on a derivative of the pANT4 plasmid that provide nearly 10× increased brightness to the commonly used UTI89 *att_{HK022}::COM-GFP* and UTI89/pANT4, respectively. We demonstrate that these derivatives, despite the markedly higher brightness, have no fitness defects relative to the strains they are intended to replace. Furthermore, we find that the plasmid-based strains (UTI89/pANT4 and SLC-638) have an equivalent fitness defect relative to UTI89 as measured by infection load. In contrast, chromosomal expression of vsfGFP-9 produces brightness approaching that of UTI89/pANT4 without a defect in IBC formation or infection load. These new, brighter strains should be useful in future studies of the pathogenic mechanisms of UTI89, and the strategies employed here can be similarly applied to improve fluorescent derivatives of other UPEC strains.

2. Results and Discussion

2.1. New vsfGFP-9 Expressing Derivatives of UTI89 Are 10× Brighter Than Former GFP Expressing Strains

We generated UTI89 derivatives carrying plasmid (SLC-638) and chromosome (SLC-719) based vsfGFP-9 constructs intended to improve on UTI89/pANT4 and UTI89 *att_{HK022}::COM-GFP*, respectively. As controls, we also created corresponding sfGFP derivatives (plasmid: SLC-634; chromosomal: SLC-717). The vsfGFP-9 strains had no gross growth defect relative to UTI89 or to their corresponding GFPmut3 strain (SLC-638 compared with UTI89/pANT4; SLC-719 compared with UTI89 *att_{HK022}::COM-GFP*) (Figure 1a). By flow cytometry, we found that the strains carrying vsfGFP-9 were the brightest; both on the plasmid and on the chromosome, the vsfGFP-9 strains were nearly 10× brighter than their corresponding GFPmut3 strains and ~1.5× brighter than the corresponding sfGFP strain (Figure 1b). Based on Western blots for GFP, the increased brightness was due to both increased expression as well as the enhanced brightness of vsfGFP-9 over sfGFP and GFPmut3 (Figure 1c). Interestingly, SLC-719 (carrying a chromosomal vsfGFP-9) approached the brightness of UTI89/pANT4 (carrying a plasmid-based GFPmut3) (Figure 1b, light and dark green traces).

2.2. New Chromosomal vsfGFP-9 Construct Has No Fitness Defects during UTI Relative to Former GFP Expressing Strains

Because plasmid carriage as well as high GFP expression can both lead to fitness defects *in vivo*, we tested the vsfGFP-9 constructs in an *in vivo* murine model of UTI. Using competitive infections against the parental (nonfluorescent and unmodified) UTI89, we generally saw no fitness defect at 6 hpi or 24 hpi for either UTI89 *att_{HK022}::COM-GFP* or SLC-719 (chromosomal vsfGFP-9) in either the bladder or the kidney (Figure 2a,b); at 24 hpi in kidneys we saw a slight (<0.5 log) but significant defect in UTI89 *att_{HK022}::COM-GFP*. In contrast, we saw a significant defect in competitive infections for both UTI89/pANT4 and SLC-638 (plasmid vsfGFP-9) relative to UTI89 at 6 hpi and 24 hpi; however, there was no significant difference in the competitive indices between these plasmid-carrying strains.

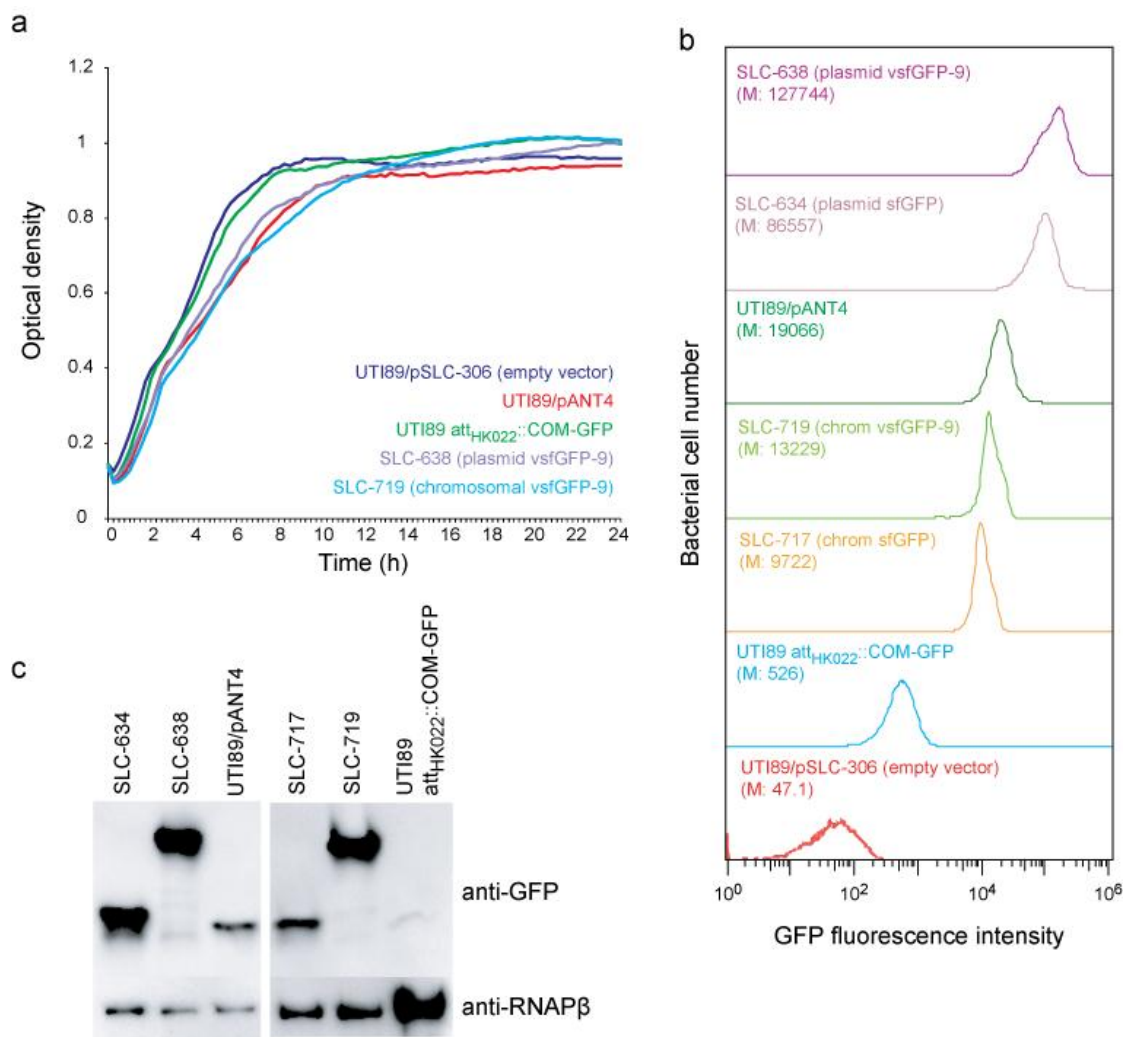


Figure 1. *In vitro* characterization of vsfGFP-9 derivatives of UTI89. (a) Growth curves in Lysogeny broth (LB) medium for the parental wt UTI89/pSLC-306 (empty vector control; dark blue), UTI89 *att*_{HK022}::COM-GFP (green), SLC-719 (chromosomal vsfGFP-9; light blue), UTI89/pANT4 (red), and SLC-638 (plasmid vsfGFP-9; purple); (b) Flow cytometry analysis of green fluorescent protein (GFP) brightness for UTI89/pSLC-306 (red), UTI89 *att*_{HK022}::COM-GFP (blue), SLC-717 (chromosomal sfGFP; brown), SLC-719 (light green), UTI89/pANT4 (dark green), SLC-634 (plasmid sfGFP; pink), and SLC-638 (purple). M indicates the median GFP fluorescence. (c) Quantification of GFP protein levels in UTI89 strains. (top) Immunoblot of samples from panel (b) using α -GFP antibody. α -RNAP β was used as a loading control.

We next tested these strains for IBC formation. Validating previous reports that used GFPmut3 expressing strains to study IBCs, we found no significant difference in the number of IBCs formed by any of the strains tested relative to wt UTI89 as quantified by LacZ staining (Figure 2c). The fluorescent strains also enabled a more convenient quantification of IBCs by direct observation under a fluorescent dissecting microscope; again no significant difference was seen between UTI89/pANT4 and SLC-638, though the number of IBCs for each strain detected was slightly higher than (though well correlated with) LacZ staining (Figure 2d). Highlighting the brightness advantage of the new strains, among the chromosomal GFP expressing strains, we were only able to detect IBCs by fluorescence in SLC-719 but not with UTI89 *att*_{HK022}::COM-GFP (except weakly in two cases; this was due to low fluorescence of this strain over background bladder fluorescence), which is consistent with the higher fluorescence of SLC-719 that we measured by FACS.

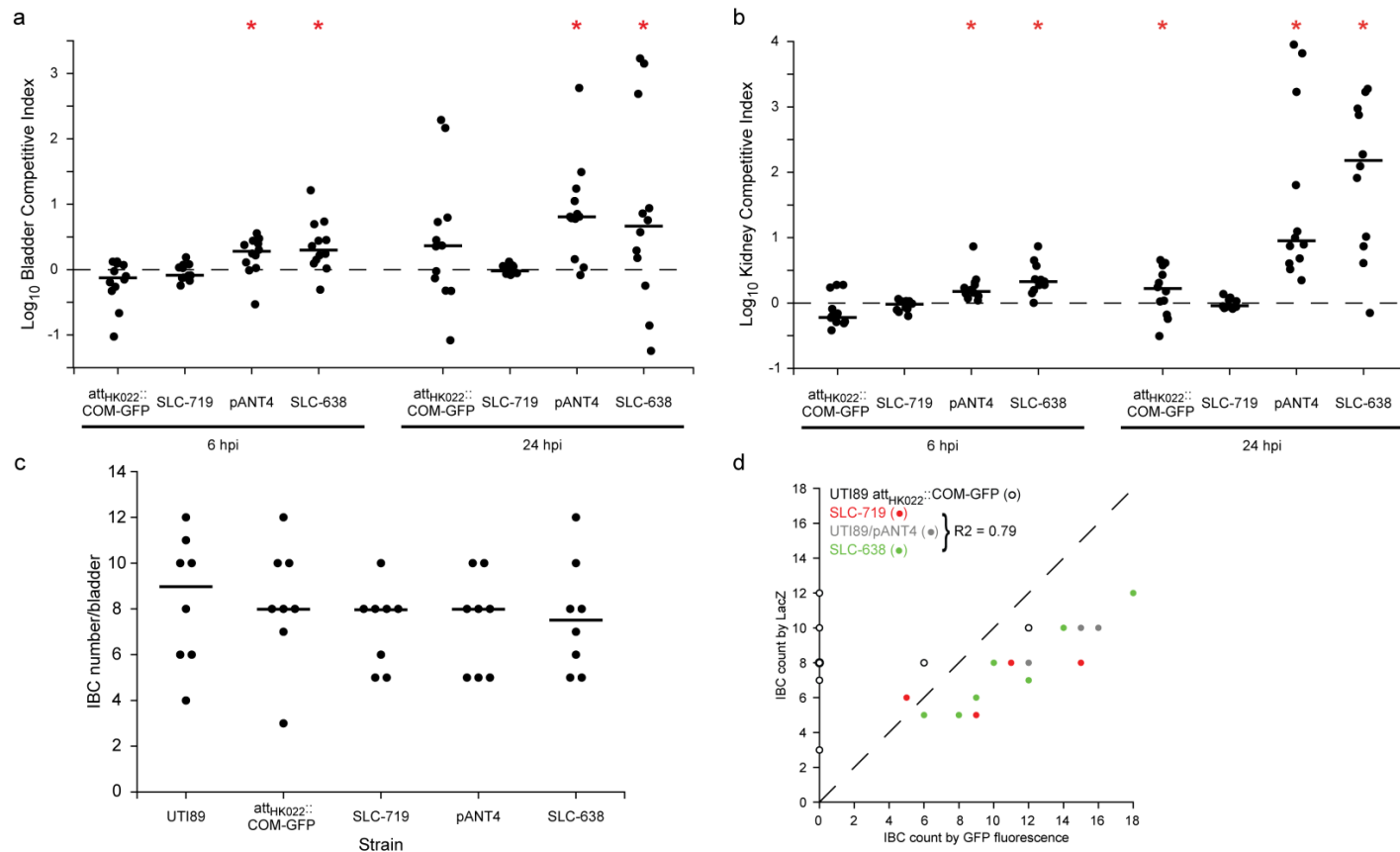


Figure 2. *In vivo* characterization of vsfGFP-9 derivatives of UTI89. Competitive infections between UTI89 $\text{att}_{HK022}::\text{COM-GFP}$, SLC-719, UTI89/pANT4, and SLC-638 and the parental wt UTI89 were performed. (a) Logarithm of the competitive index against wt UTI89 in the bladder at 6 and 24 hpi. Higher values indicate wt UTI89 outcompetes the GFP strain; (b) Log competitive index against wt UTI89 in kidneys at 6 and 24 hpi. Red bars indicate median values. *, $p < 0.05$ (two-tailed Wilcoxon signed-rank test whether log competitive indices are different from 0). (c) Quantification of IBCs at 6 hpi; (d) Comparison of GFP versus LacZ staining to quantify IBCs for plasmid-based GFP expressing strains. Data from UTI89 $\text{att}_{HK022}::\text{COM-GFP}$ (open circles), SLC-719 (red), UTI89/pANT4 (gray), and SLC-638 (green) are shown. R^2 coefficient for combined data from SLC-719, UTI89/pANT4, and SLC-638 is indicated in the legend.

3. Experimental Section

Strains and growth conditions. UTI89 [8], UTI89 *att_{HK022}::COM-GFP* [10], and UTI89/pANT4 [9,16] have been previously described. Lysogeny broth (LB) was used for all growth. Ampicillin was supplemented as required at 100 µg/mL, and kanamycin at 50µg/mL. For mouse infections, bacteria were grown under type 1 pili inducing conditions as previously described [17]. Bacterial growth curves were measured using a Bioscreen C MBR (Bioscreen, Finland) in LB medium at 37 °C for 24 h.

Construction of sfGFP and vsfGFP-9 expressing strains. All plasmids used in this study are listed in Table S1. All primers used for cloning and homologous recombination were purchased from Sigma (Singapore) and are listed in Table S2. The genes encoding sfGFP and vsfGFP-9 were amplified by PCR using primer pairs P1-P2 from plasmids pSLC-253 and pSLC-255 respectively. The PCR products were then digested and cloned into pANT4 (replacing GFPmut3) using *XbaI* and *HindIII* restriction sites to produce pSLC-282 (containing sfGFP) and pSLC-284 (containing vsfGFP-9). Nonfluorescent colonies from the pSLC-282 cloning contained plasmids without GFP, giving pSLC-306 (empty vector control for pANT4). pSLC-282 was transformed into UTI89 to give SLC-634; pSLC-284 was transformed into UTI89 to give SLC-638.

For chromosomal GFP strains, annealed complementary pairs of oligonucleotides (primer P3 and its reverse complement) containing a *rrnB_T2* transcriptional terminator [18] upstream of a modified $\sigma 70$ promoter [19] were digested with *Clal* and *XbaI* and cloned into plasmids pSLC-253 and pSLC-255 digested with the same enzymes to generate pSLC-293 (sfGFP) and pSLC-294 (vsfGFP-9). We integrated sfGFP or vsfGFP-9 between chromosomal coordinates 1044461-62 in UTI89, which is the *att_{HK022}* region, using a two-step positive-negative selection system [20]. Briefly, the kan-*P_{rhaB}-relE* cassette was amplified from pSLC-217 [20] using primers P4 and P5 and integrated into the *att_{HK022}* region using Red recombinase-mediated recombineering [21] to generate SLC-653. Primer pairs P5 and P6 were used to amplify fragments containing sfGFP or vsfGFP-9 from plasmids pSLC-293 or pSLC-294; these PCR products were used to replace the kan-*P_{rhaB}-relE* cassette by Red recombinase-mediated recombineering with negative selection on M9 media supplemented with 2% rhamnose, giving SLC-717 (sfGFP) and SLC-719 (vsfGFP-9).

Western blots. Primary antibodies (and dilutions) used were anti-GFP mouse monoclonal antibody (1:6000; Santa Cruz Biotechnology, Shanghai, China), anti-RNA Polymerase β (1:6000; Santa Cruz Biotechnology, Shanghai, China). The secondary antibodies were ECL™ Anti-rabbit IgG and ECL™ Anti-mouse IgG, both conjugated with Horseradish peroxidase (HRP) and used at a 1:10,000 dilution.

Mouse infections. Infections were performed as previously described [17]. Data shown is the result of two separate experiments performed on separate days with four to six mice per strain and per time point. Six to seven-week old C3H/HeN female mice were obtained from Harlan (Israel). In co-infections with UTI89 and SLC-719 (chromosomal vsfGFP-9), SLC-719 colonies were differentiated from UTI89 cells by visualization of green fluorescence under 10× magnification because neither strain carries an antibiotic resistance cassette. In coinfections with UTI89 and UTI89 *att_{HK022}::COM-GFP*, UTI89 *att_{HK022}::COM-GFP* colonies were quantified by plating on LB supplemented with kanamycin, and UTI89 colonies were calculated by subtraction of UTI89 *att_{HK022}::COM-GFP* titers from total titers quantified on LB plates. In each experiment, two to three of these LB plates were also used to quantify UTI89 *att_{HK022}::COM-GFP* titers using green fluorescence under 10× magnification; in all cases, the fluorescence-based quantification was within 10% of the kanamycin-based quantification.

IBCs were quantified in single infections for UTI89, UTI89 *att_{HK022}::COM-GFP*, UTI89/pANT4, SLC-717, SLC-719, and SLC-638. At 6 hpi, infected mice were sacrificed, and harvested bladders were hemisected, stretched onto silicone pads, and fixed with 3% paraformaldehyde (Sigma) as previously described [17]. IBCs from GFP-expressing strains were quantified by fluorescence at 10× magnification. All samples were then subjected to X-Gal staining for independent quantification as previously reported [11]. Data shown is the result of two separate experiments performed on separate days with four mice per strain per experiment.

Flow cytometry. Flow cytometry analysis for GFP expression was performed on a S3™ Cell Sorter (Bio-Rad, Hercules, CA, USA).

4. Conclusions

We have created new plasmid- and chromosome-based GFP expressing derivatives of UTI89. Through using a new vsfGFP-9 gene as well as optimization of expression, these are nearly 10× brighter than previously published GFP UTI89 derivatives, and have no *in vitro* or *in vivo* defects compared with previous GFP-expressing strains. These strains should be useful for future studies of UTI89 pathogenesis and for the creation of vsfGFP-9 derivatives of other UPEC and *E. coli*.

Supplementary Materials: The following are available online at www.mdpi.com/2076-0817/5/1/3/s1, Table S1: Plasmids used in this work, Table S2: Primer sequences used in this study.

Acknowledgments: This project was funded by National Research Foundation, Prime Minister's Office, Singapore under its NRF Research Fellowship Scheme (NRF-RF2010-10 to Swaine L. Chen) and the Genome Institute of Singapore (GIS)/Agency for Science, Technology and Research (A*STAR).

Author Contributions: Majid Eshaghi and Swaine L. Chen designed the experiments, analyzed data, and wrote the paper. Majid Eshaghi and Kurosh Mehershahi performed experiments.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Valdivia, R.H.; Hromockyj, A.E.; Monack, D.; Ramakrishnan, L.; Falkow, S. Applications for green fluorescent protein (gfp) in the study of hostpathogen interactions. *Gene* **1996**, *173*, 47–52. [[CrossRef](#)]
- Hautefort, I.; Proença, M.J.; Hinton, J.C. Single-copy green fluorescent protein gene fusions allow accurate measurement of salmonella gene expression *in vitro* and during infection of mammalian cells. *Appl. Environ. Microbiol.* **2003**, *69*, 7480–7491. [[CrossRef](#)] [[PubMed](#)]
- Foxman, B. The epidemiology of urinary tract infection. *Nat. Rev. Urol.* **2010**, *7*, 653–660. [[CrossRef](#)] [[PubMed](#)]
- Hautefort, I.; Hinton, J.C. Measurement of bacterial gene expression *in vivo*. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2000**, *355*, 601–611. [[CrossRef](#)] [[PubMed](#)]
- Garofalo, C.K.; Hooton, T.M.; Martin, S.M.; Stamm, W.E.; Palermo, J.J.; Gordon, J.I.; Hultgren, S.J. *Escherichia coli* from urine of female patients with urinary tract infections is competent for intracellular bacterial community formation. *Infect. Immun.* **2007**, *75*, 52–60. [[CrossRef](#)] [[PubMed](#)]
- Justice, S.S.; Hung, C.; Theriot, J.A.; Fletcher, D.A.; Anderson, G.G.; Footer, M.J.; Hultgren, S.J. Differentiation and developmental pathways of uropathogenic *Escherichia coli* in urinary tract pathogenesis. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 1333–1338. [[CrossRef](#)] [[PubMed](#)]
- Mysorekar, I.U.; Hultgren, S.J. Mechanisms of uropathogenic *Escherichia coli* persistence and eradication from the urinary tract. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 14170–14175. [[CrossRef](#)] [[PubMed](#)]
- Chen, S.L.; Hung, C.-S.; Xu, J.; Reigstad, C.S.; Magrini, V.; Sabo, A.; Blasiar, D.; Bieri, T.; Meyer, R.R.; Ozersky, P. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: A comparative genomics approach. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 5977–5982. [[CrossRef](#)] [[PubMed](#)]
- Lee, A.K.; Falkow, S. Constitutive and inducible green fluorescent protein expression in bartonella henselae. *Infect. Immun.* **1998**, *66*, 3964–3967. [[PubMed](#)]
- Wright, K.J.; Seed, P.C.; Hultgren, S.J. Uropathogenic *Escherichia coli* flagella aid in efficient urinary tract colonization. *Infect. Immun.* **2005**, *73*, 7657–7668. [[CrossRef](#)] [[PubMed](#)]
- Justice, S.S.; Lauer, S.R.; Hultgren, S.J.; Hunstad, D.A. Maturation of intracellular *Escherichia coli* communities requires *sra*. *Infect. Immun.* **2006**, *74*, 4793–4800. [[CrossRef](#)] [[PubMed](#)]
- Shaner, N.C.; Steinbach, P.A.; Tsien, R.Y. A guide to choosing fluorescent proteins. *Nat. Methods* **2005**, *2*, 905–909. [[CrossRef](#)] [[PubMed](#)]
- Pédelacq, J.-D.; Cabantous, S.; Tran, T.; Terwilliger, T.C.; Waldo, G.S. Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.* **2006**, *24*, 79–88. [[CrossRef](#)] [[PubMed](#)]

14. Kirchhofer, A.; Helma, J.; Schmidthals, K.; Frauer, C.; Cui, S.; Karcher, A.; Pellis, M.; Muyldermans, S.; Casas-Delucchi, C.S.; Cardoso, M.C. Modulation of protein properties in living cells using nanobodies. *Nat. Struct. Mol. Biol.* **2010**, *17*, 133–138. [[CrossRef](#)] [[PubMed](#)]
15. Eshaghi, M.; Sun, G.; Gruter, A.; Lim, C.L.; Chee, Y.C.; Jung, G.; Jauch, R.; Wohland, T.; Chen, S.L. Rational structure-based design of bright gfp-based complexes with tunable dimerization. *Angew. Chem. Int. Ed. Engl.* **2015**, *127*, 14158–14162.
16. Anderson, G.G.; Palermo, J.J.; Schilling, J.D.; Roth, R.; Heuser, J.; Hultgren, S.J. Intracellular bacterial biofilm-like pods in urinary tract infections. *Science* **2003**, *301*, 105–107. [[CrossRef](#)] [[PubMed](#)]
17. Hung, C.S.; Dodson, K.W.; Hultgren, S.J. A murine model of urinary tract infection. *Nat. Protoc.* **2009**, *4*, 1230–1243. [[CrossRef](#)] [[PubMed](#)]
18. Orosz, A.; Boros, I.; Venetianer, P. Analysis of the complex transcription termination region of the *Escherichia coli* *rrnB* gene. *Eur. J. Biochem.* **1991**, *201*, 653–659. [[CrossRef](#)] [[PubMed](#)]
19. Anderson, J.; Dueber, J.E.; Leguia, M.; Wu, G.C.; Goler, J.A.; Arkin, A.P.; Keasling, J.D. Bglbricks: A flexible standard for biological part assembly. *J. Biol. Eng.* **2010**, *4*, 1–12. [[CrossRef](#)] [[PubMed](#)]
20. Khetrapal, V.; Mehershahi, K.; Rafee, S.; Chen, S.; Lim, C.L.; Chen, S.L. A set of powerful negative selection systems for unmodified enterobacteriaceae. *Nucleic Acids Res.* **2015**, *43*, e83. [[CrossRef](#)] [[PubMed](#)]
21. Datsenko, K.A.; Wanner, B.L. One-step inactivation of chromosomal genes in *Escherichia coli* k-12 using pcr products. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 6640–6645. [[CrossRef](#)] [[PubMed](#)]



© 2016 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons by Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).