

# High-throughput Experimental and Computational Studies of Bacterial Evolution



Lars Barquist  
Queens' College  
University of Cambridge

A thesis submitted for the degree of

*Doctor of Philosophy*

XX August 2013



# Declaration

## HIGH-THROUGHPUT EXPERIMENTAL AND COMPUTATIONAL STUDIES OF BACTERIAL EVOLUTION

The work presented in this dissertation was carried out at the Wellcome Trust Sanger Institute between October 2009 and August 2013. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. This dissertation does not exceed the limit of 60,000 words as specified by the Faculty of Biology Degree Committee. This dissertation has been typeset in 12pt Computer Modern font using L<sup>A</sup>T<sub>E</sub>X according to the specifications set by the Board of Graduate Studies and the Faculty of Biology Degree Committee. No part of this dissertation or anything substantially similar has been or is being submitted for any other qualification at any other university.



## **Abstract**

---

# Contents

Declaration	i
Contents	v
List of Figures	ix
List of Tables	xi
List of Symbols	xviii
Introduction	xix
<b>1 Querying bacterial genomes with transposon-insertion sequencing</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 Protocols . . . . .	5
1.2.1 Transposon mutagenesis . . . . .	6
1.2.2 Pool construction . . . . .	7
1.2.3 Enrichment of transposon-insertion junctions . . . . .	7
1.2.4 Sequencing . . . . .	8
1.3 Reproducibility, accuracy, and concordance with previous methods . . . . .	9
1.4 Identifying gene requirements . . . . .	10
1.5 Determining conditional gene requirements . . . . .	12
1.6 Monitoring ncRNA contributions to fitness . . . . .	14
1.7 Limitations . . . . .	16
1.8 The future of transposon-insertion sequencing . . . . .	17

---

<b>2 A comparison of dense transposon insertion libraries in the <i>Salmonella</i> serovars Typhi and Typhimurium</b>	<b>19</b>
2.1 Introduction . . . . .	19
2.1.1 The genus <i>Salmonella</i> . . . . .	20
2.1.2 Host adaptation and restriction . . . . .	22
2.1.3 Serovars Typhi and Typhimurium . . . . .	24
2.2 Materials and Methods . . . . .	26
2.2.1 Strains . . . . .	27
2.2.2 Annotation . . . . .	27
2.2.3 Creation of <i>S. Typhimurium</i> transposon mutant library . . . . .	27
2.2.4 DNA manipulations and sequencing . . . . .	27
2.2.5 Sequence analysis . . . . .	28
2.2.6 Statistical analysis of required genes . . . . .	28
2.3 Results and Discussion . . . . .	30
2.3.1 TraDIS assay of every <i>Salmonella</i> Typhimurium protein-coding gene	30
2.3.2 Cross-species comparison of genes required for growth . . . . .	32
2.3.3 Serovar-specific genes required for growth . . . . .	35
2.3.4 TraDIS provides resolution sufficient to evaluate ncRNA contributions to fitness . . . . .	46
2.3.5 sRNAs required for competitive growth . . . . .	50
2.4 Conclusions . . . . .	52
<b>3 Methods for the analysis of TraDIS experiments, with an application to <i>Salmonella</i> macrophage invasion</b>	<b>55</b>
3.1 Introduction . . . . .	55
3.1.1 <i>Salmonella</i> interactions with macrophage . . . . .	56
3.1.2 Conditional gene fitness . . . . .	57
3.2 Experimental methods . . . . .	59
3.2.1 Strains and cell lines . . . . .	59
3.2.2 Preparation of THP-1 cells . . . . .	59
3.2.3 Preparation of transposon libraries . . . . .	60
3.2.4 Infection assay . . . . .	60
3.3 Analysis of conditional gene fitness using TraDIS . . . . .	61

---

3.3.1	Experimental design . . . . .	61
3.3.2	Mapping insertion sites . . . . .	61
3.3.3	Quality control . . . . .	62
3.3.4	Inter-library normalization . . . . .	65
3.3.5	Identifying fitness effects . . . . .	66
3.3.5.1	Theory . . . . .	66
3.3.5.2	Application to macrophage infection data . . . . .	69
3.3.6	Functional analysis of gene sets that affect fitness . . . . .	71
3.4	Results and Discussion . . . . .	73
<b>4</b>	<b>Detecting Rho-independent terminators in genomic sequence with covariance models</b>	<b>83</b>
4.1	Introduction . . . . .	83
4.1.1	Rho-independent termination . . . . .	84
4.1.2	Previous approaches to identifying intrinsic terminators . . . . .	85
4.1.3	Covariance models . . . . .	87
4.2	Methods . . . . .	89
4.2.1	Construction of a covariance model for Rho-independent terminators	89
4.2.2	RNIE run modes . . . . .	90
4.2.3	Definitions . . . . .	91
4.3	Results . . . . .	93
4.3.1	Alpha benchmark . . . . .	93
4.3.2	Beta benchmark . . . . .	95
4.3.3	A novel termination motif in <i>Mycobacterium tuberculosis</i> . . . . .	95
<b>5</b>	<b>Kingdom-wide discovery of bacterial intrinsic termination motifs</b>	<b>99</b>
5.1	Introduction . . . . .	99
5.2	Methods . . . . .	100
5.2.1	Genome-wise motif discovery . . . . .	100
5.2.2	Clustering covariance models . . . . .	101
5.2.3	Building consensus covariance models . . . . .	102
5.2.4	Genome annotation . . . . .	102
5.2.5	Analysis of expression data . . . . .	103

---

5.3	Results . . . . .	104
5.3.1	Kingdom-wide motif discovery . . . . .	104
5.3.2	Lineage-specific differences in RIT utilization . . . . .	108
5.3.3	Non-canonical putative attenuators of transcription . . . . .	111
5.4	Discussion . . . . .	111
	<b>Publications</b>	<b>113</b>
	<b>References</b>	<b>116</b>

# List of Figures

1.1	Transposon-insertion sequencing protocols . . . . .	5
1.2	Applications of transposon-insertion sequencing to non-coding RNAs . . . . .	15
2.1	Genomic acquisitions in the evolution of the salmonellae . . . . .	21
2.2	The distribution of gene-wise insertion indexes in <i>S. Typhi</i> . . . . .	29
2.3	Genome-wide transposon mutagenesis of <i>S. Typhimurium</i> . . . . .	31
2.4	Comparison of required genes . . . . .	33
2.5	Comparison of cell surface operon structure and requirements . . . . .	41
2.6	H-NS enrichment across the SPI-2 locus . . . . .	42
2.7	Proposed differences in sRNA utilization . . . . .	51
3.1	Biogenesis of the <i>Salmonella</i> containing vacuole . . . . .	58
3.2	Principal component analysis of TraDIS macrophage infection assays . . . . .	64
3.3	Smear plot of differences in logFC over macrophage infection between <i>S. Typhimurium</i> and <i>S. Typhi</i> . . . . .	70
3.4	Smear plot of logFC in mutant prevalences over macrophage infection in <i>S. Typhimurium</i> . . . . .	71
3.5	Smear plot of logFC in mutant prevalences over macrophage infection in <i>S. Typhi</i> . . . . .	72
3.6	Walking hypergeometric test for depletion of insertions in the <i>S. Typhimurium</i> flagellar subsystem . . . . .	74
3.7	Mutant depletion in the <i>S. Typhimurium</i> flagellar subsystem . . . . .	76
4.1	Rho-independent termination . . . . .	84
4.2	TransTermHP motif . . . . .	86

---

4.3	Covariance model architecture . . . . .	88
4.4	Alpha benchmark . . . . .	94
4.5	Beta benchmark . . . . .	96
4.6	Putative mycobacterial transcription termination motif . . . . .	97
5.1	Example alignment of cluster consensus sequences . . . . .	106
5.2	Most informative sequences for nine canonical RIT clusters . . . . .	109
5.3	Analysis of diverse RNA-seq datasets confirm canonical terminator activity	111

# List of Tables

1.1	Summary of transposon-insertion sequencing studies to date . . . . .	4
2.1	Core genome functions in <i>S. Typhimurium</i> . . . . .	34
2.2	Phage elements in <i>S. Typhimurium</i> . . . . .	36
2.3	Phage elements in <i>S. Typhi</i> . . . . .	37
2.4	Genes uniquely required in <i>S. Typhimurium</i> . . . . .	39
2.5	Candidate required genes affected by H-NS binding in <i>S. Typhimurium</i> .	43
2.6	Genes uniquely required in <i>S. Typhi</i> . . . . .	45
2.7	Candidate required ncRNAs . . . . .	49
3.1	Summary statistics for macrophage infection assay sequencing runs . .	62
3.2	Pearson's <i>r</i> between replicated TraDIS experiments . . . . .	63
3.3	<i>S. Typhimurium</i> pathways putatively involved in macrophage infection .	75
3.4	Bacterial secretion system genes implicated in <i>S. Typhimurium</i> infection of macrophages . . . . .	77
3.5	Genes putatively involved in <i>S. Typhi</i> infection of macrophages . . . .	78
4.1	Control genomes . . . . .	92

---

# List of Symbols

## Roman Symbols

A, C, G, T, U	Adenine, Cytosine, Guanine, Thymine, Uracil
Fe(II)	Ferrous iron
Fe(III)	Ferric iron

## Greek Symbols

$\lambda$	Phage lambda
$\sigma^E$	$\sigma^{24}$ , extracytoplasmic stress sigma factor
$\sigma^S$	$\sigma^{38}$ , starvation/stationary phase sigma factor

## Amino Acids

Ala, A	Alanine
Arg, R	Arginine
Asn, N	Asparagine
Asp, D	Aspartic acid (Aspartate)
Cys, C	Cysteine
Gln, Q	Glutamine
Glu, E	Glutamic acid (Glutamate)

---

Gly, G	Glycine
His, H	Histidine
Ile, I	Isoleucine
Leu, L	Leucine
Lys, K	Lysine
Met, M	Methionine
Phe, F	Phenylalanine
Pro, P	Proline
Ser, S	Serine
Thr, T	Threonine
Trp, W	Tryptophan
Tyr, Y	Tyrosine
Val, V	Valine

### **Acronyms and Abbreviations**

BALB	Bagg albino (mouse)
BLAST	Basic local alignment search tool
bp	Base pair
CCAL	Creative Commons attribution license
CCD	Charge-coupled device
CDP	Cytidine diphosphate glucose
ChIP-seq	Chromatin immunoprecipitation sequencing
cI	Clear 1 ( $\lambda$ repressor protein)

---

CM	Covariance model
CPM	Counts per million (reads)
CYK	Cocke-Younger-Kasami (algorithm)
ddNTP	dideoxynucleotide
DeADMAn	Designer microarrays for defined mutant analysis
DNA	Deoxyribonucleic acid
dNTP	deoxynucleotide
DSB	Disulfide bond
E-value	Expect value
ECA	Enterobacterial common antigen
EHEC	Enterohemorrhagic <i>Escherichia coli</i>
EM	Expectation-maximization
FASTA	Fast alignment
FDR	False discovery rate
FMN	Flavin mononucleotide
FPR	False positive rate
GEBA	Genomic encyclopedia of bacteria and archaea
GLM	Generalized linear model
GO	Gene ontology
GTM	Global transposon mutagenesis
HIRAN	HIP116, Rad5p N-terminal
HITS	High-throughput insertion tracking by deep sequencing

---

HMM	Hidden Markov model
iid	Independent identically distributed (random variable)
INSeq	Insertion sequencing
kb	Kilobase
KEGG	Kyoto encyclopedia of genes and genomes
LEE	Locus of enterocyte effacement
LLR	$\log_2$ -likelihood ratios
logFC	$\log_2$ fold-change
LPS	Lipopolysaccharide
MATT	Microarray tracking of transposon mutants
Mb	Megabase
MCC	Matthews correlation coefficient
MFE	Minimum free energy
MIS	Most informative sequence
ncRNA	non-coding RNA
NOGD	Nonorthologous gene displacement
OMP	Outer membrane protein
ORF	Open reading frame
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate-buffered saline
PCA	Principal component analysis

---

PCR	Polymerase chain reaction
PMA	Phorbol myristate acetate
PPV	Positive predictive value
RIT	Rho-independent terminator
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RNAP	RNA polymerase
RNase	Ribonuclease
RPMI	Rosewell Park Memorial Institute (cell culture medium)
RSS	Reciprocal similarity score
SAGE	Serial analysis of gene expression
SCV	<i>Salmonella</i> containing vacuole
SPI	<i>Salmonella</i> pathogenicity island
SPV	<i>Salmonella</i> plasmid virulence (genes)
sRNA	Bacterial small RNA
SRP	Signal recognition particle
STM	Signature-tagged mutagenesis
T3SS	Type III secretion system
TMM	Trimmed mean of M-values
tmRNA	Transfer-messenger RNA
Tn-seq	Transposon mutagenesis and sequencing
TNF- $\alpha$	Tumor necrosis factor $\alpha$

---

TraDIS	Transposon directed insertion sequencing
TraSH	Transposon site hybridization
TRIT	Tuberculosis Rho-independent terminator
tRNA	Transfer RNA

# Introduction

Bacteria possess a remarkable ability to adapt. This ability has allowed bacteria to colonize almost every environment on Earth, from deep sea hydrothermal vents (Jørgensen, Isaksen, and Jannasch, 1992) to cryogenic brine lakes (Murray et al., 2012) to animal hosts (Finlay and Cossart, 1997). Indeed, the ability of bacteria to establish symbiotic relationships with host cells was a critical step in the origin of so-called “higher” eukaryotic life (Sagan, 1967). While the origins of some bacterial adaptations are buried in the deep time of over 1.5 billion years of evolution (Doolittle, Feng, et al., 1996), such as the differing bauplans observed across phyla, others are far more recent, such as the emergence of *Yersinia pestis* as a human pathogen around 20,000 years ago (Achtman, Zurth, et al., 1999) or the contemporary development of specialized invasive lineages of non-typhoidal *Salmonella* in immunocompromised individuals in sub-Saharan Africa (Feasey et al., 2012; Okoro et al., 2012). Many factors likely contribute to this continuous adaptation, including large population sizes, short generation times, wide-spread homologous recombination between related strains, and a capacity for horizontal gene transfer. These factors, particularly homologous recombination and horizontal gene transfer, make the definition of species in bacteria contentious (Achtman and Wagner, 2008; Doolittle and Zhaxybayeva, 2009), and have led to some questioning the viability of a bacterial species concept altogether. For the present I will leave these matters to those better informed than myself, and work within the established, though flawed, taxonomy.

The work in this thesis is concerned with the study of bacterial evolution and adaptation on two very different time scales. In the first section, consisting of chapters 1, 2, and 3, I describe a recently emerged high-throughput technology for probing gene function, transposon-insertion sequencing (Barquist, Boinett, and Cain, 2013), and its application to the study of functional differences in two important human pathogens, *Salmonella enterica* subspecies *enterica* serovars Typhi and Typhimurium. These two

---

serovars diverged only approximately 50,000 years ago (Kidgell et al., 2002), yet have developed very different host ranges and cause very different diseases, with *S. Typhi* causing a life-threatening system disease exclusively in humans, and *S. Typhi* causing primarily a mild gastrointestinal disease in a wide range of hosts. Chapter 2 uses transposon-insertion sequencing to probe differences in gene requirements during growth on rich laboratory media, revealing differences in requirements for genes involved in iron-utilization and cell-surface structure biogenesis, as well as in requirements for non-coding RNA (Barquist, Langridge, et al., 2013). In chapter 3 I more directly probe the genomic features responsible for differences in serovar pathogenicity by analyzing transposon-insertion sequencing data produced following a two hour infection of human macrophage, revealing large differences in the selective pressures felt by these two closely related strains in the same environment.

The second section, chapters 4 and 5, uses statistical models of sequence variation, i.e. covariance models, to examine the evolution of intrinsic termination across the bacterial kingdom. Chapter 4 provides background and motivation in the form of a method for identifying Rho-independent terminators using covariance models built from deep alignments of experimentally-verified terminators from *Escherichia coli* and *Bacillus subtilis* (Gardner, Barquist, et al., 2011). In the course of the development of this method I discovered a novel putative intrinsic terminator in *Mycobacterium tuberculosis*. In chapter 5, I extend this approach to *de novo* discovery of intrinsic termination motifs across the bacterial phylogeny. I present evidence for lineage-specific variations in canonical Rho-independent terminator composition, as well as discover seven non-canonical putative termination motifs. Using a collection of 40 publicly available RNA-seq datasets, I provide evidence for the function of these elements as *bona fide* transcriptional attenuators.

# Chapter 1

## Querying bacterial genomes with transposon-insertion sequencing

*This chapter is an expansion of the previously published article “Approaches to querying bacterial genomes using transposon-insertion sequencing” (Barquist, Boinett, and Cain, 2013). Amy K. Cain and Christine J. Boinett (Pathogen Genomics, Wellcome Trust Sanger Institute) contributed to the research of the original article. All final language is my own.*

### 1.1 Introduction

The study of gene essentiality has its roots in evolutionary theory, systems biology, and comparative genomics, and has been instrumental in the development of the emerging discipline of synthetic biology. Koonin summarizes the major scientific motivation behind this line of research succinctly: “When reverse-engineering a complex machine, one basic goal is to draw up a list of essential parts” (Koonin, 2003). The earliest attempt at constructing such a minimal gene set involved a comparison between the first two complete genomes sequenced: *Mycoplasma genitalium* and *Haemophilus influenzae* (Mushegian and Koonin, 1996). Both of these organisms are pathogens with highly reduced genomes; however, they are derived from distant branches of the bacterial phylogeny being Gram-positive and -negative, respectively. Orthology prediction based on sequence similarity identified 240 genes shared between the two organisms. However,

a number of essential pathways were found to be incomplete in this set due to non-orthologous gene displacement (NOGD), and a true minimal gene set was estimated to contain 256 genes. NOGD apparently occurs when an unrelated but functionally analogous gene is introduced in a lineage, and subsequently the ancestral gene is lost. The sequencing of complete genomes has shown that this phenomena is surprisingly wide-spread, and only ~60 genes appear to be universally conserved (Koonin, 2003). Rather obviously in hindsight, it appears that gene essentiality is highly dependent on the evolutionary and systems context in which the gene occurs - our essential parts list depends on the machine we wish to build.

Large-scale experimental studies seem to confirm this. A range of approaches have been taken to experimentally determining the ‘essential’ genes of a diverse array of organisms. These include plasmid-insertion mutagenesis in *Bacillus subtilis* (Kobayashi et al., 2003), antisense-mediated gene inactivation in *Haemophilus influenzae* (Akerley et al., 2002), transposon mutagenesis in *Pseudomonas aeruginosa* (Jacobs et al., 2003), and insertion-duplication mutagenesis in *Salmonella enterica* (Knuth et al., 2004). However, the “gold standard” for the determination of gene essentiality is repeated failure to generate targeted single gene deletions. Comprehensive single gene deletion libraries have been created for the  $\gamma$ -proteobacteria *E. coli* and *Acinetobacter baylyi* (Baba et al., 2006; Berardinis et al., 2008) where  $\lambda$ -red mediated recombineering has simplified the generation of defined deletions (Datsenko and Wanner, 2000), though the process is still extremely labor-intensive. Typical estimates for essential gene sets determined by these various techniques range from less than 300 to 600 genes, depending on the organism. This variability is likely dependent on a variety of factors, including false positives and negatives due to experimental techniques, the growth conditions of the experiment, intrinsic properties of the cell being manipulated, and accidents of evolution. Now that it has become feasible to synthesize a viable bacterial chromosome (Gibson et al., 2010), a deeper understanding of the factors affecting gene requirements in diverse conditions is the next hurdle on the road to engineering truly synthetic life.

A common approach to identifying genomic regions required for survival under a particular set of conditions is to screen large pools of mutants simultaneously. This can be done with defined mutants (Baba et al., 2006; Hobbs, Astarita, and Storz, 2010), but this is both labor-intensive and requires accurate genomic annotation, which can be particularly difficult to define for non-coding regions. An alternative to defined

libraries is the construction and analysis of random transposon-insertion libraries. The original application of this method used DNA hybridization to track uniquely tagged transposon-insertions in *Salmonella enterica* serovar Typhimurium over the course of BALB/c mouse infection (Hensel, Shea, Gleeson, et al., 1995). DNA hybridization was eventually superseded by methods that used microarray detection of the genomic DNA flanking insertion sites, variously known as TraSH, MATT, and DeADMAn (reviewed in Mazurkiewicz et al., 2006). However, these methods suffered from many of the problems microarrays generally suffer from: difficulty detecting low-abundance transcripts, mis-hybridization, probe saturation, and difficulty identifying insertion sites precisely.

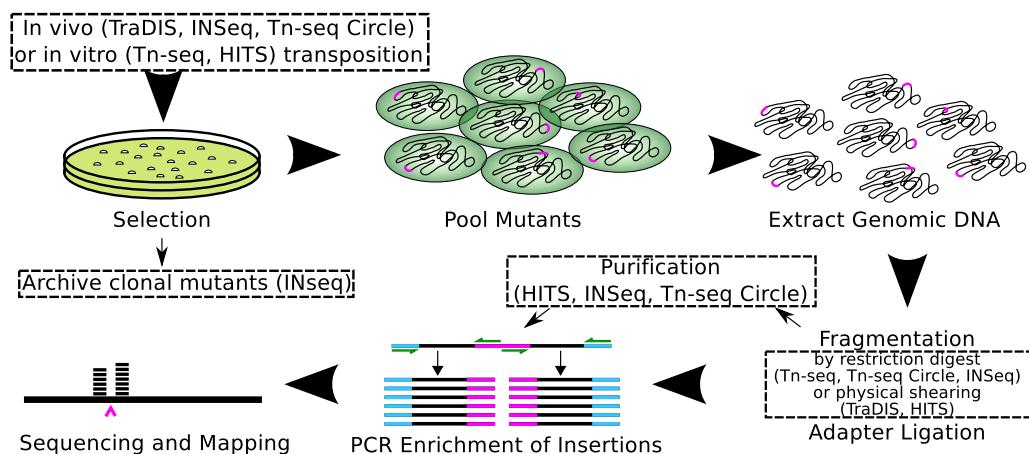
The application of high-throughput sequencing to the challenge of determining insertion location and prevalence solves many of these problems. Interestingly, the first application of transposon-insertion sequencing, developed by Hutchison et al. (1999), actually predates the development of microarray-based methods. However, this was applied to libraries of only approximately 1000 transposon mutants in highly reduced *Mycoplasma* genomes, and the difficulty of sequencing at the time prevented wide spread adoption or high resolution. Modern high-throughput sequencing technology allows the methods discussed in this chapter to routinely monitor as many as one million mutants simultaneously in virtually any genetically tractable microorganism.

**Table 1.1: Summary of transposon-insertion sequencing studies to date.** Columns: 1) study reference, 2) organism mutagenized, 3) number of mutants generated, 4) insertion density, 5) brief description of the application, 6) transposon used, 7) method name coined, if any.

Study	Organism	Total Mutants	Density	Application	Tn used	Name Coined
Hutchison et al. (1999)	<i>M. genitalium</i> <i>M. pneumoniae</i>	1291 918	1/850 bp 1/850 bp	Gene requirements	Tn4001	GTM
Goodman, McNulty, et al. (2009)	<i>B. thetaiotaomicron</i>	2 X 35,000	1/182 bp	Establishment in a murine model of the human gut	Mariner	INSeq
Gawronska et al. (2009)	<i>H. influenzae</i>	75,000	1/32 bp	Prolonged survival in murine lung	Mariner	HITS
Opipinen, Bodl, and Camilli (2009)	<i>S. pneumoniae</i>	6 x 25,000	1/91 bp	Transcriptional regulation and carbohydrate transport	Mariner	Tn-seq
Langridge et al. (2009)	<i>S. Typhi</i>	1.1 million	1/13 bp	Gene requirements, bile tolerance	Tn5	TraDIS
Gallagher, Shendure, and Manoil (2011)	<i>P. aeruginosa</i>	100,000	1/65 bp	Tobramycin resistance	Mariner	Tn-seq (circle method)
Eckert et al. (2011)	<i>E. coli</i>	19 x 95	N/A	Colonization of bovine intestinal tract; retrospective re-evaluation of a STM study	Tn5	-
Christen et al. (2011)	<i>C. crescentus</i>	800,000	1/8 bp	Genomic requirements	Tn5	-
Giffin et al. (2011)	<i>M. tuberculosis</i>	2 X 100,000	1/120 bp	Gene requirements and cholesterol utilization	Mariner	-
Khatiwara et al. (2012)	<i>S. Typhimurium</i>	16,000	1/610 bp	Bile, starvation, and heat tolerance	Tn5	-
Mann et al. (2012)	<i>S. pneumoniae</i>	9,000-24,000	Varying	Determining roles of sRNAs in pathogenesis	Mariner	-
Opipinen and Camilli (2012)	<i>S. pneumoniae</i>	4,000 - 30,000	Varying	Stress response and metabolism <i>in vitro</i> and murine <i>in vivo</i> colonization	Mariner	-
Brutinel and Granick (2012)	<i>S. oneidensis</i>	50,000	1/191 bp	Gene requirements and metabolism	Mariner	-
Zhang et al. (2012)	<i>M. tuberculosis</i>	2 x 100,000	1/120 bp	Genomic requirements	Mariner	-
Klein et al. (2012)	<i>P. gingivalis</i>	N/A	1/43 bp	Gene requirements	Mariner	-
Pickard, Kingsley, et al. (2013)	<i>S. Typhi</i>	1.1 million	1/13 bp	Bacteriophage infection	Tn5	-
Barquist, Langridge, et al. (2013)	<i>S. Typhi</i>	1.1million	1/13 bp	Comparison of genomic requirements between two <i>Salmonella</i> serovars	Tn5	-
	<i>S. Typhimurium</i>	930,000	1/9 bp			

## 1.2 Protocols

Several methods were developed concurrently for high-throughput sequencing of transposon insertion sites: TraDIS (Langridge et al., 2009), INSeq (Goodman, McNulty, et al., 2009), HITS (Gawronski et al., 2009), and Tn-seq (Opijken, Bodi, and Camilli, 2009) followed by Tn-seq Circle (Gallagher, Shendure, and Manoil, 2011) and refinements to the INSeq protocol (Goodman, Wu, and Gordon, 2011). All of these protocols follow the same basic workflow with minor variations (see Figure 1.1; Table 1.1): transposon mutagenesis and construction of pools of single insertion mutants; enrichment of transposon-insertion junctions; and finally, in some protocols a purification step either precedes or follows PCR enrichment before sequencing.



**Figure 1.1: Transposon-insertion sequencing protocols.** An illustration of the workflow typical of transposon-insertion sequencing protocols. Transposons are represented by pink lines, sequencing adaptors by blue, genomic DNA by black, and PCR primers by green. Mutants are generated through either *in vivo* or *in vitro* transposition and subsequent selection for antibiotic resistance. These mutants are pooled, and optionally competed in test conditions, then genomic DNA is extracted and fragmented by restriction digest or physical shearing. Sequencing adaptors are ligated, some protocols then perform a step to purify fragments containing transposon insertions, and PCR with transposon- and adapter-specific primers is used to specifically enrich for transposon-containing fragments. The fragments are then sequenced and mapped back to a reference genome to uniquely identify insertion sites with nucleotide-resolution. Dashed boxes indicate steps which differ between protocols.

### 1.2.1 Transposon mutagenesis

Most studies have used either Tn5 or Mariner transposon derivatives. Tn5 originated as a bacterial transposon which has been adapted for laboratory use. Large-scale studies have shown that Tn5, while not showing any strong preference for regional GC-content, does have a weak preference for a particular insertion motif (Shevchenko et al., 2002; Adey et al., 2010; Green et al., 2012). Transposon-insertion sequencing studies performed with Tn5 transposons in *S. enterica* serovars have reported a slight bias towards AT-rich sequence regions (Langridge et al., 2009; Barquist, Langridge, et al., 2013). However, this preference does not appear to be a major obstacle to analysis given the extremely high insertion densities obtained with this transposon (Langridge et al., 2009; Christen et al., 2011; Barquist, Langridge, et al., 2013) (see Table 1.1). Additionally, Tn5 has been shown to be active in a wide range of bacterial species, though the number of transformants obtained can vary significantly depending on the transformation efficiency of the host.

Mariner *Himar1* transposons on the other hand originate from eukaryotic hosts and have an absolute requirement for TA bases at their integration site (Lampe, Grant, and Robertson, 1998; Rubin et al., 1999), with no other known bias besides a possible preference for bent DNA (Lampe, Grant, and Robertson, 1998). This can be a disadvantage in that it limits the number of potential insertion sites, particularly in GC-rich sequence. However, this specificity can also be used in the prediction of gene essentiality in near-saturated libraries: as every potential integration site is known and the probability of integration at any particular site can be assumed to be roughly equal, it is straight-forward to calculate the probability that any particular region lacks insertions by chance. *Himar1* transposition can also be conducted *in vitro* in the absence of any host factors (Lampe, Churchill, and Robertson, 1996), and inserted transposons can then be transferred to the genomes of naturally transformable bacteria through homologous recombination (Johnsborg, Eldholm, and Havarstein, 2007). This can be advantageous when working with naturally transformable bacteria with poor electroporation efficiency (Gawronski et al., 2009; Opijken, Bodi, and Camilli, 2009). It is worth noting that Tn5 is also capable of transposition *in vitro* (Goryshin and Reznikoff, 1998), and could potentially be used to increase insertion density and hence the resolution of the assay, particularly in GC-rich genomic regions.

### 1.2.2 Pool construction

Once mutants have been constructed, they are plated on an appropriate selective media for the transposon chosen, and colonies are counted, picked, and pooled. A disadvantage of this is that the mutants must be recreated for follow up or validation studies. Goodman et al. introduced a clever way around this in the INSeq protocol: by individually archiving mutants, then sequencing combinatorial mutant pools it is possible to uniquely characterize  $2n$  insertion mutants by sequencing only  $n$  pools (Goodman, McNulty, et al., 2009). Each mutant is labelled with a unique binary string that indicates which pools it has been added to. These binary strings can then be reconstructed for each insertion observed in these pools by recording their presence or absence in sequencing data, providing a unique pattern relating insertions to archived mutants. The authors control false identifications due to errors in sequencing by requiring that each binary label have a minimum edit distance to every other label, allowing for a robust association of labels with insertions despite sometimes noisy sequencing data. As a proof of concept, the authors were able to identify over 7,000 *Bacteroides thetaiotaomicron* mutants from only 24 sequenced pools. This effectively uses methods for the generation of random transposon pools to rapidly generate defined mutant arrays, though it is heavily dependent on liquid-handling robotics.

### 1.2.3 Enrichment of transposon-insertion junctions

Once pools have been constructed they are grown in either selective or permissive conditions, depending on the experiment, and then genomic DNA is extracted. Fragmentation proceeds either through restriction digestion in the case of transposons modified to contain appropriate sites (Goodman, McNulty, et al., 2009; Opijnen, Bodi, and Camilli, 2009; Gallagher, Shendure, and Manoil, 2011) or via physical shearing (Langridge et al., 2009; Gawronski et al., 2009), then sequencing adapters are ligated to the resulting fragments. PCR is performed on these fragments using a transposon-specific primer and a sequencing adapter-specific primer to enrich for fragments spanning the transposon-genomic DNA junction.

Some protocols purify fragments containing transposon insertions using biotinylated primers (Gallagher, Shendure, and Manoil, 2011; Goodman, Wu, and Gordon, 2011) or PAGE (Goodman, McNulty, et al., 2009) before and/or after PCR enrichment. The

purification step from the Tn-seq Circle protocol is particularly unusual in that restriction digested fragments containing transposon sequence are circularized before being treated with an exonuclease that digests all fragments without transposon insertions, theoretically completely eliminating background (Gallagher, Shendure, and Manoil, 2011). Given the success of protocols that do not include a purification step and the lack of systematic comparisons, it is currently unclear whether including one provides any major advantages.

#### **1.2.4 Sequencing**

The protocol steps described so far are broadly similar to those used in microarray-based studies of transposon mutant pools. The major advancement that has driven transposon-insertion sequencing has been the recent development of second generation DNA sequencing technologies. For 30 years, DNA sequencing was dominated by dideoxynucleotide, or Sanger, sequencing, first described by Sanger, Nicklen, and Coulson (1977). Sanger sequencing requires a clonal population of template DNA molecules, to which a primer and a full complement of four deoxynucleotides (dNTPs) and a single species of dideoxynucleotide (ddNTP) are added. DNA polymerase is then used to perform rounds of DNA extension, with ddNTPs stochastically terminating the reaction, before the resulting fragments are denatured and separated with gel electrophoresis. By running four such reactions with each species of ddNTP, the sequence of the template molecule can be determined by reading off bands on the gel. A number of advancements progressively improved the throughput and decreased the cost of Sanger sequencing, including the substitution of capillary electrophoresis for gel electrophoresis and the use of fluorescently labelled ddNTP (fluorescent dye-terminator sequencing) enabling sequencing in a single reaction. However, even with these advances the throughput of Sanger sequencing remained in the range of kilobases of sequence per hour, and costs remained high due to requirements for template cloning and inherent limitations in the technology (Morozova and Marra, 2008).

The development of second generation sequencing technologies in the early-mid 2000's broke these barriers to the adoption of sequencing as a routine experimental technique. These technologies include Roche 454 pyrosequencing, Illumina/Solexa reversible terminator sequencing, and ABI SOLiD parallel sequencing by ligation. While in principle any of these technologies could be applicable to transposon-insertion sequencing, all studies

to date have used Solexa sequencing. Solexa sequencing is similar in principle to Sanger sequencing, with two major innovations: the ability to generate arrayed clonal clusters of template molecules on a glass flow cell (described by Fedurco et al. (2006)) allowing for hundreds of thousands of simultaneous sequencing reactions, and the adoption of reversible dye terminator chemistry (described by Bentley et al. (2008)) which allows for fluorescently labelled terminators to be rapidly stripped of their fluorophore, their termination reversed, and extension continued. By monitoring successive rounds of these hundreds of thousands of parallel sequencing reactions with a CCD camera, the sequence of a large population of template molecules can be determined quickly and simultaneously, leading to current throughputs of megabases to gigabases of sequence per hour. As each resulting read corresponds to a single template molecule, this technology is ideally suited to monitoring populations of transposon mutants, providing an accurate digital count of insertion prevalence.

### 1.3 Reproducibility, accuracy, and concordance with previous methods

A number of studies have looked at the reproducibility of transposon-insertion sequencing. Multiple studies using different protocol variations have repeatedly shown extremely high reproducibility in the number of insertions per gene (correlations of 90%) in replicates of the same library grown and sequenced independently (Goodman, McNulty, et al., 2009; Opijnen, Bodi, and Camilli, 2009; Gallagher, Shendure, and Manoil, 2011), and good reproducibility (correlations between 70-90%) in independently constructed unsaturated libraries (Opijnen, Bodi, and Camilli, 2009; Opijnen and Camilli, 2012). Opijnen and Camilli (2012) compared traditional 1 X 1 competition experiments between wild-type and mutant *Streptococcus pneumoniae* to results obtained by transposon-insertion sequencing and showed that there was no significant difference in results over a range of tested conditions. The accuracy of transposon-insertion sequencing in determining library composition has also been assessed. Zhang et al. (2012) constructed a library of identified transposon-insertion mutants in known relative quantities, and then were able to recover the relative mutant prevalence with transposon-insertion sequencing. Additionally, by estimating the number of PCR templates prior to enrichment, this study showed that

there is a high correlation between enrichment input and sequencing output.

Two studies have evaluated concordance between results obtained with transposon-insertion sequencing and microarray monitoring of transposon insertions in order to demonstrate the enhanced accuracy and dynamic range of sequencing over previous methods. In the first, 19 libraries of 95 enterohemorrhagic *Escherichia coli* (EHEC) transposon mutants that had previously been screened in cattle using signature-tagged mutagenesis (STM) were pooled and re-evaluated using the TraDIS protocol (Eckert et al., 2011). The original STM study had identified 13 insertions in 11 genes attenuating intestinal colonization in a type III secretion system located in the locus of enterocyte effacement (LEE) (Dziva et al., 2004). By applying sequencing to the same samples, an additional 41 mutations in the LEE were identified, spanning a total of 21 genes. Additional loci outside the LEE which have been previously implicated in intestinal colonization but had not been detected by STM were also reported by TraDIS.

The second study re-evaluated genes required for optimal growth determined by TraSH in *Mycobacterium tuberculosis* (Sassetti, Boyd, and Rubin, 2003; Griffin et al., 2011). The greater dynamic range of sequencing as compared to microarrays allowed easier discrimination between insertions that were truly unviable and those that were only significantly underrepresented. The authors estimate that genes called as required by sequencing in their study are at least 100-fold underrepresented in the pool. In comparison, the threshold in the previous microarray experiment reported genes that had log probe ratios at least 5-fold lower than average between transposon-flanking DNA hybridization and whole genomic DNA hybridization. Additionally, the nucleotide-resolution of insertion sequencing allowed the authors to identify genes which had required regions, likely corresponding to required protein domains (Zhang et al., 2012), but which tolerated insertions in other regions. Altogether the authors increase the set of genes predicted to be required for growth in laboratory conditions in *M. tuberculosis* by more than 25% (from 614 to 774).

## 1.4 Identifying gene requirements

The earliest application of transposon-insertion sequencing, and indeed the earliest genome-wide experimental study of gene essentiality, was to determine the minimal set of genes necessary for the survival of *Mycoplasma* (Hutchison et al., 1999). This

essential genome is of great interest in synthetic and systems biology where it is seen as a foundation for engineering cell metabolism as described previously, and also in infection biology and medicine where it is seen as a promising target for therapies. However, it is important to remember that essentiality is always relative to growth conditions: a biosynthetic gene that is non-essential in a growth medium supplying a particular nutrient may become essential in a medium that lacks it. Traditionally, gene essentiality has been determined in clonal populations (Baba et al., 2006; Jacobs et al., 2003; Glass et al., 2006); since the high-throughput transposon sequencing protocols described here necessarily contain a short period of competitive growth before DNA extraction, many of these studies prefer to refer to the required genome for the particular conditions under evaluation.

Because of this short period of competitive growth, and because many otherwise required genes tolerate insertions in their terminus (Goodman, McNulty, et al., 2009; Griffin et al., 2011; Zomer et al., 2012) or outside essential domains (Zhang et al., 2012) the determination of required genomic regions is not completely straight-forward and a number of approaches have been taken to counter this. These include only calling genes completely lacking insertions as required (Opijken, Bodi, and Camilli, 2009), or determining a cut-off based on the empirical or theoretical distribution of gene-wise insertion densities (Langridge et al., 2009; Barquist, Langridge, et al., 2013; Griffin et al., 2011; Zomer et al., 2012). Additionally, windowed methods have been developed which can be used to identify essential regions in the absence of gene annotation (Zhang et al., 2012; DeJesus et al., 2013), and have had success in identifying required protein domains, promoter regions, and non-coding RNAs (ncRNAs). The organisms that have been evaluated for gene requirements under standard laboratory conditions are summarized in Table 1.1. In agreement with previous studies (Baba et al., 2006; Jacobs et al., 2003), many required genes identified by transposon-insertion sequencing are involved in fundamental biological processes such as cell division, DNA replication, transcription and translation (Langridge et al., 2009; Goodman, McNulty, et al., 2009; Barquist, Langridge, et al., 2013; Griffin et al., 2011), and many of these requirements appear to be conserved between genera and classes (Barquist, Langridge, et al., 2013; Christen et al., 2011).

However, a recent study defining required gene sets in *Salmonella* serovars (described in detail in the next chapter) has found that phage repressors, necessary for maintaining the lysogenic state of the prophage, are also required (Barquist, Langridge, et al., 2013),

even though mobile genetic elements such as phage are usually considered part of the accessory genome. This study also highlights the need for temperance when interpreting the results of high-throughput assays of gene requirements. For example, many genes in *Salmonella* Pathogenicity Island 2 (SPI-2) did not exhibit transposon-insertions, despite clear evidence from directed knockouts showing that these genes are non-essential for viability or growth. Under laboratory conditions, SPI-2 is silenced by the nucleoid-forming protein H-NS (Lucchini, Rowley, et al., 2006; Navarre et al., 2006), which acts by oligomerizing along silenced regions of DNA blocking RNA polymerase access. A previous study has shown that transposon insertion cold spots can be caused by competition between high-density proteins and transposases for DNA (Manna et al., 2007). This suggests that H-NS may be restricting transposase access to DNA, though this has not previously been observed in transposon-insertion sequencing data, and will require additional work to confirm.

## 1.5 Determining conditional gene requirements

One of the most valuable applications of the transposon-insertion sequencing method is the ability to identify genes important in a condition of interest, by comparing differences in the numbers of sequencing reads from input (control) mutant pools to output (test) pools that have been subject to passaging in a certain growth condition. Insertion counts are compared from cells in the input pool and those after passage, thereby identifying genes that either enhance or detract from survival and/or growth in the given condition, defined by decreased or increased insertion frequency, respectively. I describe a pipeline for analyzing such experiments in chapter **XX TraDIS PIPELINE CHAPTER HERE**. A further application of this method involves comparing insertions between biologically linked conditions, such as cellular stresses or different stages of a murine infection, to gain insight into complex systems (Opijken and Camilli, 2012).

So far, transposon-insertion sequencing has been used to investigate a number of interesting biologically relevant conditions: bile tolerance in *S. Typhi* (Langridge et al., 2009) and *S. Typhimurium* (Khatiwara et al., 2012), bacteriophage infection of *S. Typhi* (Pickard, Kingsley, et al., 2013), antibiotic resistance in *P. aeruginosa* (Gallagher, Shendure, and Manoil, 2011), cholesterol utilization in *M. tuberculosis* (Griffin et al., 2011) and survival in number of stress and nutrient conditions in *S. pneumoniae* (Opijken and

Camilli, 2012). Transposon-insertion sequencing of populations passed through murine models have been used to assess genes required for *H. influenzae* infection (Gawronski et al., 2009). A further extension of the method examined double mutant libraries, that is transposon mutant libraries generated in a defined deletion background, to tease apart complex networks of regulatory genes (Opijken, Bodi, and Camilli, 2009).

Two studies in particular illustrate the power of using transposon-insertion sequencing to identify conditionally required genes. In the first, Goodman, McNulty, et al. (2009) set out to determine the genes necessary for the establishment of the commensal *B. thetaiotaomicron* in a murine model. First, the growth requirements of transposon mutant populations in the cecum of germ-free mice was assessed, and genes required for growth in monoassociation with the host were found to be enriched in functions such as energy production and amino acid metabolism. By further comparing monoassociated transposon mutant libraries with those grown in the presence of three defined communities of human gut-associated bacteria, the authors identified a locus up-regulated by low levels of vitamin B12 that is only required in the absence of other bacteria capable of synthesizing B12. This showed that the gene requirements of any particular bacterium in the gut are at least partially dependent on the metabolic capabilities of the entire community and emphasizes the importance of testing *in vivo* conditions to complement *in vitro* study.

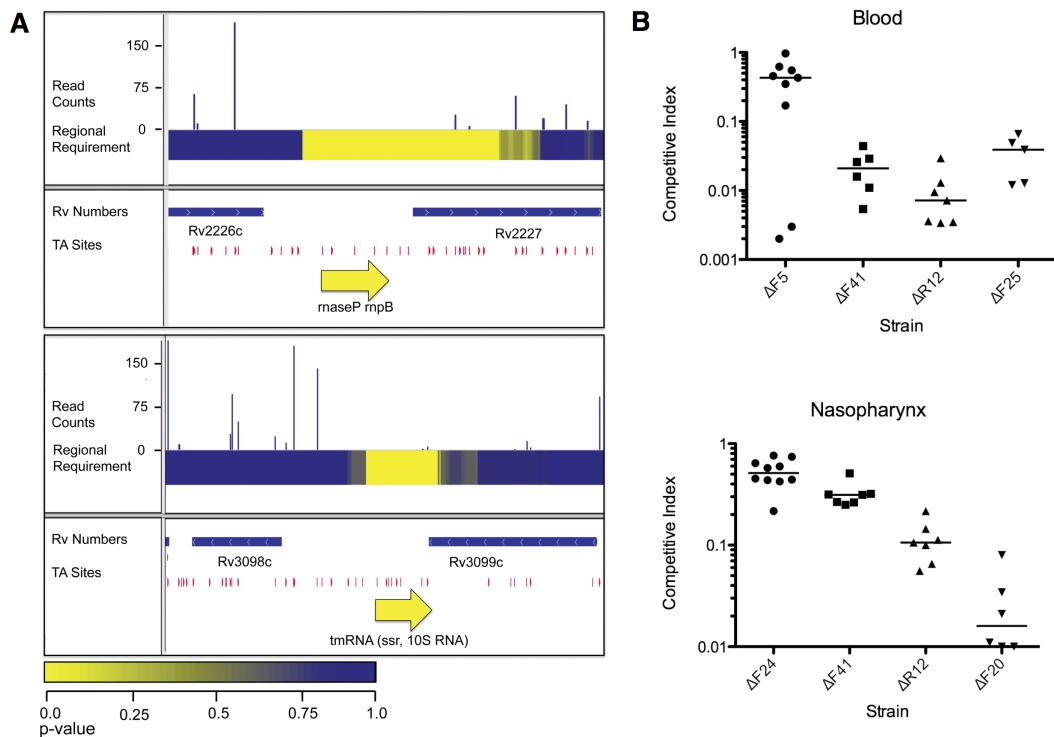
The second study, conducted by Opijken and Camilli (2012), aimed to map the genetic networks involved in a range of cellular stress responses in *S. pneumoniae*. Seventeen *in vitro* conditions were tested, including: pH, nutrient limitation, temperature, antibiotic, heavy metal, and hydrogen peroxide stress. Approximately 6% of disrupted genes resulted in increased fitness in some condition, suggesting that some genes are maintained despite being detrimental to the organism under particular conditions. These would be interesting candidates for further functional and evolutionary study, as the maintenance of these genes is presumably highly dependent on the conditions the bacteria faces, and may have implications for our understanding of e.g. gene loss in the process of bacterial host adaptation (Toft and Andersson, 2010). Two additional *in vivo* experiments were performed in a murine model, where cells were recovered from the lung and nasopharynx. Combining this data, over 1,800 genotype-phenotype genetic interactions were identified. These interactions were mapped and pathways identified. Between the two *in vivo* niches, certain stress responses pathways were markedly different. For example, temperature

stress produced a distinct response in the lung, compared to the nasopharynx, which is perhaps to be expected as temperature varies greatly between these two sites. By further examining sub-pathways required in the two different niches and comparing them to *in vitro* requirements, the authors were able to draw conclusions regarding the condition *S. pneumoniae* faces when establishing an infection. This comprehensive mapping of genotype-phenotype relationships will serve as an important atlas for further studies of the factors affecting *S. pneumoniae* carriage and virulence.

## 1.6 Monitoring ncRNA contributions to fitness

To date, four studies (including one described in detail in the next chapter) have used transposon-insertion sequencing to examine the contribution of non-coding RNAs (ncRNAs) and other non-coding regions to organismal fitness (see Table 1.1). Two of these examined requirements for non-coding regions in the relatively under-explored bacterial species *Caulobacter crescentus* (Christen et al., 2011) and *M. tuberculosis* (Zhang et al., 2012). Both utilized analytical techniques that allowed for the identification of putative required regions in the absence of genome annotation. Twenty-seven small RNAs (sRNAs) had previously been detected in *C. crescentus* (Landt et al., 2008); 6 were found to be depleted in transposon insertions indicating an important role in basic cellular processes. Additionally, the well-characterized ncRNAs tmRNA and RNase P, as well as 29 non-redundant tRNAs were found to be required. An additional 90 unannotated non-disruptable regions were identified throughout the genome, implying an abundance of unexplored functional non-coding sequence.

While the non-coding transcripts of *M. tuberculosis* have been explored more thoroughly than those of *C. crescentus*, most remain functionally uncharacterized, though there are hints that some of these may be involved in pathogenicity (Arnvig and Young, 2012). Using a Mariner transposon-based assay and a windowed statistical analysis that accounted for the distribution of potential TA integration sites, 35 intergenic regions were identified as putatively required in the *M. tuberculosis* genome (Zhang et al., 2012). In common with the *C. crescentus* study, the RNA component of RNase P, required for the maturation of tRNAs, and tmRNA, involved in the freeing of stalled ribosomes, were identified as required (Figure 1.2 A) together with 10 non-redundant tRNAs and potential promoter regions. However, due to the lower overall insertion density and lack



of TA sites in some GC-rich regions, there were some regions that could not be assayed and the resolution was limited to 250 bases.

A particularly exciting study has been conducted in *S. pneumoniae* TIGR4 combining RNA-seq with transposon-insertion sequencing (Mann et al., 2012). To identify sRNA loci the authors first sequenced size-select RNA from wild type TIGR4 and three two-component system knockouts, identifying 89 putative sRNAs, 56 of which were novel. Fifteen of these candidates, selected on the basis of high expression and low predicted folding free energy, were assayed for their ability to establish invasive disease in a murine model. Of these 8 sRNA deletions showed a significant attenuation of disease. To more broadly establish the roles of sRNAs in infecting particular organs, transposon insertion libraries were administered directly to the nasopharynx, lungs, or blood of mice, and bacteria were harvested following disease progression. Twenty-six, 28, and 18 sRNAs were found to attenuate infection in the nasopharynx, lung and blood respectively. These results were then validated with targeted deletions of 11 sRNAs (Figure 1.2 B). In addition to establishing the role of sRNAs in *S. pneumoniae* virulence, this study illustrated the power of combining RNA-seq and transposon-insertion sequencing to rapidly assign phenotypes to non-coding sequences.

## 1.7 Limitations

In this chapter, I have largely focused on the potential of transposon insertion sequencing. However, this technology does have a number of important limitations. As discussed previously, requirements for particular nucleotides at insertion sites, such as the TA required by Mariner transposons, or preference for certain sequence composition, such as the AT bias exhibited by Tn5, can limit the density of observed insertions in certain genomic regions. This may impact any down-stream analysis, and can potentially bias results, particularly the determination of gene requirements. Even if this bias has been accounted for, transposon-insertion screens will always over-predict gene requirements in comparison to targeted deletion libraries as discussed previously. However, this over-prediction can be controlled either through careful consideration of known insertion biases as in many Mariner-based studies, or by high insertion densities, such as those achieved in several Tn5-based studies (Table 1.1). Once the library has been created, only regions that have accumulated insertions in the conditions of library creation will be able to be

assayed for fitness effects in further conditions. This means that regions that lead to slow growth phenotypes when disrupted in standard laboratory conditions may be difficult to assay in other conditions. Additionally, the dynamic range of fitness effects detected will depend on the complexity of the input library(s). The absence of insertions may be a particular problem for assaying small genomic elements, such as sRNAs or short ORFs. Finally, the validation of hypotheses derived from transposon-insertion sequencing will require the construction of targeted deletions, as individual mutants cannot be recovered from pools unless specialized protocols have been followed during library construction (as in Goodman, McNulty, et al., 2009).

## **1.8 The future of transposon-insertion sequencing**

Transposon-insertion sequencing is a robust and powerful technique for the rapid connection of genotype to phenotype in a wide range of bacterial species. Already, a number of studies have demonstrated the effectiveness of this method and the results have been far-reaching: enhancing our understanding of basic gene functions, establishing requirements for colonization and infection, mapping complex metabolic pathways, and exploring non-coding genomic dark matter. Due to the range of potential applications of transposon-insertion sequencing, along with the decreasing cost and growing accessibility of next-generation sequencing, I believe that this method will become increasingly common in the near future.

A number of bacterial species have already been subjected to transposon-insertion sequencing (Table 1.1). Microarray-based approaches to monitoring transposon mutant libraries have even been applied to eukaryotic systems (Ross-Macdonald et al., 1999), and similarly transposon-insertion sequencing can potentially be applied to any system where the creation of large-scale transposon mutant libraries is technologically feasible. Recently the Genomic Encyclopedia of Bacteria and Archea (GEBA) (Wu et al., 2009) has been expanding our knowledge of bacterial diversity through targeted genomic sequencing of underexplored branches of the tree of life. Applying transposon-insertion sequencing in a comparative manner across the bacterial phylogeny will provide an unprecedented view of the determinants for survival in diverse environments - the next chapter describes a study taking the first steps toward this eventual goal (Barquist, Langridge, et al., 2013). While most transposon-insertion sequencing studies to date have focused on pathogenic bacteria,

these techniques could also have applications in energy production, bioremediation, and synthetic biology.

The combination of transposon-insertion sequencing with other high-throughput and computational methods is already proving to be fertile ground for enhancing our understanding of bacterial systems. For instance, by using transposon-insertion sequencing in a collection of relatively simple conditions combined with a computational pathway analysis, Opijnen and Camilli (2012) were able to provide a holistic understanding of the genetic subsystems involved in a complex process such as *S. pneumoniae* pathogenesis. In the future, methods to assay phenotype in a high-throughput manner (Bochner, 2009; Nichols et al., 2011; See also chapter **XX BIOLOG CHAPTER HERE**) may be combined with transposon-insertion sequencing to provide exhaustive simple genotype-phenotype associations with which to understand complex processes in a systems biology framework.

# Chapter 2

## A comparison of dense transposon insertion libraries in the *Salmonella* serovars Typhi and Typhimurium

*This chapter is a modified version of the previously published article “A comparison of dense transposon insertion libraries in the Salmonella serovars Typhi and Typhimurium” (Barquist, Langridge, et al., 2013). This work is a result of collaboration with Gemma C. Langridge (Pathogen Genomics, Wellcome Trust Sanger Institute), who constructed the Salmonella Typhimurium transposon mutant library and contributed to a draft manuscript. In particular, portions of the analyses in sections 2.3.1-3 have their origins in Langridge (2010), though have been significantly elaborated on here.*

### 2.1 Introduction

*Salmonella enterica* subspecies *enterica* serovars Typhi (*S. Typhi*<sup>1</sup>) and Typhimurium (*S. Typhimurium*) are important, closely related, human pathogens with very different lifestyles. In this chapter, I describe a study comparing dense transposon insertion libraries created in these two serovars. The results of this study demonstrate that orthologous genes can have dramatically different effects on the fitness of recently diverged organisms

---

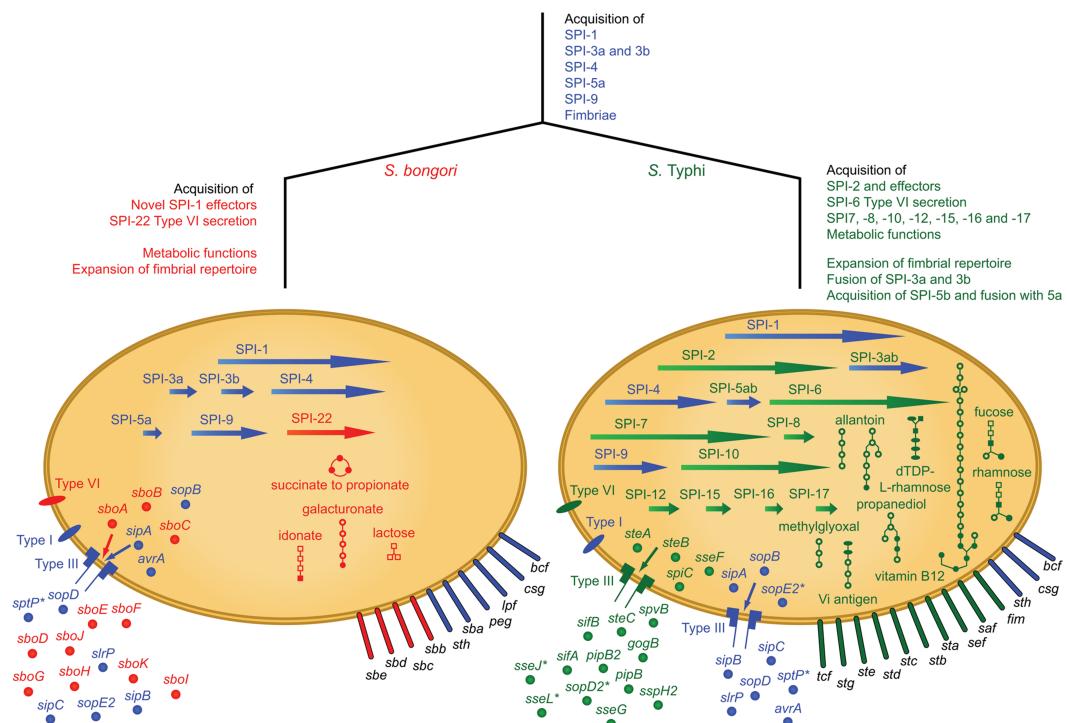
<sup>1</sup>Note that the complicated *Salmonella* taxonomy and nomenclature make abbreviation difficult (and at times contentious). Here I have adopted the practice of referring to individual serovars as *S. Serovar* once they have been introduced, following the advice of Brenner et al. (2000).

in rich media. These differences in fitness effects are indicative of changes in the network architecture of the cell which may partially underlie the dramatically different diseases caused by each organism and their different host ranges. Additionally, *S. Typhimurium* has served as a model organism for the discovery and functional characterization of ncRNAs. Comparing ncRNA requirements between it and a closely related serovar provides a glimpse of the functional evolution of non-coding regulatory networks.

### 2.1.1 The genus *Salmonella*

*Salmonella* is a Gram-negative,  $\gamma$ -proteobacterial genus within the order Enterobacterales, consisting of two species: *Salmonella enterica* and *Salmonella bongori*, though a contested third species, *Salmonella subterranea*, has recently been proposed (Shelobolina et al., 2004). Based on phylogenetic analyses of 16S and conserved amino acid sequences, *Salmonella* is most closely related to the generaes *Escherichia*, *Shigella*, and *Citrobacter* (Paradis et al., 2005; Pham et al., 2007; Wu et al., 2009). Molecular clock analyses suggest that *Salmonella* and *Escherichia* shared a common ancestor between 100 and 160 million years ago (Ochman and Wilson, 1987; Doolittle, Feng, et al., 1996), though complete genetic isolation of the two genera may have taken 70 million years (Retchless and Lawrence, 2007). During the time since their divergence *Escherichia* has become established as a mammalian gut commensal, though multiple independent origins of the *Shigella* and other pathogenic phenotypes within the genus show that a disease phenotype can be developed fairly easily through the horizontal acquisition of virulence determinants and the silencing of anti-virulence loci (Kaper, Nataro, and Moseley, 2004; Prosseda et al., 2012). Despite sharing the majority of their genomes with *Escherichia* and having broadly similar metabolic capabilities (AbuOun et al., 2009), the salmonellae exist primarily as pathogens, though are possibly commensal in some reptiles (Mermin et al., 2004; Bauwens et al., 2006).

The difference in dominant phenotype between *Escherichia* and *Salmonella* appears to be largely due to the acquisition of virulence determinants which opened new niches to ancestral salmonellae (see figure 2.1). Many of the virulence determinants characteristic of the salmonellae are encoded on large genomic islands with sizes between  $\sim$ 6 and 140 kilobases, termed *Salmonella* Pathogenicity Islands (SPIs) (Hensel, 2004). These islands encode a diverse array of pathogenicity-related functions including secretion systems,



**Figure 2.1: Genomic acquisitions in the evolution of the salmonellae.** Traits shared by the common ancestor are depicted in blue; those unique to *S. bongori* are shown in red and those unique to *S. enterica* subspecies *enterica* serovar Typhi in green. Arrows, *Salmonella* Pathogenicity Islands (SPIs); extended ovals, fimbriae; circles, effectors; small ovals and needle complexes, secretion systems. Metabolic pathways: lines, enzymatic reactions; open squares, carbohydrates; ovals, pyrimidines; open circles, other substrates; filled shapes, phosphorylated. Novel effectors acquired by *S. bongori* are secreted by the type III secretion system encoded on SPI-1. SPI-3a and 3b carry the same genes in both organisms but are fused into one island in *S. Typhi*. SPI-5a also carries the same genes in both organisms, but a further 3 kb (termed SPI-5b) has fused to SPI-5a in *S. Typhi*. \*indicates a pseudogene. Reproduced from Fookes et al. (2011) under a Creative Commons Attribution License (CCAL).

toxins, antibiotic resistances, and lipopolysaccharide (LPS) and capsular modifications. In particular, the acquisition of SPI-1, encoding a type 3 secretion system (T3SS), and various fimbriae by the ancestral *Salmonella* likely enabled invasion of cells in the intestinal epithelium and escape from competition with other members of the gut microbiota (Bäumler, 1997). *S. bongori* appears to have only acquired a single additional SPI since its divergence from *S. enterica* and likely retains a lifestyle more similar to the ancestral *Salmonella*, though there is evidence for additional adaptation to its niche in the reptilian gut (Fookes et al., 2011).

*S. enterica* meanwhile has diversified into 6 distinct subspecies: *enterica*, *salamae*, *arizonae*, *diarizonae*, *houtenae*, and *indica*. These subspecies are further divided into over 2000 serovars based on the cell-surface O, flagellar H, and capsular Vi antigens (Grimont and Weill, 2007). The acquisition of SPI-2, involved in survival inside macrophages and an enabling factor for systemic infection (Kuhle and Hensel, 2004; Abrahams and Hensel, 2006), by the ancestral *S. enterica* is thought to have been a driving force in this diversification (Bäumler, 1997). Subspecies besides *enterica* are thought to be primarily restricted to cold-blooded animals (Bäumler, 1997), though sporadic reports of zoonotic disease show these subspecies are capable of transiently colonizing the mammalian gut under certain conditions (Mermin et al., 2004; Hilbert et al., 2012). However, here I will be primarily concerned with the subspecies *enterica* and its adaptation to the mammalian, and more specifically human, host.

### 2.1.2 Host adaptation and restriction

Bacterial adaptation to a pathogenic lifestyle is a complex process involving both the acquisition of virulence factors and gene loss through both passive decay and positive selection (Pallen and Wren, 2007; Grosseda et al., 2012). In the previous section I discussed how the acquisition of SPI-1 and -2, among other factors, have enabled *S. enterica* subspecies *enterica* to establish a niche in the mammalian gut. Access to this new niche has enabled serovars of subspecies *enterica* to explore a range of pathogenic modalities. The most common form of disease caused by *enterica* serovars is a self-limiting gastroenteritis, exemplified by the serovars Typhimurium and Enteriditis (Santos, Raffatellu, et al., 2009). These serovars can infect a wide range of mammals and birds, but are only capable of causing serious disease in the very young (Bäumler, Tsolis, et al.,

1998), and are generally thought to exhibit a phenotype similar to the ancestral *enterica*.

A number of subspecies *enterica* serovars have adapted to causing invasive disease in specific organisms. These include Typhi and Paratyphi in humans, Dublin in cattle, Gallinarum in chickens, Abortusovis in sheep, Choleraesuis in pigs, and Abortusequi in horses. These adaptations appear to be the result of the acquisition of host-specific virulence factors (Bäumler, Tsolis, et al., 1998). Interestingly, those serovars associated with the most severe forms of disease appear to be most highly restricted in terms of host range. This appears to be the result of three processes: positive selection against anti-virulence loci (Pallen and Wren, 2007; Prosseda et al., 2012), and two more passive processes termed “use it or lose it” and “use it, but lose it anyway” by Moran (2004).

Selection against anti-virulence loci presumably occurs during host-adaptation, and generally involved the loss of loci that provoke an antigenic response or interfere with the infective process. Once a bacterium has escaped competition in the gut microbiota and gained access to a rich intracellular niche through horizontal acquisitions, the “use it or lose it” principle leads to the loss of metabolic pathways no longer required in this environment presumably due to the lifting of selective pressure for their maintenance. The “use it, but lose it anyway” principle is a consequence of the severe bottleneck imposed by adaptation to a particular host, which will often drastically reduce the effective population size of the bacterium. This can cause fixation of inactivating mutations in potentially beneficial genes simply as an accident of the adaptive process. Together these processes may eventually prevent the bacterium from living independently of its host; particularly extreme examples are *Mycobacterium leprae* with its thousands of inactivated pseudogenes (Cole et al., 2001), *Mycoplasma* species with their highly reduced genomes (Fraser et al., 2001), and most strikingly the endosymbiont-derived mitochondria and plastid organelles (Sagan, 1967; Andersson et al., 1998). While no *Salmonella* serovars appear to have been subject to this degree of genome degradation, it is not unusual for as much as 7% of the protein-coding genes of host-restricted serovars to be inactivated (Parkhill et al., 2001; Thomson, Clayton, et al., 2008; Holt, Teo, et al., 2009; McClelland, Sanderson, Clifton, et al., 2004).

The serovars of *S. enterica* subspecies *enterica* exhibit a spectrum of pathogenic lifestyles, from low-pathogenicity and wide host range to high-pathogenicity and narrow host range. Recent studies examining host adaptation of Typhimurium strains to immunocompromised populations (Feasey et al., 2012; Okoro et al., 2012) demonstrate

that the process of host-adaptation is both on-going and highly relevant to human health. In this study, we have used transposon-insertion sequencing to examine two recently diverged (circa 50,000 years ago (Kidgell et al., 2002)) serovars at extreme ends of this pathogenicity spectrum: Typhi and Typhimurium.

### 2.1.3 Serovars Typhi and Typhimurium

*Salmonella enterica* subspecies *enterica* serovars Typhi (*S. Typhi*) and Typhimurium (*S. Typhimurium*) are important human pathogens with distinctly different lifestyles. *S. Typhi* is host-restricted to humans and causes typhoid fever. This potentially fatal systemic illness affects at least 21 million people annually, primarily in developing countries (Crump, Luby, and Mintz, 2004; Bhutta and Threlfall, 2009; Kothari, Pruthi, and Chugh, 2008), and is capable of colonizing the gall bladder creating asymptomatic carriers; such individuals are the primary source of this human restricted infection, exemplified by the case of “Typhoid Mary” (Soper, 1939). Mary Mallon was an Irish-American cook in New York City at the turn of the twentieth century, and an (at least initially) unwitting carrier of Typhi. A series of typhoid outbreaks were traced to her by city public health authorities. She was offered removal of her gall bladder, which she refused, and was ordered to refrain from working as a cook following release from three years of quarantine. After a number of additional outbreaks – including several deaths – were traced to Mary, who had continued working as a cook under a pseudonym, she was involuntarily quarantined on North Brother Island in the East River for 23 years until her death.

*S. Typhimurium*, conversely, is a generalist, causing relatively mild disease in a wide range of mammals and birds in addition to being a leading cause of foodborne gastroenteritis in human populations. Control of *S. Typhimurium* infection in livestock destined for the human food chain is of great economic importance, particularly in swine and cattle (CDC, 2009; Majowicz et al., 2010). Additionally, *S. Typhimurium* causes an invasive disease in mice, which has been used extensively as a model for pathogenicity in general and human typhoid fever specifically (Santos, Zhang, et al., 2001).

Despite this long history of investigation, the genomic factors that contribute to these differences in lifestyle remain unclear. Over 85% of predicted coding sequences are conserved between the two serovars in sequenced genomes of multiple strains (McClelland, Sanderson, Spieth, et al., 2001; Parkhill et al., 2001; Holt, Parkhill, et al., 2008; Deng,

Liou, et al., 2003). The horizontal acquisition of both plasmids and pathogenicity islands during the evolution of the salmonellae is believed to have impacted upon their disease potential. A 100kb plasmid, encoding the *Salmonella* plasmid virulence (SPV) genes, is found in some *S. Typhimurium* strains and contributes significantly towards systemic infection in animal models (Gulig and Curtiss, 1987; Gulig, Danbara, et al., 1993). *S. Typhi* is known to have harbored IncHI1 plasmids conferring antibiotic resistance since the 1970s (Phan et al., 2009), and there is evidence that these strains present a higher bacterial load in the blood during human infection (Wain et al., 1998). Similar plasmids have been isolated from *S. Typhimurium* (Datta, 1962; Holt, Thomson, et al., 2007; Cain and Hall, 2012). *Salmonella* pathogenicity islands 1 and 2 are common to all *Salmonella enterica* subspecies, and are required for invasion of epithelial cells (reviewed in Darwin and Miller (1999)) and survival inside macrophages respectively (Ochman and Groisman, 1996; Shea et al., 1996; Kuhle and Hensel, 2004; Abrahams and Hensel, 2006). *S. Typhi* additionally incorporates SPI-7 and SPI-10, which contain the Vi surface antigen and a number of other putative virulence factors (Pickard, Wain, et al., 2003; Seth-Smith, 2008; Townsend et al., 2001).

Acquisition of virulence determinants is not the sole explanation for the differing disease phenotypes displayed in humans by *S. Typhimurium* and *S. Typhi*; genome degradation is an important feature of the *S. Typhi* genome, in common with other host-restricted serovars such as *S. Paratyphi A* (humans) and *S. Gallinarum* (chickens). In each of these serovars, pseudogenes account for 4-7% of the genome (Parkhill et al., 2001; Thomson, Clayton, et al., 2008; Holt, Teo, et al., 2009; McClelland, Sanderson, Clifton, et al., 2004). Loss of function has occurred in a number of *S. Typhi* genes that have been shown to encode intestinal colonisation and persistence determinants in *S. Typhimurium* (Kingsley et al., 2003). Numerous sugar transport and degradation pathways have also been interrupted (Parkhill et al., 2001), but remain intact in *S. Typhimurium*, potentially underlying the restricted host niche occupied by *S. Typhi*.

In addition to its history as a model organism for pathogenicity, *S. Typhimurium* has recently served as a model organism for the elucidation of non-coding RNA (ncRNA) function (Vogel, 2009a). These include cis-acting switches, such as RNA-based temperature and magnesium ion sensors (Waldminghaus et al., 2007; Cromie et al., 2006), together with a host of predicted metabolite-sensing riboswitches. Additionally, a large number of trans-acting small RNAs (sRNAs) have been identified within the *S. Typhimurium*

genome (Kröger et al., 2012), some with known roles in virulence (Hebrard et al., 2012). These sRNAs generally control a regulon of mRNA transcripts through an antisense binding mechanism mediated by the protein Hfq in response to stress. The functions of these molecules have generally been explored in either *S. Typhimurium* or *E. coli*, and it is unknown how stable these functions and regulons are over evolutionary time (Richter and Backofen, 2012).

Transposon mutagenesis has previously been used to assess the requirement of particular genes for cellular viability. The advent of next-generation sequencing has allowed simultaneous identification of all transposon insertion sites within libraries of up to 1 million independent mutants (reviewed in Barquist, Boinett, and Cain (2013); see also the previous chapter), enabling us to answer the basic question of which genes are required for *in vitro* growth with extremely fine resolution. By using transposon mutant libraries of this density, which in *S. Typhi* represents on average  $> 80$  unique insertions per gene (Langridge et al., 2009), shorter regions of the genome can be interrogated, including ncRNAs (Christen et al., 2011). In addition, once these libraries exist, they can be screened through various selective conditions to further reveal which functions are required for growth/survival.

Illumina-based transposon directed insertion-site sequencing (TraDIS (Langridge et al., 2009)) with large mutant libraries of both *S. Typhimurium* and *S. Typhi* was used to investigate whether these salmonellae require the same protein-coding and non-coding RNA (ncRNA) gene sets for competitive growth under laboratory conditions, and whether there are differences which reflect intrinsic differences in the pathogenic niches these bacteria inhabit.

## 2.2 Materials and Methods

Gemma Langridge created the *S. Typhimurium* library described here, and performed all the laboratory experiments described here. Duy Phan and Keith Turner created the *S. Typhi* library. Duy Phan and Gemma Langridge performed the read mapping.

### 2.2.1 Strains

*S. Typhimurium* strain SL3261 contains a deletion relative to the parent strain, SL1344, was used to generate the large transposon mutant library. The 2166bp deletion ranges from 153bp within *aroA* (normally 1284bp) to the last 42bp of *cmk*, forming two pseudogenes and deleting the intervening gene SL0916 completely. For comparison, our previously generated *S. Typhi* Ty2 transposon library (Langridge et al., 2009) was used.

### 2.2.2 Annotation

For *S. Typhimurium* strain SL3261, I used feature annotations drawn from the SL1344 genome (EMBL-Bank accession FQ312003.1), ignoring the deleted *aroA*, *ycaL*, and *cmk* genes. I re-analyzed the *S. Typhi* Ty2 transposon library with features drawn from an updated genome annotation (EMBL-Bank accession AE014613.1.) I supplemented the EMBL-Bank annotations with non-coding RNA annotations drawn from Rfam 10.1 (Burge et al., 2013), Sittka et al. (2008), Chinni et al. (2010), Raghavan, Groisman, and Ochman (2011), and Kröger et al. (2012). Selected protein-coding gene annotations were supplemented using the HMMER webserver (Finn, Clements, and Eddy, 2011) and Pfam (Punta et al., 2012).

### 2.2.3 Creation of *S. Typhimurium* transposon mutant library

*S. Typhimurium* was mutagenized using a Tn5-derived transposon as described previously (Langridge et al., 2009; a detailed protocol is available in Langridge, 2010). Briefly, the transposon was combined with the EZ-Tn5 transposase (Epicenter, Madison, USA) and electroporated into *S. Typhimurium*. Transformants were selected by plating on LB agar containing 15 µg/mL kanamycin and harvested directly from the plates following overnight incubation. A typical electroporation experiment generated a batch of between 50,000 and 150,000 individual mutants. 10 batches were pooled together to create a mutant library comprising approximately 930,000 transposon mutants.

### 2.2.4 DNA manipulations and sequencing

Genomic DNA was extracted from the library pool samples using tip-100g columns and the genomic DNA buffer set from Qiagen (Crawley, UK). DNA was prepared for nucleotide

sequencing as described previously (Langridge et al., 2009). Prior to sequencing, a 22 cycle PCR was performed as previously described (Langridge et al., 2009). Sequencing took place on a single end Illumina flowcell using an Illumina GAI<sup>I</sup> sequencer, for 36 cycles of sequencing, using a custom sequencing primer and 2x Hybridization Buffer (Langridge et al., 2009). The custom primer was designed such that the first 10 bp of each read was transposon sequence.

### 2.2.5 Sequence analysis

The Illumina FASTQ sequence files were parsed for 100% identity to the 5' 10bp of the transposon (TAAGAGACAG). Sequence reads which matched were stripped of the transposon tag and subsequently mapped to the *S. Typhimurium* SL1344 or *S. Typhi* Ty2 chromosomes using MAQ version maq-0.6.8 (Li, Ruan, and Durbin, 2008). Approximately 12 million sequence reads were generated from the sequencing run which used two lanes on the Illumina flowcell. Precise insertion sites were determined using the output from the Maq mapview command, which gives the first nucleotide position to which each read mapped. The number and frequency of insertions mapping to each nucleotide in the appropriate genome was then determined.

### 2.2.6 Statistical analysis of required genes

The number of insertion sites for any gene is dependent upon its length, so the values were made comparable by dividing the number of insertion sites by the gene length, giving an “insertion index” for each gene. As before (Langridge et al., 2009) the distribution of insertion indices was bimodal, corresponding to the required (mode at 0) and non-required distributions (See Figure 2.2). I fitted gamma distributions for the two modes using the R MASS library (<http://www.r-project.org>).  $\text{Log}_2$ -likelihood ratios (LLR) were calculated between the required and non-required distributions and I called a gene required if it had an LLR of less than -2, indicating it was at least 4 times more likely according to the required model than the non-required model. “Non-required” genes were assigned for an LLR of greater than 2. Genes falling between the two thresholds were considered “ambiguous” for the purpose of this analysis. This procedure lead to genes being called as required in *S. Typhimurium* when their insertion index was less than 0.020, or 1 insertion

in every 50 bases, and ambiguous between 0.020 and 0.027. The equivalent cut-offs for the *S. Typhi* library are 0.0147 and 0.0186, respectively.



**Figure 2.2: The distribution of gene-wise insertion indexes in *S. Typhi*.** Bars report the density of genes with insertion indexes within each range, black lines show gamma distributions fitted to the required (left, mode at 0) and non-required (right) peaks, and red lines report associated LLR-based cut-offs for calling gene ambiguity (left) and requirement (right). The distribution of insertion indexes in *S. Typhimurium* is similar, though with a wider separation between the required and non-required peaks due to the higher insertion density attained.

I calculated a p-value for the observed number of insertion sites per gene using a

Poisson approximation with rate  $R = \frac{N}{G}$  where  $N$  is the number of unique insert sites (549,086) and  $G$  is the number of bases in the genome (4,878,012). The p-value for at least  $X$  consecutive bases without an insert site is  $e^{(-RX)}$ , giving a 5% cut-off at 27 bp and a 1% cut-off at 41 bp.

For every gene  $g$  with  $n_{g,A}$  reads observed in *S. Typhi* and  $n_{g,B}$  reads observed in *S. Typhimurium*, I calculated the  $\log_2$  fold change ratio  $S_{g,A,B} = \log_2(\frac{n_{g,A}+100}{n_{g,B}+100})$ . The correction of 100 reads smoothes out the high scores for genes with very low numbers of observed reads. I fitted a normal distribution to the mode +/- 2 sample standard deviations of the distribution of  $S_{A,B}$ , and calculated p-values for each gene according to the fit. I considered genes with a p-value of 0.05 or less under the fitted normal distribution to be uniquely required by one serovar.

## 2.3 Results and Discussion

### 2.3.1 TraDIS assay of every *Salmonella* Typhimurium protein-coding gene

Approximately 930,000 mutants of *S. Typhimurium* were generated using a Tn5-derived transposon. 549,086 unique insertion sites were recovered from the mutant library using short-read sequencing with transposon-specific primers. This is a substantially higher density than the 371,775 insertions recovered from *S. Typhi* previously (Langridge et al., 2009). The *S. Typhimurium* library contains an average of one insertion every 9bp, or over 100 unique inserts per gene (figure 2.3). The large number of unique insertion sites allowed every gene to be assayed; assuming random insertion across the genome, a region of 41bp without an insertion was statistically significant ( $P < 0.01$ ). As previously noted in *S. Typhi*, the distribution of length-normalized insertions per gene is bimodal (see figure 2.2), with one mode at 0. Genes falling in to the distribution around this mode are interpreted as being required for competitive growth within a mixed population under laboratory conditions (hereafter “required”). Of these, 57 contained no insertions whatsoever and were mostly involved in core cellular processes (see table 2.1).

There was a bias in the frequency of transposon insertion towards the origin of replication. This likely occurred as the bacteria were in exponential growth phase immediately prior to transformation with the transposon. In this phase of growth,



**Figure 2.3: Genome-wide transposon mutagenesis of *S. Typhimurium*.** A) Circular plot showing gene content, distribution of required genes, and insertion density along the *S. Typhimurium* chromosome. The outer scale is marked in megabases. Circular tracks range from 1 (outer track) to 6 (inner track). Track 1, all forward-strand genes (color-coded according to function: dark blue, pathogenicity/adaptation; black, energy metabolism; red, information transfer; dark green, membranes/surface structures; cyan, degradation of macromolecules; purple, degradation of small molecules; yellow, central/intermediary metabolism; light blue, regulators; pink, phage/IS elements; orange, conserved hypothetical; pale green, unknown function; brown, pseudogenes.); track 2, all reverse-strand genes (color-coded as for forward-strand genes); track 3, *S. Typhimurium* required genes (red); track 4, 56 genes required by *S. Typhimurium* but not by *S. Typhi* (dark blue, see also table 1); track 5, transposon insertion density; track 6, GC bias ( $\frac{G-C}{G+C}$ ), khaki indicates values  $>1$ ; purple  $<1$ .

multiple replication forks would have been initiated, meaning genes closer to the origin were in greater copy number and hence more likely to be a target for insertion. There was a bias for transposon insertions in A+T rich regions, as was previously observed in the construction of an *S. Typhi* mutant library (Langridge et al., 2009). However, the insertion density achieved is sufficient to discriminate between required and non-required genes easily. As was first seen in *S. Typhi* (Langridge et al., 2009), there were transposon insertions into genes upstream of required genes in the same operon, suggesting that most insertions do not have polar effects leading to the inactivation of downstream genes.

Analysis of the *S. Typhimurium* mutant library allowed the identification of 353 coding sequences required for growth under laboratory conditions, and 4,112 non-required coding sequences (see [need to put supp data in appendix](#)). Sixty-five genes could not be assigned to either the required or non-required category. 60 of these genes, which I will refer to as “ambiguous”, had log-likelihood ratios (LLRs) between -2 and 2. The final 5 unassigned genes had lengths less than 60 bases, and they were removed from the analysis. All other genes contained enough insertions or were of sufficient length to generate credible LLR scores. Thus, every gene was assayed and I was able to draw conclusions for 98.7% of the coding genome in a single sequencing run (figure 2.3).

### 2.3.2 Cross-species comparison of genes required for growth

Gene essentiality has previously been assayed in *S. Typhimurium* using insertion-duplication mutagenesis. Knuth et al. (2004) estimated 490 genes are essential to growth in clonal populations, though 36 of these have subsequently been successfully deleted (Santiviago, Reynolds, et al., 2009). While TraDIS assays gene requirements after a brief period of competitive growth on rich media, I identify a smaller required set than Knuth et al. (2004) of approximately 350 genes in each serovar, closer to current estimates of approximately 300 essential genes in *E. coli* (Baba et al., 2006).

To demonstrate that TraDIS does identify genes known to have strong effects on growth, as well as to test our predictive power for determining gene essentiality, I compared our required gene sets in *S. Typhimurium* and *S. Typhi* to essential genes determined by systematic single-gene knockouts in the *Escherichia coli* K-12 Keio collection (Baba et al., 2006). I identified orthologous genes in the three data sets by best reciprocal FASTA hits exhibiting over 30% sequence identity over at least 80% of the amino acid sequence.



**Figure 2.4: Comparison of required genes.** Venn diagrams showing (A) the overlap of all genes (outer circles, light colors) and required genes (inner circles, dark colors) between *S. Typhimurium* and *S. Typhi* (excluding genes required in one serovar only which do not have significantly different read-counts). Black numbers refer to all genes, white numbers to required genes. (B) the overlap of all required genes between *S. Typhimurium* (blue), *S. Typhi* (green) and *E. coli* (purple). White numbers refer to genes with Keio essentiality scores  $\geq 0.5$ .

Required orthologous genes identified in this manner share a significantly higher average percent sequence identity with their *E. coli* counterparts than expected for a random set of orthologs, at  $\sim 94\%$  identity as compared to  $\sim 87\%$  for all orthologous genes. In 100,000 randomly chosen gene sets of the same size as our required set I did not find a single set where the average shared identity exceeded 90%, indicating that required genes identified by TraDIS are more highly conserved at the amino acid level than other orthologous protein coding sequences.

Baba et al. (2006) have defined an essentiality score for each gene in *E. coli* based on evidence from four experimental techniques for determining gene essentiality: targeted knock-outs using  $\lambda$ -red mediated homologous recombination, genetic footprinting (Gerdes et al., 2003; Tong et al., 2004), large-scale chromosomal deletions (Hashimoto et al., 2005), and transposon mutagenesis (Kang et al., 2004). Scores range from -4 to 3, with negative scores indicating evidence for non-essentiality and positive scores indicating evidence for essentiality. Comparing the overlap between essential gene sets in *E. coli*, *S. Typhi*, and *S. Typhimurium*, I found a set of 228 *E. coli* genes which have a Keio essentiality score of at least 0.5 (i.e. there is evidence for gene essentiality; See Figure 2.4.)

**Table 2.1: Core genome functions in *S. Typhimurium*.** Protein-coding genes providing fundamental biological functions in *S. Typhimurium*. Genes in bold are required in *S. Typhi* (log-likelihood ratio (LLR) between required and non-required models < -2; see Methods.) \* indicates genes ambiguous in *S. Typhimurium*, having a LLR between -2 and 2.

Biological Process	Sub-process	Required genes	Non-required genes
Cell division		<i>ftsALKQWYZ, minE, mukB, SL2391</i>	<i>ftsHJNX*</i> , <i>minCD</i> , <i>sdIA, cedA, sulA</i>
DNA replication	Polymerases I, II, and III	<i>dnaENQX, holAB</i>	<i>polAB, holCDE</i>
	Supercoiling	<i>gyrAB, parCE</i>	
	Primosome-associated	<i>dnaBCGT, priA, ssb</i>	<i>priB*C, rep</i>
Transcription	RNA polymerase	<i>rpoABC</i>	
	Sigma, elongation, anti- and termination factors	<i>nusBG, rpoDH, rho</i>	<i>nusA, rpoENS</i>
Translation	tRNA-synthetases	<i>alaS, argS, asnS, aspS, cysS, glnS, gltX, glyQS, hisS, ileS, leuS, lysS, metG, pheST, proS, serS, thrS, tyrS, valS</i>	<i>trpS, trpS2</i>
	Ribosome components	<i>rplBCDEFJKLMNPQRSTUUVWXYZ, rpmABCDHI, rpsABCDEFGHIJKLMNPQRST</i>	<i>rplAI, rpmEE2, rpmFHJJ2, rpsOR*U*V</i>
	Initiation, elongation, and peptide chain release factors	<i>fusA, infABC, prfAB, tsf, yrdC</i>	<i>efp, prfCH, selB, tuf</i>
<b>Biosynthetic pathways</b>			
Peptidoglycan		<i>murABCDEFGI</i>	<i>ddl, dlla</i>
Fatty acids		<i>accABCD, fabABDGHI</i>	

that have TraDIS-predicted required orthologs in both *S. Typhi* and *S. Typhimurium*, constituting ~85% of *E. coli* genes with evidence for essentiality indicating that gene requirements are largely conserved between these genera. Including orthologous genes that are only predicted to be essential by TraDIS in *S. Typhi* or *S. Typhimurium* raises this figure to nearly 93%. The majority of shared required genes between all three bacteria are responsible for fundamental cell processes, including cell division, transcription and translation. A number of key metabolic pathways are also represented, such as fatty acid and peptidoglycan biosynthesis (Table 2.1). A recent study in the  $\alpha$ -proteobacteria *Caulobacter crescentus* reported 210 shared essential genes with *E. coli*, despite *C. crescentus* sharing less than a third as many orthologous genes with *E. coli* as *Salmonella* serovars (Christen et al., 2011). This suggests the existence of a shared core of approximately 200 essential proteobacterial genes, with the comparatively rapid turnover of 150 to 250 non-core lineage-specific essential genes.

By making the simplistic assumption that gene essentiality should be conserved between *E. coli* and *Salmonella*, I can use the overlap of our predictions with the Keio essential genes to provide an estimate of our TraDIS libraries accuracy for predicting that a gene will be required in a clonal population. Of the 2632 orthologous *E. coli* genes which have a Keio essentiality score of less than -0.5 (i.e. there is evidence for gene

non-essentiality), only 33 are predicted to be required by TraDIS in both *Salmonella* serovars. *S. Typhi* contains the largest number of genes predicted by TraDIS to be required with *E. coli* orthologs with negative Keio essentiality scores. However, even if it is assumed these are all incorrect predictions of gene essentiality, this still gives a gene-wise false positive rate (FPR) of ~2.7% (81 out of 2981 orthologs) and a positive predictive value (PPV) of ~75% (247 with essentiality scores greater than or equal to 0.5 out of 328 predictions with some Keio essentiality score.) Under these same criteria the *S. Typhimurium* data set has a lower gene-wise FPR of ~1.6% (51 out of 3122 orthologs) and a higher PPV of ~82% (234 out of 285 predictions as before), as would be expected given the library's higher insertion density. In reality these FPRs and PPVs are only estimates; genes which are not essential in *E. coli* may become essential in the different genomic context of *Salmonella* serovars and vice versa, particularly in the case of *S. Typhi* where wide-spread pseudogene formation has eliminated potentially redundant pathways (Holt, Teo, et al., 2009; McClelland, Sanderson, Clifton, et al., 2004). Additionally, TraDIS will naturally over-predict essentiality in comparison to targeted knockouts, as the library creation protocol necessarily contains a short period of competitive growth between mutants during the recovery from electro-transformation and selection. As a consequence, genes which cause major growth defects, but not necessarily a complete lack of viability in clonal populations, may be reported as 'required.'

### 2.3.3 Serovar-specific genes required for growth

Many of the required genes present in only one serovar encoded phage repressors, for instance the cI proteins of Fels-2/SopE and ST35 (see tables 2.2 and 2.3). Repressors maintain the lysogenic state of prophage, preventing transcription of early lytic genes (Echols, 1971). Transposon insertions into these genes will relieve this repression and trigger the lytic cycle, resulting in cell death, and consequently mutants are not represented in the sequenced library. This again broadens the definition of 'required' genes; such repressors may not be required for cellular viability in the traditional sense, but once present in these particular genomes, their maintenance is required for continued viability, as long as the rest of the phage remains intact.

*S. Typhimurium* and *S. Typhi* both contain 8 apparent large phage-derived genomic regions (Thomson, Baker, et al., 2004; Kropinski et al., 2007). I was able to identify

**Table 2.2: Phage elements in *S. Typhimurium*.** Genomic coordinates determined from annotations in the EMBL annotation for FQ312003 and manual inspection. Repressor domains and architecture were determined using the HMMER webserver (Finn, Clements, and Eddy, 2011) and Pfam (Punta et al., 2012). Phage types were determined using repressor sequence similarity searches and information from Thomson, Baker, et al. (2004) and Kropinski et al. (2007).

Element name	Genomic coordinates	Repressor	Repressor domain(s)	Repressor domain architecture	Predicted active?	Phage type	Required cargo
Gifsy-2 SLP105	1054795 - 1100036	SL0950	HTH_3 (PF01381)		Yes	lambdaoid	N/A
N/A	1913364 - 1925490	N/A	N/A	N/A	No	remnant	SL1799
SLP203	2039803 - 2079890	SL1967	HTH_19 (PF12844) and Peptidase_S24 (PF00717)		Yes	P22-like	N/A
Gifsy-1 SLP272	2726717 - 2777229	SL2593	HTH_3 (PF01381)		Yes	lambdaoid	SL2549
SLP281	2815382 - 2825915	SL2633	2 Phage_CI_repr (PF07022)		Yes	degenerate P2-like	N/A
Fels-2 SLP285	2855616 - 2888522	SL2708	Phage_CI_repr (PF07022)		Yes	P2-like	SL2695
SLP289	2890073 - 2900377	IsrK RNA (RF01394)	N/A	N/A	No	P4-like	N/A
SLP443	4437731 - 4459844	N/A	N/A	N/A	No	remnant	SL4132

required repressors in all the intact lambdaoid, P2-like, and P22-like prophage in both genomes, including Gifsy-1, Gifsy-2, and Fels-2/SopE (see tables 2.2 and 2.3). With the exception of the SLP203 P22-like prophage in *S. Typhimurium*, all of these repressors lack the peptidase domain of the classical  $\lambda$  repressor gene cI. This implies that the default anti-repression mechanism of *Salmonella* prophage may be more similar to a trans-acting mechanism recently discovered in Gifsy phage (Lemire, Figueroa-Bossi, and Bossi, 2011) than to the  $\lambda$  repressor's RecA-induced self-cleavage mechanism. I was also able to confirm that most phage remnants and fusions contained no active repressors, with the exception of the SLP281 degenerate P2-like prophage in *S. Typhimurium*. This degenerate prophage contains both intact replication and integration genes, but appears to lack tail and head proteins, suggesting it may depend on another phage for production of viral particles. Both genomes also encode P4-like satellite prophage, which rely on ‘helper’ phage for lytic functions and utilize a complex antisense-RNA based regulation mechanism for decision pathways regarding cell fate (Briani et al., 2001) using structural homologs of the IsrK (Padalon-Brauch et al., 2008) and C4 ncRNAs (Forti

**Table 2.3: Phage elements in *S. Typhi*.** Genomic coordinates determined from Thomson, Baker, et al. (2004) and manual inspection. Repressor domains and architecture were determined using the HMMER webserver (Finn, Clements, and Eddy, 2011) and Pfam (Punta et al., 2012). Phage types were determined using repressor sequence similarity searches and information from Thomson, Baker, et al. (2004) and Kropinski et al. (2007).

Element name	Genomic coordinates	Repressor	Repressor domain(s)	Repressor domain architecture	Predicted active?	Phage type	Required cargo
ST15	1408790 - 1441377	N/A	N/A	N/A	No	Mu/P2 fusion	N/A
Gifsy-2	1929572 - 1972330	t1920	HTH_3 (PF01381)		Yes	lambdoid	N/A
ST2-27	2735054 - 2745321	IsrK RNA (RF01394)	N/A	N/A	Yes	P4-like	N/A
ST27	2745477 - 2768221	N/A	N/A	N/A	No	P2/iroA fusion	N/A
ST35	3500854 - 3536047	t3402	Phage_CI_repr (PF07022)		Yes	P2-like	t3415
SopE	4457346 - 4491316	t4337	Phage_CI_repr (PF07022)		Yes	P2-like	N/A
N/A	4519423 - 4519501	IsrK RNA (RF01394)	N/A	N/A	No	remnant	N/A
ST46	4666579 - 4677433	IsrK RNA (RF01394)	N/A	N/A	Yes	P4-like	N/A

et al., 2002), known as seqA and CI RNA in the P4 literature, respectively. While the mechanism of P4 lysogenic maintenance is not known, the IsrK-like ncRNAs of two potentially active P4-like prophage in *S. Typhi* are required under TraDIS. This sequence element has previously been shown to be essential for the establishment of the P4 lysogenic state (Sabbattini et al., 1995), and we predict based on our observations that it may be necessary for lysogenic maintenance as well. The fact that some lambdoid prophage in *S. Typhimurium* encode non-coding genes structurally similar to the IsrK-C4 immunity system of P4 raises the possibility that these systems may be acting as a defense mechanism of sorts, protecting the prophage from predatory satellite phage capable of co-opting its lytic genes.

In addition to repressors, 4 prophage cargo genes in *S. Typhimurium* and one in *S. Typhi* are required (See tables 2.2, 2.3, 2.4, and 2.6). The *S. Typhimurium* prophage cargo genes encode a PhoPQ regulated protein, a protein predicted to be involved in natural transformation, an endodeoxyribonuclease, and a hypothetical protein. The *S. Typhi* prophage cargo gene encodes a protein containing the DNA-binding HIRAN domain (Iyer, Babu, and Aravind, 2006), believed to be involved in the repair of damaged DNA. These warrant further investigation, as they are genes that have been recently

acquired and become necessary for survival in rich media.

To compare differences between requirements for orthologous genes in both serovars, I calculated log-fold read ratios to eliminate genes which were classified differently in *S. Typhi* and *S. Typhimurium* but did not have significantly different read densities (see Methods.) Even after this correction, 36 *S. Typhimurium* genes had a significantly lower frequency of transposon insertion compared to the equivalent genes in *S. Typhi* ( $P < 0.05$ ), including four encoding hypothetical proteins (table 2.4). This indicates that these gene products play a vital role in *S. Typhimurium* but not in *S. Typhi* when grown under laboratory conditions.

**Table 2.4: Genes uniquely required in *S. Typhimurium*.** Genes determined to be uniquely required in *S. Typhimurium*. SL, *S. Typhimurium*; Ty, *S. Typhi*; inserts refer to the number of unique insertion sites within a gene; reads refer to the number of sequence reads over all insertions sites within a gene. †, P-value (associated with log2 read ratio) < 0.05. ‡, sse is a pseudogene in *S. Typhi*. Shaded rows indicate genes shown to be H-NS repressed in Navarre et al. (2006)

Ty inserts	Ty reads	SL inserts	SL reads	SL ID	SL gene length	Ty ID	Ty gene length	Name	Function
-	-	18	123	SL0742	1269	-	-	-	putative cation transporter
-	-	9	80	SL0830	516	-	-	-	conserved hypothetical protein
-	-	4	21	SL0831	855	-	-	-	putative electron transfer flavoprotein (beta subunit)
-	-	0	0	SL0950	323	-	-	-	predicted bacteriophage protein, potential phage repressor Gifsy-2
-	-	11	75	SL1179	789	-	-	-	lipoprotein
-	-	3	18	SL1480	249	-	-	-	antitoxin Phd.YefM, type II toxin-antitoxin system
-	-	4	32	SL1527	264	-	-	-	putative inner membrane protein
-	-	1	3	SL1560	717	-	-	-	putative membrane protein
-	-	7	50	SL1601	859	-	-	-	putative transcriptional regulator (pseudogene)
-	-	4	36	SL1799	201	-	-	-	bacteriophage encoded pagK (phoPQ-activated protein)
-	-	5	22	SL1830A	434	-	-	-	conserved hypothetical protein (pseudo-gene)
-	-	3	27	SL1967	677	-	-	-	predicted bacteriophage protein, potential phage repressor SLP203
-	-	1	15	SL2045A	63	-	-	-	short ORF
-	-	17	107	SL2066	900	-	-	-	CDP-abequose synthase
-	-	3	34	SL2549	209	-	-	-	endodeoxyribonuclease
-	-	4	149	SL2593	449	-	-	-	putative DNA-binding protein, potential phage repressor Gifsy-1 SLP272
-	-	3	7	SL2633	846	-	-	-	putative repressor protein, SLP281
-	-	2	21	SL2695	978	-	-	-	putative competence protein
-	-	5	39	SL4132	291	-	-	-	hypothetical protein
-	-	5	45	SL4354A	303	-	-	-	conserved hypothetical protein
36	474	5	26	SL0032	441	t0033	306	-	putative transcriptional regulator
71	349	11	48	SL0623	642	t2232	576	lpB	lipopeptide ligase B
151	3546	10	64	SL0702	897	t2156	894	-	putative glycosyl transferase
194	3007	9	61	SL0703	1134	t2155	1134	-	galactosyltransferase
231	3499	15	67	SL0706	1779	t2152	1780	-	putative glycosyltransferase, cell wall biogenesis
84	1041	2	4	SL0707	834	t2151	834	-	putative glycosyltransferase, cell wall biogenesis
49	367	14	70	SL0722	1569	t2136	1569	cydA	cytochrome d ubiquinol oxidase subunit I
74	1613	5	22	SL1069	693	t1789	693	-	putative secreted protein
20	199	1	1	SL1203	150	t1146	156	-	hypothetical protein

20	290	1	5	SLI1264	315	t1209	315	-
84	384	6	26	SLI1327	402	t1261	384	spiC
66	769	5	35	SLI1331	270	t1265	327	putative membrane protein putative pathogenicity island 2 secreted effector protein
36	307	2	5	SLI1341	228	t1275	228	T3SS chaperone
47	407	1	3	SLI1342	249	t1276	249	putative pathogenicity island protein
144	3197	5	14	SLI1343	750	t1277	750	putative pathogenicity island protein
63	847	5	26	SLI1354	267	t1288	267	putative type III secretion protein
73	762	4	44	SLI1355	780	t1289	780	putative type III secretion protein
30	226	12	48	SLI1386	693	t1322	693	rnfE/ydgQ
265	3337	29	165	SLI1473	1557	t1463	1557	Electron transport complex protein rnfE
85	765	6	35	SLI1532	951	t1511	951	PhoPQ-activated protein
22	156	16	174	SLI1561	1227	t1334†	141	putative virulence effector protein
119	1639	10	44	SLI1563	762	t1536	762	Salmonella translocated effector protein (SseJ)
107	2440	5	44	SLI1564	648	t1537	648	putative periplasmic amino acid-binding protein
183	1646	20	118	SLI1628	1355	t1612	1364	putative ABC amino acid transporter permease
23	177	1	5	SLI1659	183	t1640	183	hypothetical protein
78	617	16	104	SLI1684	1014	t1664	1014	conserved hypothetical protein
37	277	4	25	SLI1785	396	t1022	396	putative regulatory protein
166	2823	9	27	SLI1793	915	t1016	915	conserved hypothetical protein
28	311	3	22	SLI1794	159	t1015	159	inner membrane protein
23	155	1	4	SLI1823	972	t0988	972	putative inner membrane protein
60	402	11	58	SLI2064	1002	t0786	1002	lipid A acyltransferase
87	524	7	59	SLI2065	1293	t0785	1299	putative glycosyl transferase
66	559	13	74	SLI2069	774	t0780	774	putative O-antigen transporter
41	204	5	14	SLI3828	1830	t3658	1830	glucose-1-phosphate cytidylyltransferase
27	288	5	23	SLI4250	288	t4220	288	glucosamine-fructose-6-phosphate aminotransferase
148	2633	16	89	SLI4251	876	t4221	876	putative GerE family regulatory protein
								araC family regulatory protein

Present in *S. Typhi* but required  
only in *S. Typhimurium*



**Figure 2.5: Comparison of cell surface operon structure and requirements.** Diagram illustrating cell surface operons with different requirement patterns in *S. Typhimurium* and *S. Typhi*. The top figure is of an uncharacterized operon putatively involved in cell wall biogenesis, while the bottom figure shows a portion of the *rfb* operon involved in O-antigen biosynthesis. Plots along the top and bottom of each figure show insertions in *S. Typhimurium* and *S. Typhi*, respectively, with read depth on the y-axis with a maximum cut-off of 100 reads. Genes in blue are required in *S. Typhimurium*, genes in white are pseudogenes; others are in grey. Grey rectangles represent BLAST hits between orthologous genes, with percent nucleotide identity colored on the scale to the right of each figure.

A major difference between the two serovars is in the requirement for genes involved in cell wall biosynthesis (see figure 2.5). A set of four genes (SL0702, SL0703, SL0706, and SL0707) in an operonic structure putatively involved in cell wall biogenesis is required in *S. Typhimurium* but not in *S. Typhi*. The protein encoded by SL0706 is a pseudogene in *S. Typhi* (Ty2 unique ID: t2152) due to a 1bp deletion at codon 62 that causes a frameshift. This operon contains an additional two pseudogenes in *S. Typhi* (t2154 and t2150), as well as a single different pseudogene (SL0700) in *S. Typhimurium*, indicating that this difference in gene requirements reflects the evolutionary adaptation of these serovars to their respective niches. Similarly, four genes (*rfbV*, *rfbX*, *rfbJ* and *rfbF*) within an O-antigen biosynthetic operon are required by *S. Typhimurium* but not *S. Typhi*. There appears to have been a shuffling of O-antigen biosynthetic genes since the divergence between the two serovars, and *rfbJ*, encoding a CDP-abequose synthase, has been lost from *S. Typhi* altogether. These broader requirements for cell wall-associated biosynthetic and transporter genes suggest that surface structure biogenesis is of greater

importance in *S. Typhimurium*.

There were seven genes from the shared pathogenicity island SPI-2 that appear to contain few or no transposon insertions only in *S. Typhimurium* under laboratory conditions. These genes (*spiC*, *sseA*, and *ssaHIJT*) are thought to encode components of the SPI-2 type III secretion system apparatus (T3SS) (Kuhle and Hensel, 2004). In addition, the effector genes *sseJ* and *sifB*, whose products are secreted through the SPI-2-encoded T3SS (Miao and Miller, 2000; Freeman, Ohl, and Miller, 2003), also fell into the ‘required’ category in *S. Typhimurium* alone. All of these genes display high A+T nucleotide sequence and have been previously shown (in *S. Typhimurium*) to be strongly bound by the nucleoid associated protein H-NS, encoded by *hns* (Lucchini, Rowley, et al., 2006; Navarre et al., 2006). Therefore, rather than being ‘required’, it is instead possible that access for the transposon was sufficiently restricted that very few insertions occurred at these sites. In further support of this hypothesis, a comparison of the binding pattern of H-NS detected in studies using *S. Typhimurium* LT2 with the TraDIS results from the SPI-2 locus indicated that high regions of H-NS enrichment correlated well with both the *ssa* genes described here and with *sseJ* (see figure 2.6). An earlier study also suggests that high-density DNA binding proteins can block Mu, Tn5, and Tn10 insertion (Manna et al., 2007); however, a genome-wide study of the effects of H-NS binding on transposition would be necessary to confirm this effect.



**Figure 2.6: H-NS enrichment across the SPI-2 locus.** Based on data from Lucchini, Rowley, et al. (2006) where a 2 fold enrichment of H-NS-bound DNA over a total genomic DNA control in a ChIP-on-chip experiment was taken to indicate regions of H-NS binding in *S. Typhimurium* strain LT2. Assuming these binding patterns are similar in the *S. Typhimurium* strain tested in this study, H-NS binding may have affected transposon access to genes in the SPI-2 locus.

**Table 2.5: Candidate required genes affected by H-NS binding in *S. Typhimurium*.** Genes identified by comparison with data from Navarre et al. (2006). Fold change values report the results of a ChIP-on-chip experiment, and indicate genes strongly bound by H-NS.

Gene	SL ID	STM ID	Fold change	Function
-	SL0830	STM0854	-16.2	conserved hypothetical protein
-	SL0831	STM0855	-33.8	putative putative electron transfer flavoprotein (beta subunit)
-	SL1069	STM1131	-13.5	putative putative secreted protein
spiC	SL1327	STM1393	-19.1	putative pathogenicity island 2 secreted effector protein
sseA	SL1331	STM1397	-46	Type three secretion system chaperone
ssaH	SL1341	STM1407	-8.8	Type three secretion system apparatus
ssaI	SL1342	STM1408	-32.4	putative putative pathogenicity island protein
ssaJ	SL1343	STM1409	-53.7	putative putative pathogenicity island lipoprotein
ssaS	SL1354	STM1420	-15.5	putative putative type III secretion protein
ssaT	SL1355	STM1421	-33.9	putative putative type III secretion protein
pqaA	SL1473	STM1544	-5.5	PhoPQ-activated protein
sifB	SL1532	STM1602	-66.8	putative putative virulence effector protein
-	SL1560	STM1630	-9.8	putative putative membrane protein
sseJ	SL1561	STM1631	-48.6	<i>Salmonella</i> translocated effector protein (SseJ)
-	SL1563	STM1633	-91.9	putative putative periplasmic amino acid-binding protein
-	SL1564	STM1634	-22.5	putative putative ABC amino acid transporter permease
-	SL1628	STM1698	-101.4	hypothetical protein
-	SL1659	STM1728	-17.3	cytochrome b561 (cytochrome b-561)
-	SL1785	STM1856	-12.1	conserved hypothetical protein
pagO	SL1793	STM1862	-11.9	inner membrane protein (PagO)
-	SL1794	STM1864	-22.9	putative inner membrane protein

Indeed, the generation of null *S. Typhimurium* mutants in *sseJ* and *sifB*, as well as many others generated at the SPI-2 locus suggest that these genes are not truly a requirement for growth in this serovar (Freeman, Ohl, and Miller, 2003; Hensel, Shea, Baumler, et al., 1997; Hensel, Shea, Waterman, et al., 1998; Ohlson et al., 2005). While this is a reminder that the interpretation of gene requirement needs to be made with care, the effect of H-NS upon transposon insertion is not genome-wide. If this were the case, there would be an under-representation of transposon mutants in high A+T regions (known for H-NS binding), which is not what was observed. In total, only 21 required genes fall into the ‘*hns*-repressed’ category described in Navarre et al. (2006)(see table 2.5); the remainder (almost 400) contained sufficient transposon insertions to conclude they were non-required. In addition, all SPI-1 genes that encode another T3SS and are of high A+T content were also found to be non-required. This phenomenon was not observed in *S. Typhi*, possibly because the strain used harbors the pHCM1 plasmid, which encodes the H-NS-like protein *sfh* and has been shown to affect H-NS binding (Doyle et al., 2007; Dillon et al., 2010).

Twenty-two *S. Typhi* genes had a significantly lower frequency of transposon insertion compared to orthologs in *S. Typhimurium* ( $P < 0.05$ ), indicating that they are required only in *S. Typhi* for growth under laboratory conditions (table 2.6), including the *fepBDGC* operon. This indicates a requirement for ferric (Fe(III)) rather than ferrous

(Fe(II)) iron. This can be explained by the presence of Fe(III) in the bloodstream, where *S. Typhi* can be found during typhoid fever (Wain et al., 1998). These genes function to recover the ferric chelator enterobactin from the periplasm, acting with a number of proteins known to aid the passage of this siderophore through the outer membrane (Rabsch et al., 1999). It has long been noted that *aroA* mutants of *S. Typhi*, deficient in their ability to synthesize enterobactin, exhibit severe growth defects on complex media, while similar mutants of *S. Typhimurium* grow normally under the same conditions (Edwards and Stocker, 1988), though the mechanism has not been clear. These results suggest that this difference in growth of *aroA* mutants is caused by a requirement for iron uptake through the *fep* system in *S. Typhi*. During host adaptation, *S. Typhi* has accumulated pseudogenes in many iron transport and response systems (McClelland, Sanderson, Clifton, et al., 2004), presumably because they are not necessary for survival in the niche *S. Typhi* occupies in the human host, which may have led to this dependence on *fep* genes. In contrast, *S. Typhimurium* generally causes intestinal rather than systemic infection and is able to utilize a wider range of iron sources, including Fe(II), a soluble form of iron present under anaerobic conditions such as those found in the intestine (Tsolis et al., 1996).

**Table 2.6: Genes uniquely required in *S. Typhi*.** Genes determined to be uniquely required in *S. Typhi*. SL, *S. Typhimurium*; Ty, *S. Typhi*; inserts refer to the number of unique insertion sites within a gene; reads refer to the number of sequence reads over all insertions sites within a gene. †, P-value (associated with  $\log_2$  read ratio)  $< 0.05$ . \*, the assignment of recA as a required gene has been described previously (Langridge et al., 2009), but briefly is believed to be due to the presence of the priC pseudogene in *Typhi*.

SL inserts	SL reads	Ty inserts	Ty reads	Ty ID	Ty gene length	SL ID	SL gene length	Name	Function
<i>S. Typhimurium</i> in <i>S. Typhi</i> orthologs									
199	1792	18	59	t1332	132	-	-	-	pseudogene
45	498	3	22	t1920	405	-	-	-	possible repressor protein, prophage 10/Gifsy-2
120	589	11	32	t0123	165	-	-	-	conserved hypothetical protein
-	-	2	12	t3157	228	-	-	-	spurious ORF annotation overlapping the RnasP/M1 RNA
-	-	2	12	t3166	-	-	-	-	repressor protein, cs 73 prophage
-	-	6	196	t3402	570	-	-	cI	HIRAN-domain family gene, potential DNA repair
-	-	4	58	t3415	741	-	-	-	hypothetical secreted protein
-	-	1	6	t4531	150	-	-	-	survival protein SurA precursor
123	982	2	25	t0224	1287	SL0093	1287	surA	glutamate-1-semialdehyde 2,1-
67	452	1	14	t0270	459	SL0119	459	yabB/mraZ	aminomutase
140	760	0	0	t0587	1281	SL0203	1281	hemL	Zinc metallopeptidase
113	641	15	42	t2140	1353	SL0224	1353	yaeL/rseP	RNA polymerase sigma-E factor
116	753	13	36	t2177	576	SL2604	576	rpoE	ribonucleoside-diphosphate reductase 1
80	542	9	15	t2276	2286	SL2246	2286	mrnA	alpha chain
93	591	2	2	t2277	2802	SL0718	2802	sucA	2-oxoglutarate dehydrogenase E1 component
64	508	5	6	t2278	795	SL0578	795	fepC	phosphoglucomutase
201	1129	12	116	t2410	2355	SL0444	1641	fepD	ferric enterobactin transport protein
95	518	8	20	t2730	1062	SL2809	1008	FepD	FepD ferric enterobactin transport protein
135	719	16	39	t2996	1992	SL3052	1947	FepG	FepG ferric enterobactin transport protein
76	358	3	9	t3120	1434	SL3173	1434	lon	Lon protease
213	1976	6	50	t3265	1071	SL3321	1062	recA*	recA protein
43	448	3	10	t3326	606	SL3925	1047	tktA	transketolase
124	571	17	36	t3384	2025	SL3872	2025	rfaE	ADP-heptose synthase
175	1208	6	21	t3621	2787	SL3947	2787	degS	serine protease
117	797	9	13	t3808	1047	SL3677	1047	yigP	conserved hypothetical protein
176	1628	14	59	t4153	1080	SL4183	1080	rep	ATP-dependent DNA helicase
140	1127	10	38	t4411	951	SL4294	951	polA	DNA polymerase I
								waaF	ADP-heptose-LPS heptosyltransferase II
								alr	alanine racemase
								miaA	tRNA delta-2-isopentenylpyrophosphate transferase

### 2.3.4 TraDIS provides resolution sufficient to evaluate ncRNA contributions to fitness

Under a Poisson approximation to the transposon insertion process, a region of 41 (in *S. Typhimurium*) or 60 bases (in *S. Typhi*) has only a 1% probability of not containing an insertion. NcRNAs tend to be considerably shorter than their protein-coding counterparts, but this gives us sufficient resolution to assay most of the non-coding complement of the *Salmonella* genome. As a proof of principle, I performed an analysis of the best-understood class of small ncRNAs, the tRNAs. Francis Crick hypothesized that a single tRNA could recognize more than one codon through wobble recognition (Crick, 1966), where a non-canonical G-U base pair is formed between the first (wobble) position of the anticodon and the third nucleotide in the codon. As a result, some codons are covered by multiple tRNAs, while others are covered non-redundantly by a single tRNA. I expect that singleton wobble-capable tRNAs, that is wobble tRNAs which recognize a codon uniquely, will be required. In addition, I inferred the requirement for other tRNAs through the non-redundant coverage of their codons and used this to benchmark our ability to use TraDIS to reliably interrogate short genomic intervals.

The *S. Typhi* and *S. Typhimurium* genomes encode 78 and 85 (plus one pseudogene) tRNAs respectively with 40 anticodons, as identified by tRNAscan-SE (Lowe and Eddy, 1997). In *S. Typhi*, 10 out of 11 singleton wobble tRNAs are predicted to be required or ambiguous, compared to 16 tRNAs below the ambiguous LLR cut-off overall (significant enrichment at the 0.05 level, two-tailed Fishers exact test p-value: 6.4e-08.) Similarly in *S. Typhimurium*, 9 of 11 singleton wobble tRNAs are required or ambiguous compared to 15 required or ambiguous tRNAs overall, again showing a significant enrichment of required tRNAs in this subset (Fishers exact test p-value: 5.2e-07.) The one singleton wobble tRNA which is consistently not required in both serovars is the tRNA-Pro(GGG), which occurs within a 4-member codon family. It has previously been shown in *S. Typhimurium* that tRNA-Pro(UGG) can read all four proline codons in vivo due to a cmo5U34 modification to the anticodon, obviating the need for a functional tRNA-Pro(GGG) (Näsvall, Chen, and Björk, 2004) and making this tRNA non-required. The other non-required singleton wobble tRNA in *S. Typhimurium*, tRNA-Leu(GAG), is similarly a member of a 4-member codon family. I predict tRNA-Leu(TAG) may also be capable of recognizing all 4 leucine codons in this serovar; such a leucine “four-way

wobble” has been previously inferred in at least one other bacterial species (Osawa et al., 1992; Marck and Grosjean, 2002).

Of the 6 required non-wobble tRNAs in each serovar, four are shared. These include two non-wobble singleton tRNAs covering codons uniquely, as well as a tRNA with the ATG anticodon which is post-transcriptionally modified by the required protein MesJ/TilS to recognize the isoleucine codon ATA (Marck and Grosjean, 2002). An additional two required tRNAs in both serovars, one shared and one with a differing anticodon, contain Gln anticodons and are part of a polycistronic tRNA operon containing other required tRNAs. This operon is conserved in *E. coli* with the exception of an additional tRNA-Gln at the 3' end that has been lost in the *Salmonella* lineage. It is possible that transposon insertions early in the operon may interfere with processing of the polycistronic transcript into mature tRNAs. Finally, I did not observe insertions in a tRNA-Met and a tRNA-Val in *S. Typhi* and *S. Typhimurium*, respectively.

Using this analysis of the tRNAs we estimate a worst-case PPV for these short molecules (~76 bases) at 81%, in line with my previous estimates for conserved protein-coding genes, and a FPR of <4%, higher than for protein-coding genes but still well within the typical tolerance of high-throughput experiments. This assumes that the “required” operonic tRNA-Glns and the serovar-specific tRNA-Met and tRNA-Val are all false positives; it is not clear that this is in fact the case.

Surveying the shared required ncRNA content of both serovars (see table 2.7), I found that the RNA components of the signal recognition particle (SRP) and RNase P, two universally conserved ncRNAs, are required as expected. The SRP is an essential component of the cellular secretion machinery, while RNase P is necessary for the maturation of tRNAs. I also found a number of required known and potential cis-regulatory molecules associated with genes required for growth under laboratory conditions in both serovars. The FMN riboswitch controls *ribB*, a 3,4-dihydroxy-2-butanone 4-phosphate synthase involved in riboflavin biosynthesis, in response to flavin mononucleotide concentrations (Winkler, Nahvi, and Breaker, 2002). Additionally, I was able to assign putative functions to a number of previously uncharacterized required non-coding transcripts through their 5' association with required genes. SroE, a 90 nucleotide molecule discovered in an early sRNA screen (Vogel et al., 2003), is consistently located at the 5' end of the required *hisS* gene across its phylogenetic distribution in the Enterobacteriaceae. Given this consistent association and the function of HisS as a histidyl-tRNA synthetase, I

hypothesize that this region may act in a manner similar to a T-box leader, inducing or repressing expression in response to tRNA-His levels. The *thrU* leader sequence, recently discovered in a deep-sequencing screen of *E. coli* (Raghavan, Groisman, and Ochman, 2011), appears to regulate a polycistronic operon of required singleton wobble tRNAs. Three additional required cis-regulatory elements, t44, S15, and StyR-8, are associated with required ribosomal proteins, highlighting the central role ncRNA elements play in regulating fundamental cellular processes.

**Table 2.7: Candidate required ncRNAs greater than 60 nucleotides in length, excluding rRNA and tRNA.** Known and putative non-coding elements classified as required or ambiguous in this screen. Required ncRNAs have a log-likelihood ratio (LLR) between required and non-required models of < -2; see Methods. \* †, ncRNAs which are ambiguous (LLR between -2 and 2) in *S. Typhi*(\*) or in *S. Typhimurium*(†). Hfq-binding annotations are taken from Chao et al. (2012). The downstream protein-coding genes columns report annotated CDS or ribosomal RNA start sites within 100 bases of each candidate required non-coding element on either strand, and whether these downstream sequences are also classified as required.

Name	Rfam accession	Function	Hfq-binding	Downstream protein-coding gene(s)	Downstream gene required	References
<b>Required or ambiguous in both <i>S. Typhi</i> and <i>S. Typhimurium</i></b>						
SRP	RF00169	RNA component of the signal recognition particle				Rosenblad et al. (2009)
RNase P	RF00010	RNA component of RNase P	<i>ybaZ</i>	N		Frank and Pace (1998)
RFN	RF00050	FMN-sensing riboswitch controlling the <i>ribB</i> gene	<i>ribB</i>	Y		Winkler, Nahvi, and Breaker (2002)
SroE	RF00371	Putative cis-regulatory element controlling the <i>hisS</i> gene	<i>hisS</i>	Y		Vogel et al. (2003)
ThrU Leader	NA	Putative cis-regulatory element controlling the ThrU tRNA operon				Raghavan, Groisman, and Ochman (2011)
t44	RF00127	Cis-regulatory element controlling the ribosomal <i>rpsB</i> gene	<i>rpsB</i>	Y		Tjaden et al. (2002); Aseev et al. (2008); Meyer et al. (2009)
S15†	RF00114	Translational regulator of the ribosomal S15 protein	<i>rpsO</i>	Y		Benard et al. (1996)
StyR-8	NA	Putative cis-regulatory element controlling the ribosomal <i>rpmB</i> gene	<i>rpmB</i>	Y		Chinni et al. (2010)
MicA	RF00078	sRNA involved in cellular response to extracytoplasmic stress	Y	<i>luxS</i>	N	Vogel (2009b)
DsrA†	RF00014	sRNA regulator of H-NS	Y	<i>mngB</i>	N	Lease, Cusick, and Belfort (1998)
STnc10	NA	Putative sRNA		<i>nhaA</i>	N	Sittka et al. (2008)
STnc60†	NA	Putative sRNA				Sittka et al. (2008)
STnc840	NA	Verified sRNA derived from 3' UTR of the <i>fgl</i> gene	Y			Chao et al. (2012)
IS0420*†	NA	Putative ncRNA		<i>rmf</i>	N	Raghavan, Groisman, and Ochman (2011); Chen, Lesnik, et al. (2002)
RGO0†	NA	Putative sRNA identified in <i>E. coli</i>				Raghavan, Groisman, and Ochman (2011)
<b>Required or ambiguous in <i>S. Typhimurium</i> only</b>						
rne5	RF00040	RNase E autoregulatory 5' element	<i>rne</i>		Y	Diwa et al. (2000)
RydC	RF00505	sRNA regulator of the <i>yejABEF</i> ABC transporter	Y			Antal et al. (2005)
RydB	RF00118	Putative ncRNA				Wassarman et al. (2001)
STnc510	NA	Putative sRNA		<i>pagD/pagC</i>	Y/N	Sittka et al. (2008)
STnc460†	NA	Putative sRNA				Sittka et al. (2008)
STnc170	NA	Putative sRNA		SL1458	N	Sittka et al. (2008)
STnc130	NA	Putative sRNA		<i>dmsA</i>	N	Sittka et al. (2008)
RseX	RF01401	sRNA regulator of OmpA and OmpC	Y			Douchin, Bohn, and Bouloc (2006)

IsrJ	RF01393	sRNA regulator of SPI-1 effector protein secretion					Sittka et al. (2008); Padalon-Brauch et al. (2008)
IsrI	RF01392	Island-encoded sRNA	Hfq-binding	Y	SL1028	Y	Sittka et al. (2008); Padalon-Brauch et al. (2008); Chao et al. (2012)
<b>Required or ambiguous in <i>S. Typhi</i> only</b>							
RybB	RF00110	sRNA involved in cellular response to extracytoplasmic stress		Y			Vogel (2009b)
tk5*	NA	Putative ncRNA					Raghavan, Groisman, and Ochman (2011); Rivas and Eddy (2001)
STnc750	NA	Verified sRNA		Y	<i>speB</i>	N	Kröger et al. (2012); Chao et al. (2012)
StyR-44*	RF01830	Putative multicopy (2/6 copies required in <i>S. Typhi</i> ) ncRNA associated with ribosomal RNA operon			23S rRNA	N	Chinni et al. (2010)
tp2	NA	Putative ncRNA			<i>aceE</i>	N	Raghavan, Groisman, and Ochman (2011); Rivas and Eddy (2001)
RdID	RF01813	RdID RNA anti-toxin, 1/2 copies required in <i>S. Typhi</i>					Kawano et al. (2002)
STnc120*	NA	Putative sRNA					Sittka et al. (2008)
tp28*	NA	Putative ncRNA		<i>fur</i>		N	Raghavan, Groisman, and Ochman (2011); Rivas and Eddy (2001)
Phe Leader*	RF01859	Phenylalanine peptide leader sequence associated with the required <i>pheST</i> operon			<i>pheS</i>	Y	Zurawski et al. (1978)
RimP Leader	RF01770	Putative cis-regulator of the <i>rimP-nusA-infB</i> operon			<i>rimP</i>	Y	Naville and Gautheret (2010)
GlmY	RF00128	Trans-acting regulator of the <i>glmS</i> gene					Urban and Vogel (2008)

### 2.3.5 sRNAs required for competitive growth

Inferring functions for potential trans-acting ncRNA molecules, such as anti-sense binding small RNAs (sRNAs), from requirement patterns alone is more difficult than for cis-acting elements, as one cannot rely on adjacent genes to provide any information. It is also important to keep in mind that TraDIS assays requirements after a brief competition within a large library of mutants on permissive media. This may be particularly important when surveying the bacterial sRNAs, which are known to participate in responses to stress (Vogel, 2009a).

This is demonstrated by two sRNAs involved in the  $\sigma^E$ -mediated extracytoplasmic stress response, RybB and RseX, both of which can be successfully knocked out in *S. Typhimurium* (83). In *S. Typhi*, *rpoE* is required, as it also is in *E. coli* (Baba et al., 2006; De Las Penas, Connolly, and Gross, 1997). However, in *S. Typhimurium*, *rpoE* tolerates a heavy insertion load, implying that  $\sigma^E$  mutants are not disadvantaged in competitive growth. In *S. Typhimurium*, the sRNA RseX is required. Overexpression of



**Figure 2.7: Proposed differences in sRNA utilization.** Diagram illustrating inferred required sRNA regulatory networks under TraDIS. Molecules required in *S. Typhi* are highlighted in yellow and in *S. Typhimurium* are highlighted in blue. RseA, in yellow/grey check, is ambiguous in *S. Typhi*. Non-required molecules are in grey. Diamonds indicate sRNAs, circles regulatory proteins, ovals proteases, oblong shapes are membrane-anchored proteins, and rounded squares are outer membrane porins.

RseX has previously been shown to compensate for  $\sigma^E$  essentiality in *E. coli* by leading to the degradation of *ompA* and *ompC* transcripts (85). This suggests that RseX may also be short-circuiting the  $\sigma^E$  stress response network in *S. Typhimurium* (figure 2.7). To our knowledge, this is the first evidence of a native (i.e. not experimentally induced) activity of RseX.

*S. Typhi* on the other hand requires  $\sigma^E$  along with its activating proteases RseP and DegS and anchoring protein RseA, as well as the  $\sigma^E$ -dependent sRNA RybB, which also regulates OmpA and OmpC in *S. Typhimurium*, along with a host of other OMPs (Papenfort et al., 2006). It is unclear why the  $\sigma^E$  response is required in *S. Typhi* but

not *S. Typhimurium*, though it may partially be due to the major differences in the cell wall and outer membrane between the two serovars. In addition, there are significant differences in the OMP content of the *S. Typhi* and *S. Typhimurium* membranes that may be driving alternative mechanisms for coping with membrane stress. For instance, *S. Typhi* completely lacks OmpD, a major component of the *S. Typhimurium* outer membrane (Santiviago, Toro, et al., 2003) and a known target of RybB (Vogel, 2009a).

Two additional sRNAs involved in stress response are also required by both *S. Typhi* and *S. Typhimurium*. The first, MicA, is known to regulate *ompA* and the *lamB* porin-coding gene in *S. Typhimurium* (Bossi and Figueroa-Bossi, 2007), contributing to the extracytoplasmic stress response. The second, DsrA, has been shown to negatively regulate the nucleoid-forming protein H-NS and enhance translation of the stationary-phase alternative sigma factor  $\sigma^S$  in *E. coli* (Lease, Cusick, and Belfort, 1998), though its regulation of  $\sigma^S$  does not appear to be conserved in *S. Typhimurium* (Jones, Goodwill, and Elliott, 2006). Both have been previously deleted in *S. Typhimurium*, and so are not essential. H-NS knockouts have previously been shown to have severe growth defects in *S. Typhimurium* that can be rescued by compensatory mutations in either the *phoPQ* two-component system or *rpoS*, implying that the lack of H-NS is allowing normally silenced detrimental regions to be transcribed (Navarre et al., 2006). As MicA has recently been shown to negatively regulate PhoPQ expression in *E. coli* (Coornaert et al., 2010), it is tempting to speculate that MicA may be moderating the effects of DsrA-induced H-NS repression; however, it is currently unclear whether sRNA regulons are sufficiently conserved between *E. coli* and *S. enterica* to justify this hypothesis.

## 2.4 Conclusions

The extremely high resolution of TraDIS has allowed the assaying of gene requirements in two very closely related salmonellae with different host ranges. I found, under laboratory conditions, that 58 genes present in both serovars were required in only one, suggesting that identical gene products do not necessarily have the same phenotypic effects in the two different serovar backgrounds. Many of these genes occur in genomic regions or metabolic systems which contain pseudogenes and/or have undergone reorganization since the divergence of *S. Typhi* and *S. Typhimurium*, demonstrating the complementarity of TraDIS and phylogenetic analysis. These changes may in part explain differences

observed in the pathogenicity and host specificity of these two serovars. In particular, *S. Typhimurium* showed a requirement for cell surface structure biosynthesis genes; this may be partially explained by the fact that *S. Typhi* expresses the Vi-antigen which masks the cell surface, though these genes are not required for survival in our assay. *S. Typhi* on the other hand has a requirement for iron uptake through the *fep* system, which enables ferric enterobactin transport. This dependence on enterobactin suggests that *S. Typhi* is highly adapted to the iron-scarce environments it encounters during systemic infections. Furthermore, this appears to represent a single point of failure in the *S. Typhi* iron utilization pathways, and may present an attractive target for narrow-spectrum antibiotics.

Of the approximately 4500 protein coding genes present in each serovar, only about 350 were sufficiently depleted in transposon insertions to be classified as required for growth in rich media. This means that over 92% of the coding genome has sufficient insertion density to be queried in future assays. Dense transposon mutagenesis libraries have been used to assay gene requirements under conditions relevant for infection, including *S. Typhi* survival in bile (Langridge et al., 2009), *Mycobacterium tuberculosis* catabolism of cholesterol (Griffin et al., 2011), drug resistance in *Pseudomonas aeruginosa* (Gallagher, Shendure, and Manoil, 2011), and *Haemophilus influenzae* survival in the lung (Gawronski et al., 2009). I expect that parallel experiments querying gene requirements under the same conditions in both serovars examined in this study will yield further insights into the differences in the infective process between *Typhi* and *Typhimurium*, and ultimately the pathways that underlie host-adaptation.

Both serovars possess substantial complements of horizontally-acquired DNA. I have been able to use TraDIS to assay these recently acquired sequences. In particular, I have been able to identify, on a chromosome wide scale, active prophage through the requirement for their repressors. The P4 phage utilizes an RNA-based system to make decisions regarding cell fate, and structurally similar systems are used by P1, P7, and N15 phage (Citron and Schuster, 1990; Ravin, Svarchevsky, and Deho, 1999). C4-like transcripts have been regarded as the primary repressor of lytic functions, though the IsrK-like sequence is known to be essential to the establishment of lysogeny in P4 and is transcribed in at least two phage types (Sabbattini et al., 1995; Ravin, Svarchevsky, and Deho, 1999). These observations in *S. Typhi* suggest an important role for the IsrK-like sequence in maintenance of the lysogenic state in P4-like phage, though the mechanism

remains unclear.

Recent advances in high-throughput sequencing have greatly enhanced our ability to detect novel transcripts, such as ncRNAs and short open reading frames (sORFs). In fact, our ability to identify these transcripts now far out-strips our ability to experimentally characterize these sequences. There have been previous efforts at high-throughput characterization of bacterial sRNAs and sORFs in enteric bacteria; however, these have relied on labor-intensive directed knockout libraries (Santiviago, Reynolds, et al., 2009; Hobbs, Astarita, and Storz, 2010). Here I have demonstrated that TraDIS has sufficient resolution to reliably query genomic regions as short as 60 bases, in agreement with a recent high-throughput transposon mutagenesis study in the  $\alpha$ -proteobacteria *Caulobacter crescentus* (Christen et al., 2011). This method has the major advantage that library construction does not rely upon genome annotation, and newly discovered elements can be surveyed with no further laboratory work.

I have been able to assign putative functions to a number of ncRNAs using TraDIS through consideration of their genomic and experimental context. In addition, ncRNA characterization generally is done in model organisms like *E. coli* or *S. Typhimurium*, and it is unclear how stable ncRNA regulatory networks are over evolutionary time. By assaying two serovars of *Salmonella* with the same method under the same conditions, I have seen hints that there may be differences in sRNA regulatory networks between *S. Typhi* and *S. Typhimurium*. In particular, I have found that under the same experimental conditions, *S. Typhi* appears to rely on the  $\sigma^E$  stress response pathway while *S. Typhimurium* does not; it is tempting to speculate that this difference in stress response is mediated by the observed difference in requirement for two sRNAs, RyB and RseX. I believe that this combination of high-throughput transposon mutagenesis with a careful consideration of the systems context of individual genes provides a powerful tool for the generation of functional hypotheses. I anticipate that the construction of TraDIS libraries in additional organisms, as well as the passing of these libraries through relevant experimental conditions, will provide further insights into the function of bacterial ncRNAs in addition to the protein-coding gene complement.

# Chapter 3

## Methods for the analysis of TraDIS experiments, with an application to *Salmonella* macrophage invasion

Section 3.2 describes a collaborative study with Gemma C. Langridge (Pathogen Genomics, Wellcome Trust Sanger Institute). Gemma performed all laboratory experiments described in this chapter unless otherwise noted.

### 3.1 Introduction

In the previous chapter, I described the results of a study predicting and comparing the genes required for robust growth of two *Salmonella* serovars in standard laboratory media. While this revealed interesting aspects of *Salmonella* biology, linking these findings to *Salmonella*'s infective niche in the human host is difficult. However, transposon-insertion sequencing can be used to interrogate infective conditions directly (reviewed previously in section 1.5): by comparing libraries passed through a condition of interest to control libraries, we can determine the genomic regions involved in survival in that condition. In this chapter, I describe a pipeline I have devised for the analysis of such experiments, illustrated with an experiment assaying genes required for *S. Typhi* and *Typhimurium* invasion of (or uptake into) human macrophage. These methods have been adopted by Pathogen Informatics at the Sanger, form the basis of the current Sanger

pipelines for analysis of TraDIS experiments, and are currently being used in a variety of transposon-insertion sequencing studies.

### 3.1.1 *Salmonella* interactions with macrophage

As previously described in section 2.1, the ability to invade and survive in host cells was a major factor in the early evolution of *S. enterica* subspecies *enterica*; survival in macrophages in particular is known to be necessary for virulence (Fields et al., 1986). This ability appears to have been largely driven by the acquisition of two horizontally-acquired pathogenicity islands, SPI-1 and -2. Due to the availability of a mouse model of systemic infection (Santos, Zhang, et al., 2001), most of what is known about *Salmonella* interactions with host cells is derived from studies of *S. Typhimurium* infection.

*S. Typhimurium* infections of either epithelial or phagocytic cells appear to follow broadly similar paths (Figueira and Holden (2012), see also figure 3.1). On encountering a suitable host cell, the bacterium adheres using an array of fimbrial adhesins (Baumler, Tsolis, and Heffron, 1996; Velden et al., 1998). The SPI-1 T3SS, a needle-like complex spanning the periplasm and presenting its tip to the exterior of the bacterial cell (Mueller, Broz, and Cornelis, 2008), induces membrane ruffling in the host cell through secretion of effector proteins (Zhou and Galán, 2001), facilitating bacterial uptake. While use of this mechanisms is not strictly necessary for entrance to phagocytic cells such as macrophage, *S. Typhimurium* strains unable to induce ruffling are taken up six to ten times less efficiently than the wild-type (Monack et al., 1996), though the entry mechanism does not ultimately affect cell fate (Rathman, Barker, and Falkow, 1997).

Once entry has been gained to the cell, through either active invasion or phagocytic engulfment, *S. Typhimurium* begins expressing a second T3SS encoded on SPI-2. The effectors secreted by this T3SS allow *S. Typhimurium* to remodel the *Salmonella* containing vacuole (SCV), and even modulate host immune signalling (see figure 3.1). There is some controversy as to whether or not the SCV undergoes fusion with lysosomes; a recent study suggests it does, but that the activity of these lysosomes is first modulated by the SPI-2 effector SifA (McGourty et al., 2012). Little is known about the growth conditions *S. Typhimurium* faces within the SCV, though transcriptomic studies suggest it is aerobic, mildly acidic, rich in gluconate, and limited in aromatic amino acids, purines, and pyrimidines (Eriksson et al., 2003; Hautefort et al., 2008).

Our understanding of how these findings relate to *S. Typhi* infections of human macrophage is limited, largely due to the lack of a non-human model organism for infection by this serovar. A recent study suggests that SPI-2 may not even be necessary for *S. Typhi* invasion of and survival in human macrophages (Forest et al., 2010), though SPI-2 genes are known to be expressed by *S. Typhi* in macrophages (Faucher et al., 2006) and a SPI-2 deletion mutant was previously shown to be attenuated under these conditions (Khan et al., 2003). Regardless, it is well established that the genotype of both the *Salmonella* strain used and the macrophage can have profound effects on the course of infection. A number of studies comparing a variety of *Salmonella* serovars infecting murine-, human-, and even chicken-derived macrophages have repeatedly shown that serovars exhibit remarkably different behaviors under the same conditions (Buchmeier and Heffron, 1989; Vladoianu, Chang, and Pechère, 1990; Schwan et al., 2000; Okamura et al., 2005); these differences appear to correlate somewhat with the degree of host-adapation exhibited by the serovar. In this study we compare our *Salmonella* TraDIS libraries following uptake by human macrophage in the hopes of uncovering genomic factors underlying these differences in behavior.

### 3.1.2 Conditional gene fitness

Determining conditional gene fitness presents a somewhat different problem to that addressed in the previous chapter, predicting and comparing “essential” genes under the conditions of library creation. In predicting gene essentiality, we had a single time point representing the initial growth of the library on rich media, while in identifying conditional gene fitness (measured as the relative expansion or contraction of mutant populations) we are always comparing changes in mutant fitness with respect to fitness in a baseline condition. The ratio of reads between the two conditions is taken as indicative of differences in relative mutant prevalences between them. In some ways, this makes the problem of identifying genes with strong fitness effects easier: as we are primarily interested in the ratio of various insertion mutants present between the two conditions, effects that may confound the prediction of simple gene essentiality are effectively “zeroed out”. More explicitly, whether low insertion density in the initial library occurs due to chance, nucleotide composition bias, or the exclusionary effects of high-density DNA-binding proteins (described in section 2.3.3) does not matter – these



**Figure 3.1: Biogenesis of the *Salmonella* containing vacuole (SCV).** *Salmonella* adheres to the outer membrane of host cells, and uses the SPI-1 T3SS and its associated effectors to induce membrane ruffling and entry into the SCV. The SPI-2 T3SS functions mainly in maintenance of the SCV, through the action of the effectors SifA, SopD2, SseJ and PipB2 (orange boxes), and its localization near the Golgi of host cells, mediated by SseF and SseG (purple boxes). Other effectors are involved in modulation of host immune signalling (SpvC, SspH1 and SseL; pink boxes) or target the host cytoskeleton (SteC, SpvB, SspH2 and SrfH; blue boxes). Reproduced from Figueira and Holden (2012) under a Creative Commons Attribution License (CCAL).

regions can simply be identified as not producing sufficient reads over insertion-sites to be assayed and removed from the analysis.

In many ways, the problem of investigating the statistical and biological significance of ratios of reads over insertion sites resembles established analyses developed for differential RNA-seq analysis. In the following sections I describe the application of these methods to the problem of determining conditional gene fitness using the *Salmonella* macrophage infection dataset as an example.

## 3.2 Experimental methods

Gemma C. Langridge performed all laboratory experiments described in this chapter, as well as read mapping; condensed descriptions are included here for completeness. Silvia Pinero prepared the THP-1 cells for infection. Sabine Eckert and Daniel Turner (Wellcome Trust Sanger Institute) performed the nucleotide sequencing. A more detailed description of the experimental methods is available in Langridge (2010), including preliminary assessments of bacterial strain ability to grow in RPMI, invade THP-1 derived macrophage, and experiment optimization.

### 3.2.1 Strains and cell lines

These experiments were performed with *S. Typhi* WT174 and *S. Typhimurium* SL3261 transposon mutant libraries, described in chapter 3. Annotations and orthology predictions used are as in chapter 3. Human monocytic cell line THP-1 was used for cell infections.

### 3.2.2 Preparation of THP-1 cells

THP-1 cells were grown up from frozen stocks in RPMI-1640 supplemented with 10% heat-inactivated foetal bovine serum and 2 mM L-glutamine, and incubated without shaking in vented flasks (VWR, Lutterworth, UK) at 37°C in the presence of 5% CO<sub>2</sub>. Culture volumes were split and given fresh media every 3-4 days until the desired volume and cell density was achieved. Phorbol myristate acetate (PMA) was used to differentiate the THP-1 monocytes. Briefly, approximately 212 cells in 4 mL supplemented RPMI

containing 0.125 ng/mL PMA were seeded into each well of a 6-well plate and incubated for six days at 37°C in 5% CO<sub>2</sub>. On the day of infection, the PMA-containing media was removed, cells were washed with dPBS and fresh warmed, supplemented RPMI was added to maintain the cells while the bacterial inoculum was prepared.

### 3.2.3 Preparation of transposon libraries

Frozen stocks of the Typhi library were found to be at half the concentration of the Typhimurium library by OD600. To ensure similar concentrations for the infection assay, a 1 in 5000 dilution of the Typhi library and a 1 in 10,000 dilution of the Typhimurium library was used to inoculate the growth medium. Cultures of each transposon library were grown with shaking in 100 mL of RPMI-1640 supplemented with 0.3 g/L L-glutamine and buffered with 10 mL 1 M MOPS at 37°C for 16 hours. These cultures were subcultured 1 in 20 into fresh RPMI supplemented and buffered as before, and grown for between 3 and 4 hours to mid-log phase (OD600 of 2.4).

### 3.2.4 Infection assay

Five 6-well plates were used for each run of the assay. In total, 29 wells were infected with the bacterial inoculum and one served as a blank control for eukaryotic cell contamination. At the start of the assay, media was removed from all wells except for the blank control, and a 3 mL bacterial inoculum was added to each experimental well. The plates were centrifuged for 5 minutes at 600 x g and incubated at 37°C in 5% CO<sub>2</sub> for 30 minutes. A 4-6 mL aliquot of the inoculum was processed for genomic DNA as the TraDIS control. After 30 minutes, media was removed from all wells, and fresh RPMI additionally supplemented with 100 µg/mL gentamicin was added. After 2 hours the wells were washed 3 times in plain dPBS. Following washing, 500 µL of 1% Triton-X-100 was added to each well to lyse the eukaryotic cells, mixed well by pipetting, and incubated at 37°C in 5% CO<sub>2</sub> for 2 minutes. The cell suspensions from all experimental wells were pooled for bacterial DNA extraction. Genomic DNA was extracted using the Qiagen DNeasy Blood and Tissue kit, according to the manufacturers protocol for Gram negative bacteria. Sequencing was performed as described in section 2.2.4.

## 3.3 Analysis of conditional gene fitness using TraDIS

### 3.3.1 Experimental design

The goal for this experiment was to determine the differences in gene requirements for human macrophage infection in two *Salmonella* serovars: Typhimurium, a host-generalist, and Typhi, host-restricted to humans as described in the previous chapter. To this end, infection assays of THP-1 monocytes were performed in triplicate with transposon libraries for each serovar at high multiplicities of infection in an attempt to avoid bottleneck effects. These were compared to libraries grown in cell culture medium (RPMI), to control for any incidental changes in library composition due to growth in this medium.

### 3.3.2 Mapping insertion sites

Read mapping is a special case of one of the oldest problems in bioinformatics, aligning a short sequence of length  $n$  to a much longer sequence, or database of sequences, of length  $m$ . An optimal solution (with respect to a particular sequence similarity scoring scheme) for this problem using dynamic programming was first proposed by Smith and Waterman (1981), building on previous work by Needleman and Wunsch (1970). Unfortunately, this method requires construction of a dynamic programming matrix of size  $n \times m$ , which quickly becomes impractical for large  $m$  due to both time and memory constraints. Heuristic solutions to this problem have been developed, starting with the FASTA and BLAST algorithms (Lipman and Pearson, 1985; Altschul et al., 1990). The basic idea behind these heuristics is to rapidly search for identical matches using a hash of the sequence database before performing a full Smith-Waterman style local alignment around this match. For the case of mapping reads to larger eukaryotic genomes, more powerful heuristics, such as the Burrows-Wheeler transform (Burrows and Wheeler, 1994; Langmead, Trapnell, et al., 2009; Li and Durbin, 2010), may be required due to time and space constraints. However as we are working with relatively small bacterial genomes, MAQ (Li, Ruan, and Durbin, 2008) has been used here, which is similar in spirit to FASTA or BLAST, but with additional refinements to deal with repetitive genomic regions and to assessing alignment quality.

**Table 3.1: Summary statistics for macrophage infection assay sequencing runs.**  
 Table columns as follows: 1, description; 2, total sequencing reads; 3, reads containing transposon tag; 4, reads mapped to chromosome with quality score greater than 20; 5, number of insertions recovered. STY: *S. Typhi*; STM: *S. Typhimurium*.

Description	Reads	Reads tagged	Reads mapped	Insertion sites
STY control 1	11107014	10534361	9722100	154356
STY control 2	10983030	10016035	8868829	193417
STY control 3	13506872	12168442	11062549	180998
STY infection 1	7526390	4193529	2304138	90218
STY infection 2	8630360	4166256	2000771	73154
STY infection 3	8215834	4323817	2459573	98894
STM control 1	14583559	14314003	9318191	365266
STM control 2	18119496	17494267	11458349	464036
STM control 3	13565707	12457266	7312946	179702
STM infection 1	3292265	2972803	2033041	41775
STM infection 2	6444469	5351193	3732480	59476
STM infection 3	13012186	12124834	9633788	43110

### 3.3.3 Quality control

We can assess the quality of TraDIS experiments on multiple levels: the number of reads containing transposon tags and mapping to the genome, the number of insertion sites recovered, the correlation between the numbers of reads recovered for each gene in replicated experiments, and clustering experiments using a dimensionality reduction technique such as principal component analysis (PCA).

Summary statistics of the sequencing runs for this study are presented in table 3.1. Total read yield varied from  $\sim 3.3$  to  $\sim 18.1$  million reads, with lower yields generally observed for the infection libraries. Similarly, the percentage of reads containing exact matches to transposon sequence is significantly lower in the *S. Typhi* infection samples, which may be a result of low read quality obscuring the sequence. However, despite these issues, over two million reads over insertion sites were recovered in every sample which provides adequate coverage for this assay. Interestingly, the number of unique insertion sites recovered from the *S. Typhimurium* infection assays was approximately half that observed in the *S. Typhi* assay in every replicate, despite having an apparently more complex inoculum. This is suggestive of either stronger selective pressure, a more severe bottleneck effect, or both for *S. Typhimurium* compared to *S. Typhi* during infection of human macrophage, as might be expected given the latter's host adaptation.

Linear correlation coefficients, reported in table 3.2 lend some credence to this idea that *S. Typhimurium* may be experiencing a more severe bottleneck leading to the incidental loss of mutants during infection, possibly due to the killing effects of the

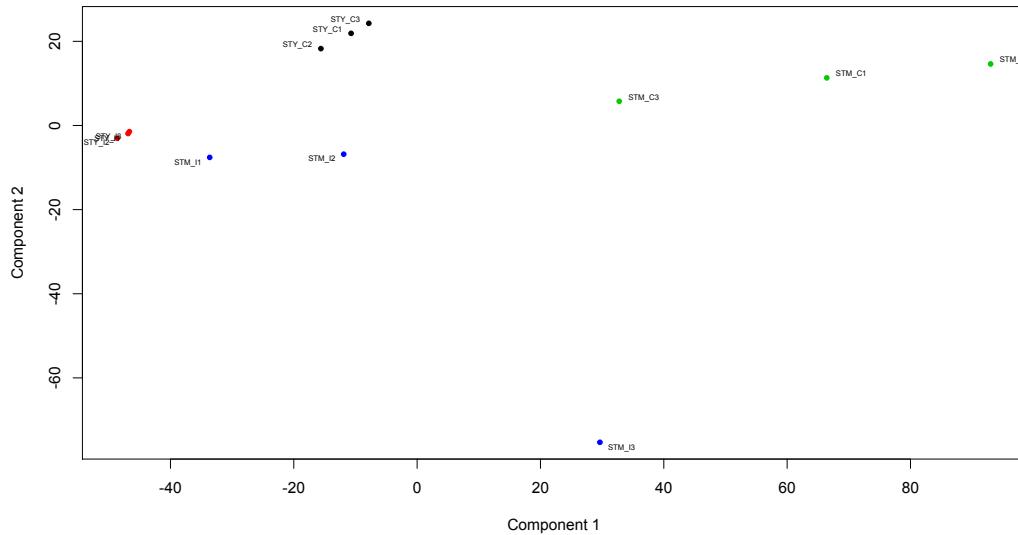
**Table 3.2: Pearson's  $r$  between replicated TraDIS experiments.** Correlations of reads over genic and non-coding RNA features between replicated control and infection assays, rounded down to nearest hundredth. Y: *S. Typhi*; M: *S. Typhimurium*; C: Control; I: Infection.

	Y C1	Y C2	Y C3	Y I1	Y I2	Y I3	M C1	M C2	M C3	M I1	M I2	M I3
Y C1	1.00	0.99	0.99	0.65	0.69	0.72	0.43	0.43	0.48	0.34	0.39	0.43
Y C2	0.99	1.00	0.99	0.65	0.70	0.72	0.42	0.43	0.48	0.33	0.39	0.43
Y C3	0.99	0.99	1.00	0.67	0.71	0.74	0.44	0.44	0.49	0.34	0.40	0.45
Y I1	0.65	0.65	0.67	1.00	0.99	0.99	0.26	0.28	0.32	0.30	0.31	0.49
Y I2	0.69	0.70	0.71	0.99	1.00	0.99	0.28	0.28	0.34	0.31	0.33	0.49
Y I3	0.72	0.72	0.74	0.99	0.99	1.00	0.29	0.29	0.35	0.31	0.33	0.50
M C1	0.43	0.42	0.44	0.26	0.28	0.29	1.00	0.99	0.93	0.74	0.85	0.76
M C2	0.43	0.43	0.44	0.26	0.28	0.29	0.99	1.00	0.93	0.73	0.85	0.77
M C3	0.48	0.48	0.49	0.32	0.34	0.35	0.93	0.93	1.00	0.69	0.80	0.75
M I1	0.34	0.33	0.34	0.30	0.31	0.31	0.74	0.73	0.69	1.00	0.74	0.68
M I2	0.39	0.39	0.40	0.31	0.33	0.31	0.85	0.85	0.80	0.74	1.00	0.72
M I3	0.43	0.43	0.45	0.49	0.49	0.50	0.76	0.75	0.75	0.68	0.72	1.00

macrophage. Correlations between replicate experiments are over .99 with two notable exceptions. The first is in the third replicate of the *S. Typhimurium* assay. Due to failure of this replicate during the current study, an earlier 2 hour time point from optimization experiments (Langridge, 2010) was used, so the lower correlation between the third control replicate and replicates 1 and 2 may be explained by this sample being handled at a different time and sequenced earlier on a different machine. However, the correlation coefficient between the third control replicate and replicates 1 and 2 is still well over .9, indicating that it still largely agrees with the later experiments.

The other discrepancy is in the correlation between *S. Typhimurium* infection experiments, with coefficients ranging between .68 and .74. This is still a high positive correlation; however it does not reach the level observed in the other replicated experiments in this study. This is again suggestive of a bottleneck effect in this assay. If the loss of particular mutants were purely due to selection, we would expect a high correlation, as these losses would presumably be reproducible under the same experimental conditions. Rather, it appears that there is some stochasticity in the loss of mutants in this particular experiment, suggesting losses that are incidental to the actual factors underlying infection of human macrophage. As mentioned previously, this may be due to a higher rate of macrophage killing of the non-host adapted *S. Typhimurium* strain used in this study. It has been observed previously that even in *S. Typhimurium* strains capable of successfully infecting macrophage, some proportion of the invading bacteria do not manage to establish a protective SCV (Monack et al., 1996) for reasons that remain unclear. A higher rate of failure in establishing the SCV in human macrophage for *S. Typhimurium* than *S. Typhi*, or even the use of an entirely different mechanism

for survival in macrophage by *S. Typhi*, may explain this difference.



**Figure 3.2: Principal component analysis of TraDIS macrophage infection assays.** Plot showing samples along first two components of a PCA, representing 55% and 18% of the total variance in the data set, respectively. Replicates appear to cluster together, with the exception of the third Typhimurium infection replicate, which was excluded from the analysis. STY: *S. Typhi*; STM: *S. Typhimurium*; C: control; I: infection.

To further investigate the potential bottleneck effect in *S. Typhimurium*, I performed a principal component analysis (PCA) on all samples. PCA is a mathematical technique for dimensional reduction which identifies linear vectors (components) in a high-dimensional dataset which capture maximal amounts of the variance between samples. This high dimensional data can then be visualized in a lower (e.g. 2 or 3) dimensional space by plotting samples against these components. Samples were centered and scaled to correct for the differences in read counts between experiments. Plots of all samples in this study on the first two principal components, accounting for 55% and 18% of the total variance respectively, are shown in figure 3.2. With the exception of the third *S. Typhimurium* infection experiment, all samples collected under the same conditions cluster on this plot, as would be expected if these results are reporting the effects of differential selection. All infection samples lie to the left of their respective control

samples on the first component, suggesting that the dominant signal in this data is due to the effects of selection during macrophage infection. The fact that two of the three *S. Typhimurium* infection experiments cluster together suggests that this signal is stronger than any stochastic bottleneck effect despite the lower correlations observed between these libraries, and that it should be possible to derive useful information about the conditions faced by *S. Typhimurium* during infection from these experiments. Unfortunately, the third *S. Typhimurium* infection replicate, which was performed separately as described earlier, does not cluster well with these. I performed a similar analysis using the plotMDS function of edgeR (Robinson, McCarthy, and Smyth, 2010), which performs multidimensional scaling using a variance-stabilized distance measure between samples, and came to a similar result. It is unclear why this replicate is so different, and it may be due to differences in experimental set up or sequencing. I excluded this replicate from further analysis on this basis.

### 3.3.4 Inter-library normalization

Normalization is a critical part of any high-throughput sequencing experiment. As observed in the previous section, even the same experiment repeated on the same machine can lead to very different read counts. The naive approach to solving this problem is simply to scale each sequencing library by some factor so that the total read counts are equivalent. This may be adequate for analysis of technical replicates where gene expression levels are identical between all samples. However, Robinson and Oshlack (2010) illustrate why this may not be the case for the comparison of sequencing libraries sampling populations under different conditions with a simple thought experiment:

Imagine we have a sequencing experiment comparing two RNA populations, A and B. In this hypothetical scenario, suppose every gene that is expressed in B is expressed in A with the same number of transcripts. However, assume that sample A also contains a set of genes equal in number and expression that are not expressed in B. Thus, sample A has twice as many total expressed genes as sample B, that is, its RNA production is twice the size of sample B. Suppose that each sample is then sequenced to the same depth. Without any additional adjustment, a gene expressed in both samples will have, on average, half the number of reads from sample A, since the reads are spread

over twice as many genes. Therefore, the correct normalization would adjust sample A by a factor of 2. (Robinson and Oshlack, 2010)

More generally, we can think of each gene in a sequencing library as representing a slice of a pie. If a particular gene increases in expression (or mutant prevalence for transposon-insertion sequencing such as ours), then the space left in this pie for other genes necessarily shrinks. A scaling normalization which does not take this fact in to account, but simply assumes all pies are the same size would necessarily underestimate expression (or prevalence) for the majority of genes which don't change, while overestimating it for the few that do. A recent study has shown that normalization methods which explicitly account for this problem perform better on both real and simulated data (Dillies et al., 2012). Here I have used the trimmed mean of M-values (TMM) method, which assumes the majority of genomic features do not change in actual expression (or mutant prevalence here) and attempts to align the read counts of these features to produce an appropriate scaling factor (Robinson and Oshlack, 2010).

### 3.3.5 Identifying fitness effects

#### 3.3.5.1 Theory

Once sequencing libraries have been normalized, the next step in determining fitness effects is the choice of a proper test to determine the significance of changes in read counts. In the previous chapter, I used two test statistics. The first was to test for gene requirements within a particular library, and this was accomplished by fitting gamma distributions to the two modes observed in the empirical distributions of insertion indexes, then setting a threshold based on a log-odds ratio (see figure 2.2). The second was to additionally test for significant differences in read depth between the *S. Typhimurium* and *S. Typhi* libraries. In this case the  $\log_2$  read ratios between genomic features in the two libraries were roughly normally distributed, and I was able to set a significance threshold based on a fitted normal curve.

Neither of these tests are entirely appropriate for the present situation of identifying reproducible changes in mutant prevalence in replicated experiments. Most obviously, neither of these test can easily be modified to accommodate replicates, which is essential for robust identification of changes in mutant prevalence. Secondly, both tests are

dependent on manual fitting of gamma or normal distributions, which can not easily or robustly be automated. Standard statistical tests, such as the two sample Student's T-test or Mann-Whitney U-test are not applicable due to the small numbers of replicates (3 here, often 2) because of high experimental overhead in replication. Fortunately, these problems have largely been addressed in modern RNA-seq differential expression analysis software.

The two leading packages for analysis of RNA-seq based differential expression analysis are DESeq (Anders and Huber, 2010) and edgeR (Robinson, McCarthy, and Smyth, 2010). Both assume that sequence count data is negative binomially distributed. The negative binomial distribution arises naturally in the case of a Poisson process sampling from gamma-distributed random variables (Fisher, 1941). Sequencing of mixed populations of oligonucleotides has long been theorized to behave as a Poisson process, and this has shown to roughly be the case for technical replicates of Illumina RNA-seq runs (Marioni et al., 2008), i.e. repeated sequencing of the same input sample. Other studies have shown that biological RNA-seq and SAGE replicates, i.e. repeated experiments, generate extra-Poisson variability (Lu, Tomfohr, and Kepler, 2005; Robinson and Smyth, 2007), possibly due to variability in the concentration of the transcripts being sampled, which can be captured by the negative binomial.

This leads naturally to the question, is transposon-insertion sequencing data negative binomially distributed? Obviously, technical replicates of TraDIS experiments will be roughly Poisson distributed, as this is identical to the case of technical replication of RNA-seq. The question then becomes whether the underlying distribution of mutant prevalences being sampled by sequencing can be effectively modelled by a gamma distribution. Theoretical considerations indicate that this model may be appropriate: as subcultures of the mutant library expand, the number of insertion mutants per gene will be the summed result of independent exponentially-expanding clones, which will be gamma distributed assuming the starting populations are roughly equal. The only way to answer this question definitively would be to repeat the same experiment a large number of times, which is impractical. However, this is not necessary. Lu, Tomfohr, and Kepler (2005) showed that the negative binomial assumption is highly robust to the actual distribution of the data being assessed. In fact, it appears that the underlying transcript prevalences being sampled by RNA-seq experiments may actually be distributed according to a sum of log-normal distributions (Bengtsson et al., 2005); this does not prevent

DESeq and edgeR from performing competitively in benchmarks of differential expression analysis (Kvam, Liu, and Si, 2012; Soneson and Delorenzi, 2013). These approaches have previously been successfully applied to other Illumina sequencing-based experiments which likely have different underlying distributions than transcriptomic data, for instance differential analysis of ChIP-seq data (Robinson, Strbenac, et al., 2012).

I have used edgeR (Robinson, McCarthy, and Smyth, 2010) for significance testing here, an R package which implements the TMM normalization (Robinson and Oshlack, 2010), an approximation to an empirical Bayes estimation of feature-wise negative binomial dispersion parameters (Robinson and Smyth, 2007), and a version of Fisher’s exact test modified to deal with overdispersed data (Robinson and Smyth, 2008) as well as a likelihood-ratio test in the case of multifactorial designs (McCarthy, Chen, and Smyth, 2012; Lund et al., 2012). After testing, we are interested primarily in two values: the P-value given by the statistical testing which tells us how confident we can be that mutant prevalence differs between two conditions given the estimated negative binomial distribution distribution of read counts, and the  $\log_2$  fold-change (logFC) which gives an estimation of the magnitude of the difference. LogFC is calculated as

$$\log FC_g = \log(n_{g,b}) - \log(n_{g,a})$$

where the index  $g$  indicates the genomic feature being tested,  $n_{g,b}$  is the normalized average read count in the test condition, and  $n_{g,a}$  is the normalized average read count in the control condition. This subtraction is equivalent to taking the log of the ratio  $\frac{n_{g,b}}{n_{g,a}}$ , and hence  $\log FC_g$  becomes unstable for small changes in  $n_{g,b}$  as  $n_{g,a} \rightarrow 0$ , and is ultimately undefined when  $n_{g,a} = 0$ . In the previous chapter I corrected for this by adding a pseudocount to each gene’s read count. I take the same approach here, as implemented in edgeR, only since each library has been normalized by a different factor, I rather use the transformation

$$n_{g,x}^T = \log\left(\frac{n_{g,x}}{L_x} + \frac{2}{\bar{L}}\right)$$

where  $L$  is the library size. This has the effect of shrinking unreliable logFCs for features with small read counts, and removing the problem of undefined logFCs.

### 3.3.5.2 Application to macrophage infection data

Returning to the macrophage infection assays, I first eliminated genomic features from consideration which did not have at least 20 counts per million normalized reads (CPM) in at least three assay or control replicates. This cut-off is arbitrary, but serves the purpose of removing features from consideration which do not have adequate read coverage to deliver biologically significant results in at least one condition. This provides two advantages: firstly, it increases statistical power by reducing the number of simultaneous hypothesis tests that need to be corrected for, and secondly, it eliminates features which may have statistically significant logFCs but may not have large enough mutant populations to determine if these effects are biologically relevant. This reduces the number of genomic features tested from 3882 (including all orthologous coding sequences and non-coding RNAs) to 3596.

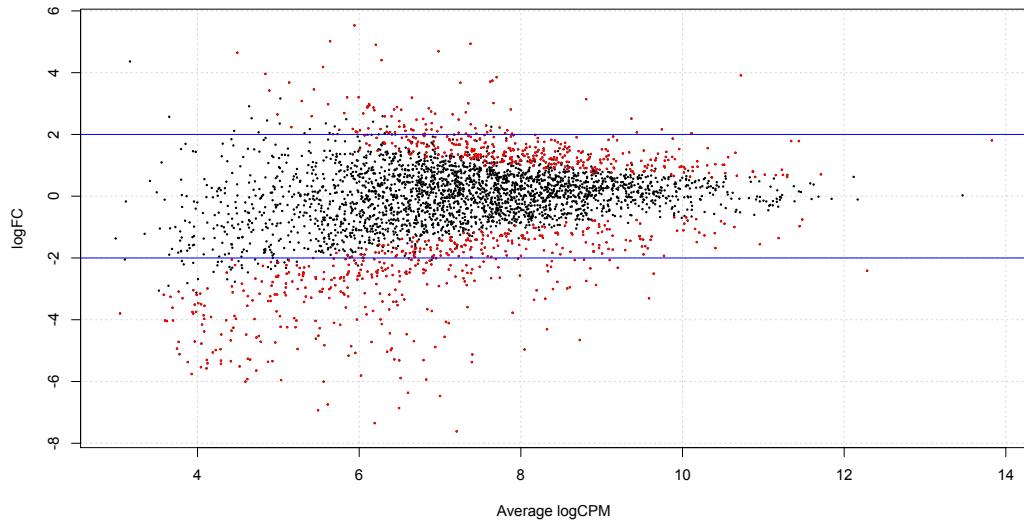
I then set up three statistical analyses within the generalized linear model (GLM) framework provided by edgeR, which allows for multi-factorial analyses. The first tests whether the logFC between *S. Typhimurium* infection and control is different from the logFC between *S. Typhi* infection and control. This allows me to discriminate between mutant populations which behave similarly during macrophage invasion in the two serovars (no or small difference in logFCs), from those which behave differently (large difference in logFCs). Of course, this does not allow me to discriminate between mutant populations which are expanding, those that are shrinking, or those which are static in both serovars during invasion - this test only tells if their behavior is similar. Similarly, using this test I can not discriminate between features with differences in logFCs that are the result of mutant expansion in one serovar, or contraction in another. For this reason I performed two additional analyses, testing the significance of logFCs between infection and control in each serovar independently. All p-values have been corrected for multiple testing using the method of Benjamini and Hochberg (1995), controlling for a false discovery rate (FDR) of 10%.

The results of the comparison between *S. Typhimurium* and *S. Typhi* changes in logFC over macrophage infection are shown in figure 3.3, and the individual changes in mutant prevalences for each serovar are shown in figures 3.4 and 3.5. On first viewing these figures, there is a striking difference in the behavior of the *S. Typhimurium* and *S. Typhi* mutant libraries: while *S. Typhimurium* displays a wide spread of changes



**Figure 3.3: Smear plot of differences in logFC over macrophage infection between *S. Typhimurium* and *S. Typhi*.** Each point in this plot represents a tested genomic feature. LogFC is reported on the Y-axis, logCPM on the X-axis; statistically significant features at a FDR of 0.1 are in red. The blue lines represent logFCs of  $|2|$ , translating to a four-fold difference in logFC in mutant prevalences between the two serovars. Negative values indicate that the *S. Typhimurium* mutant population has shrunk relative to the *S. Typhi* mutant population, and vice versa.

in mutant prevalence, 938 of them statistically significant NEED TO PUT IN SUPP DATA, indicating a strong selective pressure operating on the library, the composition of the *S. Typhi* library appears nearly unchanged after infection, with only 28 features showing a statistically significant change in mutant prevalence (see table 3.5). In fact it appears that nearly all of the statistically significant differences in logFC between the two libraries over macrophage infection are due to changes in mutant prevalences in the *S. Typhimurium* library. This seems to indicate, on a gross level, that *S. Typhi* is somehow avoiding the brunt of the gauntlet imposed on *S. Typhimurium* in the first two hours of macrophage infection. This may partially be due to the presence of the Vi capsule on *S. Typhi*, which has previously been shown to enhance survival in THP-1 derived macrophage (Hirose et al., 1997) through the creation of a ‘stealth’ phenotype which reduces the expression of inflammatory factors, such as TNF- $\alpha$ , by the macrophage.



**Figure 3.4: Smear plot of logFC in mutant prevalences over macrophage infection in *S. Typhimurium*.** Each point in this plot represents a tested genomic feature. LogFC is reported on the Y-axis, logCPM on the X-axis; statistically significant features at a FDR of 0.1 are in red. The blue lines represent logFCs of  $|2|$ , translating to a four-fold change in mutant prevalences between infection and control. Negative values indicate a reduction over infection in mutant prevalence, positive values an increase.

### 3.3.6 Functional analysis of gene sets that affect fitness

Now that I have determined the changes in mutant prevalences in each library over macrophage infection, I am left with the task of determining the biological context and importance of these changes. The traditional approach, taken in the previous chapter with regards to genomic features required for survival under standard laboratory conditions, would be to create a ranked list and work through these features one at a time, researching what is known about them and building a picture of their contribution to survival in the macrophage. This has some distinct advantages, as it allows the investigator to piece together new hypotheses as to gene function from the existing literature. However, while it was possible with  $<100$  genomic features identified as significantly affecting growth in a single serovar, the task becomes extremely time consuming when faced with the  $\sim 1000$  genes potentially involved in macrophage infection in *S. Typhimurium*. Hence an alternative, automated approach is required, at least for a first scan of the data.



**Figure 3.5: Smear plot of logFC in mutant prevalences over macrophage infection in *S. Typhi*.** Each point in this plot represents a tested genomic feature. LogFC is reported on the Y-axis, logCPM on the X-axis; statistically significant features at a FDR of 0.1 are in red. The blue lines represent logFCs of  $|2|$ , translating to a four-fold change in mutant prevalences between infection and control. Negative values indicate a reduction over infection in mutant prevalence, positive values an increase.

A number of resources exist which could provide a basis for this sort of automated functional analysis of high-throughput experimental data. These include the Gene Ontology (GO) (Gene Ontology Consortium, 2013), MetaCyc (Caspi et al., 2012), TIGRFAM and Genome Properties (Haft et al., 2013), and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2012). Each of these databases has different goals in its curation, and their own unique advantages and disadvantages. For instance, TIGRFAM provides hidden Markov models (HMMs) with attached pathway information, which can be used to annotate pathways and subsystems present in a genome in the absence of annotation. MetaCyc provides similar resources, including tools for filling ‘hole’ in pathways and subsystems annotated in a genome, based on large manually curated pathway databases covering much of the diversity of life. However, here I have chosen to use the KEGG database as the basis for my analysis, as it is relatively comprehensive, contains annotations for both of the serovars being studied here, and has

a readily available R interface.

Many techniques have been developed for purposes of pathway analysis (Khatri, Sirota, and Butte, 2012), however, few of them have readily available implementations and many of those that do are tailored to Eukaryotic data. So instead of using a previously developed method, I implemented what I have previously described as a ‘walking hypergeometric test’ (Croucher, Harris, et al., 2012) in the context of determining the effect of sequence identity on recombination in *Streptococcus pneumoniae*. In a standard hypergeometric test, one labels genes as being members of a category (in this case a pathway or subsystem), then asks if a random draw of the same size as the significant gene set were taken without replacement, whether one would expect to draw this many (or more) labelled genes by chance. The walking hypergeometric test extends this by walking down an ordered list, in this case sorted by logFC, and performing a hypergeometric test for category enrichment at each entry. This technique is inspired by the test used in the Sylamer microRNA target prediction tool (Dongen, Abreu-Goodger, and Enright, 2008), itself inspired by GSEA (Subramanian et al., 2005). An illustration of this test can be seen in figure 3.6.

This method has a number of important advantages over traditional gene enrichment testing. Normally, one would first choose significance cut-offs based on a p-value and logFC, then perform a hypergeometric test using the resulting set as the draw. This can fail to detect enriched categories if the size of the draw is large, as in the case of *S. Typhimurium* here. Additionally, these cut-off are by their nature somewhat arbitrary, and it is possible that a large number of genes with individually (statistically) non-significant effects could be representative of a (biologically) significant effect on an entire pathway or subsystem. Finally, this method also provides an intuitive graphical representation of the test statistic, which allows the viewer to understand the distribution of gene categories in the data.

## 3.4 Results and Discussion

I applied the walking hypergeometric test to *S. Typhimurium* in order to discover pathways and subsystems involved in the infective process of this organism. Six pathways were found to be significantly enriched in genes with mutant populations undergoing either expansion or contraction during macrophage infection, summarized in table 3.3. The pathways with significantly expanding mutant populations were LPS biosynthesis



**Figure 3.6: Walking hypergeometric test for depletion of insertions in the *S. Typhimurium* flagellar subsystem.** The x-axis shows the index of genes sorted on logFC from highest (enrichment in insertions over the experiment) on the left to lowest (depleted in insertions over the experiment). The y-axis shows the  $-\log_{10}$  p-value derived from a hypergeometric test at each gene for a higher than expected number of genes in the flagellar subsystem to the right of that position. The red line at index 2839 indicates the position of the minimum p-value of  $\sim 6.7 \times 10^{-9}$ .

and purine metabolism. LPSs are well known to be antigenic, and in fact are commonly used to activate macrophages for infection assays in the laboratory. It seems likely that mutants defective in LPS biosynthesis are able to survive better due to a reduction in the inflammatory response provoked in the host cell. The SCV is known to be limited in purines (Eriksson et al., 2003; Hautefort et al., 2008), so mutants which do not waste resources synthesizing genes involved in purine metabolism would also have a selective advantage.

Flagellar assembly, bacterial secretion systems, and RNA degradation on the other hand were all found to be enriched in genes with contracting mutant populations. Flagellar assembly is particularly striking (figure 3.7), with 28 of 34 genes in the subsystem exhibit negative logFCs over macrophage invasion. Interestingly, three genes in this subsystem exhibited statistically significant positive fold changes. Most strongly among these

**Table 3.3:** *S. Typhimurium* pathways putatively involved in macrophage infection.

Pathways and subsystems with a walking hypergeometric minimum p-value less than  $1 \times 10^{-3}$ . Table columns as follows: 1, pathway description; 2, identified as being relatively enriched or depleted in mutants; 3, minimum p-value from walking hypergeometric test; 4, number of genes in pathway significant enriched/depleted in mutants; 5, number of genes in pathway with significantly different logFCs compared to *S. Typhi*.

KEGG Pathway	Enriched/Depleted	P-value	Genes	Different from Typhi
Lipopolysaccharide Biosynthesis	Enriched	$6.67 \times 10^{-6}$	6	9
Purine Metabolism	Enriched	$9.51 \times 10^{-6}$	10	11
Flagellar Assembly	Depleted	$6.72 \times 10^{-9}$	16	19
Bacterial Secretion System	Depleted	$3.21 \times 10^{-4}$	13	17
RNA Degradation	Depleted	$2.65 \times 10^{-4}$	4	8

was *fliT*, which is known to produce a hyperflagellated phenotype in deletion mutants (Yokoseki et al., 1995). It seems likely that other genes in this subsystem with expanding mutant populations produce similar paradoxical effects. Flagella have long been known to be important for *S. Typhimurium* infection of macrophage (Weinstein et al., 1984; Bäumler, Kusters, et al., 1994; Schmitt et al., 2001), and our results agree.

Bacterial secretion systems were also enriched in contracting mutant populations, see tables 3.3 and 3.4. Most prominent among these were SPI-1 and SPI-2 T3SSs, known to be involved in invasion of and survival in macrophage, respectively. SPI-2 genes had relatively low mutant prevalences in our initial library, likely due to the exclusion of transposase by the nucleoid-forming protein H-NS (see chapter 2); despite this SPI-2 genes are still enriched in contracting mutant populations, and the effect of SPI-2 genes on macrophage infection is likely underestimated by these results. Additionally this KEGG system does not include many of the effector proteins secreted by these T3SSs, so again their effect is likely to be underestimated. Finally, the RNA degradation system was also enriched in contracting mutant populations. The four genes in this system found to be significantly depleted in mutants after the infection assay were *pcnB*, involved in polyadenylation of transcripts; *hfq*, involved in the activity of bacterial sRNAs; *dnaK*, a molecular folding chaperone implicated in heat-shock responses; and *rnr*, encoding RNase R, a component of the bacterial RNA degradosome. This underlines the importance of RNA-based regulation to the infective process in *S. Typhimurium*, an emerging theme in infection biology (Hebrard et al., 2012).

Overall, the picture emerging from this high-level analysis of the *S. Typhimurium* macrophage infection assay recapitulates much of what is already known from the literature



**Figure 3.7: Mutant depletion in the *S. Typhimurium* flagellar subsystem.** Genes in grey are relatively depleted in mutants over infection, while genes in red have mutant populations that have expanded. Figure adapted from the KEGG database (Kanehisa et al., 2012).

(see figure 3.1). It is an active process, involving flagella, manipulation of host cells through the SPI-1 and -2 T3SSs, and rapid RNA-based regulatory changes in response to the changing conditions during infection, and induces large changes in the population structure of our mutant library. In contrast, the structure of the *S. Typhi* library after infection is almost entirely unchanged (see figure 3.5). Perhaps most interestingly, the genes of the SPI-1 and -2 T3SSs displayed no significant differences in mutant prevalence, and were all significantly different in behavior from the same genes in *S. Typhimurium* (see table 3.4). This confirms a previous controversial study (Forest et al., 2010) which claimed that SPI-2 had no effect on *S. Typhi* survival in macrophage. In addition SPI-1 does not appear to have an appreciable effect on macrophage entry or survival, suggesting uptake through phagocytosis as a primary entry mechanism. As only 28 genes were significantly changed in mutant prevalence over the assay (table 3.5), I did not perform a pathway analysis and instead examined genes individually. A comparatively small

**Table 3.4: Bacterial secretion system genes implicated in *S. Typhimurium* infection of macrophages.** Genes in the KEGG bacterial secretion system category with statistically significant changes in mutant prevalence over macrophage infection. Columns: 1, SL1344 gene ID; 2, gene name; 3, SPI gene resides in; 4, logFC over *S. Typhimurium* infection of macrophage, negative values indicate a contraction of the mutant population, positive an expansion; 5, adjusted p-value for difference to *S. Typhimurium* control; 6, logFC between *S. Typhi* and *S. Typhimurium* experiments, negative values indicate faster contraction of the *S. Typhimurium* population and/or expansion of the *S. Typhi* population and vice versa; 7, p-value for difference from *S. Typhi* logFC.

SL ID	Name	SPI	logFC	P-value	Δ logFC	Δ P-value
SL1343	ssaJ	SPI-2	-4.65	$1.29 \times 10^{-7}$	-4.83	$1.79 \times 10^{-8}$
SL1355	ssaT	SPI-2	-4.46	$1.41 \times 10^{-4}$	-5.04	$7.55 \times 10^{-6}$
SL2868	spaQ	SPI-1	-4.10	$2.37 \times 10^{-6}$	-3.95	$5.55 \times 10^{-6}$
SL2650	ffh	N/A	-4.04	$4.28 \times 10^{-2}$	-2.11	$1.43 \times 10^{-1}$
SL2869	spaP	SPI-1	-3.30	$3.75 \times 10^{-15}$	-3.09	$1.80 \times 10^{-11}$
SL1352	ssaQ	SPI-2	-2.33	$4.68 \times 10^{-8}$	-2.38	$2.80 \times 10^{-7}$
SL1353	SL1353	SPI-2	-2.30	$1.40 \times 10^{-2}$	-2.48	$6.99 \times 10^{-3}$
SL1354	ssaS	SPI-2	-2.18	$2.60 \times 10^{-2}$	-2.28	$1.53 \times 10^{-2}$
SL2867	spaR	SPI-1	-1.93	$2.50 \times 10^{-7}$	-1.84	$6.16 \times 10^{-6}$
SL2853	prgI	SPI-1	-1.72	$9.23 \times 10^{-3}$	-1.65	$1.21 \times 10^{-2}$
SL2870	spaO	SPI-1	-1.70	$2.62 \times 10^{-5}$	-1.91	$9.31 \times 10^{-6}$
SL2876	invE	SPI-1	-1.67	$1.36 \times 10^{-3}$	-1.61	$2.80 \times 10^{-3}$
SL2873	SL2873	SPI-1	-1.25	$9.00 \times 10^{-3}$	-1.12	$2.01 \times 10^{-2}$
SL3928	tatB	N/A	1.45	$5.99 \times 10^{-2}$	2.73	$9.11 \times 10^{-4}$
SL1340	ssaG	SPI-2	1.51	$7.97 \times 10^{-2}$	1.47	$8.95 \times 10^{-2}$
SL1328	SL1328	SPI-2	1.82	$9.68 \times 10^{-2}$	1.39	$2.04 \times 10^{-1}$
SL3264	secG	N/A	2.67	$6.16 \times 10^{-3}$	3.70	$5.26 \times 10^{-4}$

number of genes appear to be actively selected for or against in the assay; however, this belies the broad effects these relatively few differences may have on the phenotypes exhibited by the population.

As in *S. Typhimurium*, a disproportionate number of genes with significant changes in mutant prevalence were involved in surface antigen and LPS biosynthesis: *rfaH*, *wecG*, *wecC*, *wecB*, *waaG*, *waaP*, *waaI*, and *waaJ*. However, in contrast to *S. Typhimurium*, many of the contracting mutant populations were in genes involved in biosynthesis of the enterobacterial common antigen (ECA). *S. Typhimurium* ECA mutants have previously

**Table 3.5: Genes putatively involved in *S. Typhi* infection of macrophages.** Genes with statistically significant changes in mutant prevalence over macrophage infection. See text for a discussion of gene function. Columns: 1, TY2 gene ID; 2, gene name; 3, logFC over *S. Typhi* infection of macrophage, negative values indicate a contraction of the mutant population, positive an expansion; 4, adjusted p-value for difference to *S. Typhi* control; 6, logFC between *S. Typhi* and *S. Typhimurium* experiments, negative values indicate faster contraction of the *S. Typhimurium* population and/or expansion of the *S. Typhi* population and vice versa; 7, p-value for difference from *S. Typhimurium* logFC.

Ty2 ID	Name	logFC	P-value	$\Delta$ logFC	$\Delta$ P-value
t0540	nuoF	-4.50	$2.01 \times 10^{-4}$	4.79	$8.23 \times 10^{-4}$
t1033	prc	-1.77	$4.70 \times 10^{-4}$	-2.28	$2.77 \times 10^{-3}$
t1038	yobG	-3.13	$4.63 \times 10^{-4}$	-0.10	$8.81 \times 10^{-1}$
t1662	hns	1.60	$2.26 \times 10^{-5}$	-3.48	$5.26 \times 10^{-6}$
t2312	t2312	1.64	$9.41 \times 10^{-6}$	-3.37	$2.87 \times 10^{-9}$
t2313	fimY	-1.49	$2.30 \times 10^{-4}$	1.83	$1.63 \times 10^{-3}$
t2317	fimD	-1.56	$2.30 \times 10^{-4}$	2.13	$4.02 \times 10^{-4}$
t2695	stpA	-1.56	$2.30 \times 10^{-4}$	-3.08	$2.81 \times 10^{-7}$
t2961	dsbC	-2.65	$3.44 \times 10^{-4}$	2.90	$3.85 \times 10^{-3}$
t2980	serA	-2.38	$1.76 \times 10^{-4}$	4.04	$2.17 \times 10^{-6}$
t3252	yhcH	2.05	$2.60 \times 10^{-11}$	-2.96	$2.05 \times 10^{-11}$
t3264	degQ	-1.48	$1.25 \times 10^{-4}$	2.53	$1.79 \times 10^{-6}$
t3320	rfaH	3.35	$4.51 \times 10^{-7}$	-1.54	$7.51 \times 10^{-2}$
t3368	wecG	-3.18	$5.34 \times 10^{-6}$	2.40	$5.56 \times 10^{-3}$
t3376	wecC	-1.59	$4.63 \times 10^{-4}$	2.02	$1.52 \times 10^{-3}$
t3377	wecB	-1.70	$3.27 \times 10^{-5}$	0.50	$3.67 \times 10^{-1}$
t3500	oxyR	1.92	$4.70 \times 10^{-4}$	-1.73	$2.98 \times 10^{-2}$
t3623	dsbA	-5.05	$1.91 \times 10^{-6}$	6.66	$9.26 \times 10^{-8}$
t3634	rbsK	-1.92	$4.70 \times 10^{-4}$	2.36	$2.30 \times 10^{-3}$
t3645	gidA	-2.89	$4.50 \times 10^{-10}$	1.61	$7.33 \times 10^{-3}$
t3677	mnmE	-2.15	$7.61 \times 10^{-7}$	1.71	$3.49 \times 10^{-3}$
t3796	waaG	-2.37	$9.01 \times 10^{-8}$	6.28	$1.22 \times 10^{-27}$
t3797	waaP	-3.36	$2.37 \times 10^{-7}$	1.63	$4.02 \times 10^{-2}$
t3801	waaI	3.41	$1.99 \times 10^{-36}$	-1.60	$1.34 \times 10^{-5}$
t3802	waaJ	1.37	$1.36 \times 10^{-4}$	0.40	$4.30 \times 10^{-1}$
t4179	actP	1.43	$7.08 \times 10^{-10}$	-0.95	$3.80 \times 10^{-3}$
t4411	miaA	-2.68	$8.39 \times 10^{-5}$	1.25	$1.29 \times 10^{-1}$
t4488	treC	-3.27	$2.10 \times 10^{-6}$	3.91	$1.97 \times 10^{-6}$

been shown to not cause acute disease in mice, though they are capable of persistently colonizing the spleen and liver (Gilbreath et al., 2012), reminiscent of *S. Typhi* infections of silent carriers, though the relevance of this to *S. Typhi* infection of macrophages is unclear. The most interesting of the LPS biosynthesis genes affected, *rfaH*, is a anti-termination factor affecting primarily LPS biosynthesis loci (Artimovitch and Landick, 2002; Santangelo and Roberts, 2002) with a >8-fold increase in mutant population size over the assay. This anti-termination factor associates with RNA polymerase and prevents pausing at both Rho-dependent and Rho-independent transcriptional terminators, promoting the expression of promoter-distal loci. As a result, a mutation in this single gene is likely to have broad pleiotropic effects, a feature common to many of the genes implicated in *S. Typhi* infection.

Other examples of genes with potentially wide-ranging pleiotropic effects include the paralogous nucleoid-forming proteins *hns* and *stpA*. H-NS has been described previously in chapter 2, but briefly it acts to condense DNA by binding to AT-rich, bent regions, and primarily regulates virulence and stress-response loci (Navarre et al., 2006; Lucchini, Rowley, et al., 2006). The paralogous StpA has similar binding affinity, but regulates a reduced regulon compared to H-NS, and in fact *hns* masks the phenotypic effects of an *stpA* deletion (Lucchini, McDermott, et al., 2009). The expansion of the population containing *hns* mutations with the concomitant contraction of *stpA* mutant populations suggests a subtle interplay between the two at work under infective conditions, which potentially wide repercussions for cellular phenotype. Another example of this theme of *S. Typhi* relying on genes with pleiotropic effects is given by *gidA* and *mnmE*. The products of these genes act together to post-transcriptionally modify a number of tRNAs (Yim et al., 2006), affecting translational fidelity (Brégeon et al., 2001). Mutations in these genes can have global effects (Kinscherf and Willis, 2002), and have recently been shown to affect *S. Typhimurium* virulence (Shippy et al., 2013).

While *hns*, *stpA*, *gidA*, and *mnmE* modulate gene expression at the transcriptional and post-transcriptional level, two other genes with depleted mutant populations, *dsbA* and *dsbC*, likely induce effects post-translationally. The *dsb* genes are involved in disulfide bond (DSB) formation, which is required for the proper folding and function of a wide range of proteins, and is known to be required for virulence in a number of bacterial species including *Shigella flexneri* and uropathogenic *Escherichia coli* (Heras et al., 2009). DsbA catalyzes the formation of DSBs in newly translated proteins translocated periplasm;

however, this process is non-specific and introduces spurious bonds. DsbC provides a proof-reading mechanism through isomerization of non-native bonds introduced by DsbA. This process is critical to the expression of a wide range of virulence factors in many species, including toxins and fimbriae (Yu and Kroll, 1999). DsbA expression is known to affect *S. flexneri* survival in macrophage (Yu, Oragui, et al., 2001), and it appears to affect virulence in *S. Typhimurium*, though this is thought to be mediated through its effects on the SPI-2 T3SS (Miki, Okada, and Danbara, 2004). The exact mechanism through which the DSB system affects *S. Typhi* survival in macrophage is unclear, though it appears likely that DSB formation is important for extracellular or cell-surface structures *S. Typhi* uses to interact with the host cell.

In conclusion, the picture that emerges from this analysis is that unlike *S. Typhimurium*, *S. Typhi* is robust to assault from a human macrophage host cell. Infection produces only small changes in the population structure of the *S. Typhi* mutant library, and those populations which are affected have mutations in genes causing broad pleiotropic effects which can not help but have a strong effect on phenotype. This suggests that *S. Typhi* is already tuned to maintain homeostasis within human macrophages, indicative of its extreme adaptation to its host. While I have only performed a systematic analysis of the orthologous genes present in both *S. Typhimurium* and *S. Typhi* here, I have also examined the effect of macrophage infection on the genes involved in synthesis of the Vi antigen, which may be responsible for some of the differences exhibited between the two serovars. This capsular antigen confers a protective effect on *S. Typhi* in macrophage (Hirose et al., 1997), and as expected mutations in these genes were not well tolerated. It appears that *S. Typhi*, with the help of its capsule, adopts a stealth phenotype whereby it can enter and replicate within macrophage unmolested. *S. Typhimurium*, on the other hand, uses its flagella and SPI-encoded T3SSs to actively invade the macrophage, and the toll of this combat can be seen in the effects on the mutant population. I am currently working with Prof. John Wain (University of East Anglia) to procure microscopy of *S. Typhimurium* and *S. Typhi* infection of macrophage to confirm and build on these results.

While I have developed the methods presented in this chapter specifically to deal with this study, they are broadly generalizable to any transposon-insertion sequencing study. I am assisting in applying them to a number of TraDIS studies in a wide range of organisms, including carbon source utilization in *S. Typhimurium* and *S. Enteritidis*;

twitching motility in *Pseudomonas aeruginosa*; whole animal infection in *Citrobacter*, *Salmonella*, and *Escherichia* strains; and drug resistance in *Klebsiella pneumoniae*. As I have shown here, with the proper analytical tools TraDIS can be a powerful technique for the rapid generation of functional hypotheses about gene function in complex processes.



# Chapter 4

## Detecting Rho-independent terminators in genomic sequence with covariance models

*Portions of this chapter are based on the previously published article “RNIE: genome-wide prediction of bacterial intrinsic terminators” (Gardner, Barquist, et al., 2011). This work is the result of collaboration with Paul P. Gardner (Wellcome Trust Sanger Institute/University of Canterbury).*

### 4.1 Introduction

Bacteria are thought to utilize two major systems for transcriptional termination: Rho-dependent termination, and Rho-independent or intrinsic termination (Peters, Vangeloff, and Landick, 2011). Rho-dependent termination relies on a protein cofactor, Rho, a homohexameric ring protein that threads its way along the newly synthesized RNA molecule before causing RNA polymerase (RNAP) to dissociate at poorly defined pause sites. Intrinsic termination on the other hand, depends primarily on the biophysical characteristics of the sequence being transcribed. The detection of these intrinsic terminators in genomic sequence is the subject of this chapter. This chapter will serve largely as background and motivation for the next, in which I develop computational methods for identifying and characterizing transcriptional termination motifs across the

bacterial phylogeny.

#### 4.1.1 Rho-independent termination

Intrinsic termination is mediated by short structured RNA motifs known as Rho-independent terminators (RITs). These are generally characterized by a G+C-rich hairpin followed by a tract of T (as DNA) / U (as RNA) residues. As RNAP transcribes the poly-U tract it pauses, possibly with assistance from the partially formed hairpin structure, allowing full nucleation of the hairpin which melts weak rU-dA bonds within the elongation complex and leads to dissociation of RNAP (Peters, Vangeloff, and Landick, 2011), see figure 4.1. This process is somewhat stochastic, and the probability of successful transcription termination depends on various features of the RIT including stem composition and length, loop composition, length of the poly-U tract, and the sequence context of the element (Larson et al., 2008; Cambray et al., 2013; Chen, Liu, et al., 2013). As is well known from the study of transcriptional attenuators and riboswitches (Henkin and Yanofsky, 2002; Barrick and Breaker, 2007; Naville and Gautheret, 2010), alternative structures formed upstream of the RIT can also affect termination efficiency, and force exerted on the upstream sequence can increase termination efficiency even in the absence of obvious alternative structures (Larson et al., 2008).



**Figure 4.1: Rho-independent termination.** A) The RNA polymerase traverses the DNA template strand from 3' to 5', synthesizing the nascent RNA molecule. B) As the polymerase nears a termination site, a G+C-rich terminator stem sequence (boxed) is transcribed. C) Formation of a hairpin structure causes the polymerase to pause, and together with a string of unstable rU-dA bonds causes the polymerase to release from the template. Reproduced from Gardner, Barquist, et al. (2011).

The degree to which bacteria rely on intrinsic termination varies widely. A bioinformatic analysis examining the computationally predicted minimum free energy (MFE) of gene terminuses showed that while some species display an enrichment of strong RNA

secondary structure potential at the 3' ends of genes, others do not (Washio, Sasayama, and Tomita, 1998). Mutagenesis studies support this conclusion: while Rho is essential in some genomes with fewer apparent intrinsic terminators (for instance, *Salmonella enterica*, see table 2.1), it is dispensable in others that are more heavily dependent on intrinsic termination, such as *Bacillus subtilis* (Quirk et al., 1993). This suggests competition between the two termination systems, leading to clade-specific skews in RIT utilization (Carafa, Brody, and Thermes, 1990; Kröger and Wahl, 1998; Hoon et al., 2005). The accurate prediction of these elements is critical to understanding the regulation of transcription, particularly in light of the  $\geq 3000$  completed bacterial genomes currently deposited in EMBL-bank. In addition to their obvious role in helping to define operon structures in genomic sequence (Salgado et al., 2013), they can also be important indicators of cis-RNA regulation (Henkin and Yanofsky, 2002; Barrick and Breaker, 2007; Naville and Gautheret, 2010). Finally, the importance of RITs in designing synthetic genetic circuits has recently been recognized, and this has driven studies attempting to broaden our understanding of the factors affecting intrinsic termination efficiency (Cambray et al., 2013; Chen, Liu, et al., 2013).

#### 4.1.2 Previous approaches to identifying intrinsic terminators

Two main approaches to detecting RITs have been taken over the years: RNA motif descriptors, both expertly constructed (Lesnik et al., 2001) and automatically generated (Naville, Ghuillot-Gaudeffroy, et al., 2011); and thermodynamic models of RNA folding to detect hairpins paired with a heuristic scoring scheme for the poly-T tail region (Ermolaeva et al., 2000; Wan and Xu, 2005; Wan, Lin, and Xu, 2006; Kingsford, Ayanbule, and Salzberg, 2007). Arguably the most popular of these methods has been TransTermHP (Kingsford, Ayanbule, and Salzberg, 2007), an example of the second approach.

The TransTermHP algorithm takes a windowed approach to detecting RITs (figure 4.2). In order to avoid the computational cost of predicting local secondary structure across the entire genome, TransTermHP first scans overlapping windows of 6 bases for those containing at least 3 T residues. Upon finding such a window, TransTermHP performs a dynamic programming procedure to predict potential hairpin structures, using a simplified version of the Zuker algorithm (Zuker and Stiegler, 1981) for *in silico* RNA folding parameterized using a set of experimentally validated *Escherichia coli* RITs



**Figure 4.2: TransTermHP motif.** Schematic of the terminator motif that TransTermHP searches for. The terminators consist of a short stem-loop hairpin followed by a thymine-rich region on their 3' side. TransTermHP is generally restricted to find terminators where each side of the stem is  $\geq 4$  nt, the length of the loop is  $\geq 3$  nt and  $\leq 13$  nt, and the total length of the stem-loop is  $\leq 59$  nt. Reproduced under a Creative Commons Attribution License (CCAL) from Kingsford, Ayanbule, and Salzberg (2007).

(Ermolaeva et al., 2000). This is then combined with a heuristic score for the quality of the poly-T tail (Carafa, Brody, and Thermes, 1990) which rewards T residues occurring closer to the closing base-pairs of the predicted hairpin structure. Candidate RITs are then filtered on stem length, loop length, and total length (see the caption of figure 4.2 for details). Finally, the combined score of surviving candidates is compared to the scores of predicted terminators in random sequence with similar GC content to that of the target genome to provide a measure of prediction quality. Search is apparently also limited to regions surrounding stop codons (Kingsford, Ayanbule, and Salzberg, 2007; see also the discussion of the beta benchmark below), though the exact boundaries on the search space are not explicitly given in the TransTermHP documentation or publication.

This methodology presents a number of problems. First, while the thermodynamic method used to predict hairpin structures likely places some implicit restrictions on the sequence composition of the hairpin structure, it does not explicitly model conservation of residue composition across terminators. Conservation of residue composition could arise due to convergent evolution of terminator structures under selection for properties that promote strong termination in the host species, or as the results of genuine homology between RITs due to their descent from a common ancestor deposited by transposable

elements, as has previously been hypothesized (Naville and Gautheret, 2010). In addition, windowed searching for and heuristic scoring of the poly-T tail is unlikely to accurately capture the true constraints on this feature. We show here that explicitly modelling residue conservation improves detection of RITs. Secondly, the comparison to random sequence with similar GC content is unlikely to be an adequate control: it has been shown that considering dinucleotide content of sequences is critical to determining the significance of their secondary structure (Workman and Krogh, 1999). Though the method of generating random sequence is not explicitly stated in Kingsford, Ayanbule, and Salzberg (2007), it seems unlikely that it was the product of dinucleotide shuffling or a first-order Markov chain, as would be required to preserve dinucleotide frequencies. In fact TransTermHP does not appear to consider base-stacking effects in its predictions whatsoever. Finally, restricting search to the regions around annotated gene terminuses, or rewarding candidate RITs for being in these regions, is both somewhat artificial and requires accurate gene annotation, which remains a challenge.

### 4.1.3 Covariance models

Our method, RNIE, overcomes many of the problems in previous RIT detection methods through the use of *covariance models* (CMs), a special case of stochastic context-free grammars (Eddy and Durbin, 1994; Sakakibara et al., 1994). CMs are sophisticated statistical models which incorporate information about both sequence and secondary structure conservation. They are perhaps most easily understood through analogy to the simpler profile hidden Markov models (HMMs) (Eddy, 1998). A typical method for HMM construction takes as its input an alignment of sequences. For each column of the alignment, a *node* is constructed, consisting of three *states*: match, insert, and delete. The match state models the residue distribution at that alignment position, while the insert and delete states model the probability and length distributions of insertions and deletions beginning at this column, respectively. The mathematics of HMMs have been well explored, and efficient dynamic programming algorithms exist for training (the Baum-Welch algorithm), assigning a probability to a sequence being produced by the model (the Forward algorithm), and finding the most probable parsing of a sequence (the Viterbi algorithm).

CMs are similar to profile HMMs, with the extension that they can additionally



**Figure 4.3: Covariance model architecture.** A) A toy multiple alignment of three RNA sequences, with 28 total columns, 24 of which will be modeled as consensus positions. The [structure] line annotates the consensus secondary structure: angle brackets mark base pairs, colons mark consensus single-stranded positions, and periods mark insert columns that will not be considered part of the consensus model because more than half the sequences in these columns contain gaps. B) The structure of one sequence from A, the same structure with positions numbered according to alignment columns, and the guide tree of nodes corresponding to that structure, with alignment column indices assigned to nodes (for example, node 5, a MATP match-pair node, will model the consensus base pair between columns 4 and 14). C) The state topology of three selected nodes of the CM, for two MATP nodes and one consensus leftwise single residue bulge node (MATL, match-left). The consensus pair and singlet states (two MPs and one ML) are white, and the insertion/deletion states are gray. State transitions are indicated by arrows. Reproduced under a Creative Commons Attribution License (CCAL) from Nawrocki and Eddy (2007).

model dependence between alignment positions (see figure 4.3); rather than nodes being constructed from alignment columns, they are built from structural elements, i.e. pairing bases, annotated in the alignment (figure 4.3B). This increases the complexity of node architecture (figure 4.3C), as each node must now contain states to match both bases in a pair, either one of a pair individually if its partner has been deleted, insertions on either side of the pair, and base pair deletions. Analogs to the Baum-Welch, Forward, and Viterbi algorithms exist for CMs: expectation-maximization using the inside-outside algorithm, the inside algorithm, and the Cocke-Younger-Kasami (CYK)

algorithm, respectively. Unfortunately, modeling the dependence between positions, that is moving from a regular grammar such as an HMM to a context-free grammar such as a CM, comes at a considerable computational cost due to the restrictions imposed by the Chomsky hierarchy (Chomsky, 1959), roughly equivalent to adding an additional dimension to the dynamic programming matrix. In this study we have used the Infernal package (Nawrocki, Kolbe, and Eddy, 2009), which implements CMs and associated algorithms for RNA sequence analysis, and includes a number of heuristics for increasing the speed of CM-based searches including adaptive banding of the dynamic programming matrix (Nawrocki and Eddy, 2007) and HMM pre-filters based on HMMER (Eddy, 2011). Importantly, Infernal also incorporates a null model for scoring sequence hits; for sequence that matches the CM, the probability of this match is compared to the probability of a match to the null model. This comparison is expressed as a  $\log_2$  odds ratio, or bitscore, and from this further statistics, such as an expect value (E-value), can be calculated. Covariance models are widely used in RNA homology search, most notably by the Rfam database (Burge et al., 2013) and the tRNAscan-SE tool (Lowe and Eddy, 1997) for predicting tRNAs in genomic sequence.

## 4.2 Methods

*Paul P. Gardner implemented and benchmarked the RNIE tool. Eric P. Nawrocki (Howard Hughes Medical Institute Janelia Farm Research Campus) assisted in optimizing Infernal parameters for RIT search. Zasha Weinberg (Howard Hughes Medical Institute/Yale University) ran the Rnall and Rnall-Brkr algorithms for benchmarking. I designed and implemented the analysis which lead to the discovery of the putative mycobacterial termination motif.*

### 4.2.1 Construction of a covariance model for Rho-independent terminators

One hundred seventy-one and 891 experimentally validated RIT sequences from *Escherichia coli* and *Bacillus subtilis*, respectively, were collected from the *E. coli* Database Collection (ECDC; Wahl and Kröger, 1995) and the supplementary information of Hoon et al. (2005) and manually curated based on evidence quality, leaving a set of 981 RIT

sequences. These sequences were subjected to iterative rounds of alignment, structure prediction, homology search and refinement. Alignments and secondary structures were predicted using WAR (Torarinsson and Lindgreen, 2008), CMfinder (Yao, Weinberg, and Ruzzo, 2006), and MLocarna (Will et al., 2007), iteratively refined using Infernal (Nawrocki, Kolbe, and Eddy, 2009), then manually refined using the RALEE emacs environment (Griffiths-Jones, 2005). Sequence searches were performed using the resulting CM against EMBL with the Rfam pipeline (Gardner, Daub, et al., 2009), and additional sequences were incorporated in to the alignment based on the following criteria: i) the maximum similarity to an existing seed sequence had to be 95% and the minimum 60%, ii) the minimum fraction of canonical base pairs had to be 75%, iii) the sequence annotation should not contain terms like contaminant, pseudogene, repeat or transposon and iv) they must score above a bitscore threshold of 20. These additional sequences were then manually checked for their position near a gene terminus. This resulted in 1117 aligned sequences, which were further split in to two groups based on how well they matched the resulting CM. Those scoring with a bitscore over 14 were placed in alignment A, those scoring less were placed in alignment B. These alignments were then again automatically refined using Infernal before a final round of manual refinement.

#### 4.2.2 RNIE run modes

As described in the introduction, algorithms for performing inference with CMs can be very slow, and as a result Infernal implements a number of filters to reduce the number of sequences which proceed to a full CM-based homology search. In response to this, two modes for RNIE were developed: genome mode meant for large-scale searches, which enables HMM filters and adaptive banding and uses the CYK algorithm with a higher threshold for reported RIT predictions; and gene mode meant for annotation of relatively short sequence regions, which disables Infernal's filtering mechanisms and uses the slower but more powerful inside algorithm with a lower threshold for reporting RIT predictions. Genome mode scans sequence at  $\sim$ 43 kb/s with a low false positive rate of  $\sim$ 1.7 FP/Mb. The sensitivity, positive predictive value and Matthews' correlation coefficient for this mode (determined in the alpha benchmark below) are 0.70, 0.79 and 0.74. Gene mode scans at  $\sim$ 1kb/s, and the false positive rate, positive predictive value and Matthews' correlation coefficient are  $\sim$ 9.6 FP/Mb, 0.45 and 0.61, respectively. The

Infernal parameters used for genome and gene mode, respectively, are

```
cmsearch -T 16 -g --fil-no-qdb --fil-T-hmm 2  
--cyk --beta 0.05 CM query_sequence.fasta
```

```
cmsearch -T 14 -g --fil-no-qdb --fil-no-hmm  
--no-qdb --inside CM query_sequence.fasta
```

### 4.2.3 Definitions

For the purposes of benchmarking, the following measures were used

$$Sensitivity = \frac{TP}{TP + FN}$$

$$PPV = \frac{TP}{TP + FP}$$

$$FPR = \frac{FP}{\text{Total length in kb}}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where any prediction that covered a known RIT by at least one nucleotide was considered a true positive (TP), any prediction that did not overlap a known RIT was considered a false positive (FP), a missed RIT was considered a false negative (FN), and the number of unclassified, non-RIT sequence were considered true negatives (TN).

**Table 4.1: Control genomes.** Columns: 1) species name, 2) EMBL-bank accession, 3) genome size in megabases, 5) number of CDSs annotated in genome, 6) genome G+C content, 7) number of RNIE predictions in genome mode on native sequence, 8) number of RNIE predictions in genome mode on dinucleotide shuffled sequence, 9) number of RNIE predictions in gene mode on native sequence, 10) number of RNIE predictions in gene mode on dinucleotide shuffled sequence.

Species	EMBL accession	Phylum	Genome size (MB)	CDSs	G+C content	Number of predictions			
						native	shuffled	native	shuffled
<i>Mycobacterium tuberculosis</i>	AE000516	Actinobacteria	4.40	4189	0.66	19	0	111	3
<i>Streptomyces griseus</i>	AP009493	Actinobacteria	8.55	7138	0.72	72	0	353	2
<i>Bacteroides thetaiotaomicron</i>	AF015628	Bacteroidetes	6.26	4778	0.43	783	2	1470	44
<i>Chlamydia pneumoniae</i>	AE001363	Chlamydiae	1.23	1052	0.41	61	3	135	19
<i>Prochlorococcus marinus</i>	AE017126	Cyanobacteria	1.75	1882	0.36	81	5	131	22
<i>Deinococcus radiodurans</i>	AE000513	Deinococcus-Thermus	2.65	2579	0.67	283	0	506	2
<i>Bacillus subtilis</i>	AL009126	Firmicutes	4.22	4245	0.44	1851	4	2540	54
<i>Clostridium difficile</i>	AM180355	Firmicutes	4.29	3777	0.29	431	8	1152	58
<i>Fusobacterium nucleatum</i>	AE00951	Fusobacteria	2.17	2067	0.27	155	1	457	34
<i>Thermodesulfobacter yellowstonii</i>	CP001147	Nitrospirae	2.00	2033	0.34	78	6	176	41
<i>Escherichia coli</i>	U00096	Proteobacteria	4.64	4321	0.51	601	6	1058	35
<i>Helicobacter pylori</i>	AE000511	Proteobacteria	1.67	1566	0.39	28	12	128	61
<i>Salmonella enterica</i>	AB014613	Proteobacteria	4.79	4323	0.52	537	4	980	32
<i>Leptospira interrogans</i>	AB016623	Spirochaetes	4.28	3394	0.35	164	18	375	132
<i>Ureaplasma parvum</i>	AF222894	Tenericutes	0.75	611	0.26	54	0	163	5
<i>Fervidobacterium nodosum</i>	CP000771	Thermotogae	1.95	1750	0.35	409	3	588	28
<i>Methylacidiphilum infernorum</i>	CP000575	Verrucomicrobia	2.29	2472	0.45	50	7	157	52

## 4.3 Results

Benchmarking a tool for RIT detection is challenging. As described in the methods section, only a relatively small number of RITs had been verified at the time of this study. While this situation is beginning to improve with the development of high-throughput techniques for RIT characterization (Cambray et al., 2013; Chen, Liu, et al., 2013), verified RITs are still largely drawn from the model bacteria *E. coli* and *B. subtilis*. As a result, two benchmarks were performed: the first, or alpha, benchmark examines method performance on known RITs, with the caveat that these RITs formed part of the training set for RNIE and many of the other methods tested. The second, or beta, benchmark is a qualitative assessment on whole genomes with unknown RIT contents, evaluating the quality of predictions by their genomic position and estimating the FPR by the relative number of predictions on shuffled sequence.

### 4.3.1 Alpha benchmark

For the first benchmark 485 known RITs, curated on the basis of experimental evidence for activity, were used, drawn from the ECDC (Wahl and Kröger, 1995) and the supplementary information of Hoon et al. (2005). Each RIT was embedded in 1000 bases of randomly selected dinucleotide shuffled sequence drawn from the genomes in table 4.1. For each known RIT a first-order Markov chain was trained on the nucleotide distribution of that sequence and 100 decoy sequences were generated and similarly embedded in 1000 bases of dinucleotide shuffled sequence. A first-order Markov chain was used rather than dinucleotide shuffling of the native RITs, as these short sequences may have a limited number of permutations with identical dinucleotide content. As TransTermHP will only run on annotated sequence, artificial gene annotations were added to each sequence, with either decoys or genuine RITs positioned at the 3' end of one of the annotations.

Four methods besides RNIE were tested (figure 4.4): TransTermHP (with 2, 4, 9, or 10 gene features; Kingsford, Ayanbule, and Salzberg, 2007), RNAmotif (using either the structural score alone (struct), or the structural score augmented with a score for hybridization between the poly-U tail and the DNA sequence (dG); Lesnik et al., 2001), Rnall (using either the score for hairpin formation (dG), or the score for hairpin formation augmented with a score for poly-U/DNA hybridization (hbG); Wan and Xu, 2005; Wan,



**Figure 4.4: Alpha benchmark.** The accuracy of RNIE compared to existing methods of terminator prediction. The left figure shows a ROC plot for four independent methods. The middle figure compares the sensitivity and PPV for the four methods. The figure on the right shows the speeds for each algorithm in kilobases per second. Reproduced from Gardner, Barquist, et al. (2011).

Lin, and Xu, 2006), and a version of Rnall modified by the Breaker lab at Yale University (Rnall-Brkr; using either the score for hairpin formation (dG), or the score for hairpin formation augmented with a score for poly-U/DNA hybridization (hbG); Barrick and Breaker, 2007; Weinberg, Barrick, et al., 2007).

The results of this benchmark show that RNIE’s performance is superior to any previous method for detecting RITs at any level of sensitivity or specificity. Interestingly, all methods which rely on poly-U/DNA hybridization scores performed extremely poorly, suggesting that the understanding of the role of RNA-DNA hybridization in intrinsic termination modelled by these methods is incorrect, or at best incomplete. Of the other methods, the only ones besides RNIE which cross the line  $y = 1 - x$ , the performance of a hypothetical ‘random’ predictor, on the sensitivity versus PPV plot were TransTermHP and RNAMotif. The scanning speed of RNIE in genome mode,  $\sim 43$  kb/s, is comparable to that of TransTermHP at  $\sim 74\text{-}186$  kb/s, depending on the number of gene annotations. Based on these results, thresholds were chosen for reporting RNIE RIT predictions in genome and gene modes at levels slightly below the maximum MCC, that is allowing for a slightly higher FPR in return for increased sensitivity with the assumption that false positives can often be determined by their genomic context.

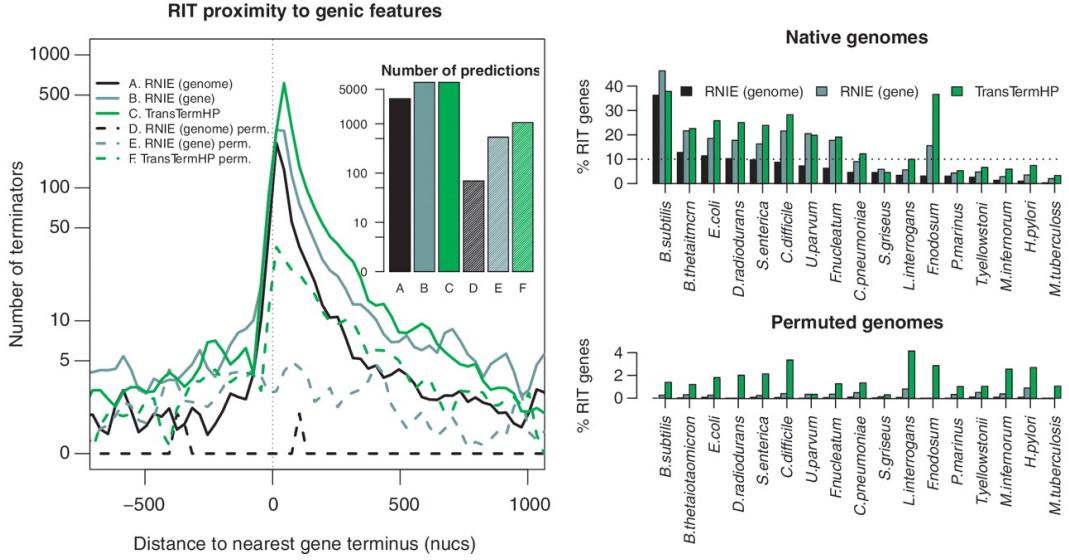
### 4.3.2 Beta benchmark

For the second benchmark 17 genomes representative of the diversity of the bacterial phylogeny (table 4.1) were scanned with both RNIE and TransTermHP, and the results compared. Additionally, dinucleotide shuffles of these genomes were scanned to provide an estimate of the FPR of each method. Genuine RITs are expected to occur preferentially in the 3' region of annotated genes. As can be seen in figure 4.5, predictions for both RNIE and TransTermHP are enriched in predictions 3' to gene annotations (solid lines). RNIE makes relatively few predictions in shuffled sequence (dashed lines), particularly in the more stringent genome mode, and these appear to be randomly distributed with respect to gene terminuses. Worryingly, TransTermHP predictions on dinucleotide shuffled sequence are also enriched at annotated gene terminus, suggesting it is giving a bonus to predictions falling in the correct genomic context. This is particularly problematic, as it suggests a higher FPR in regions where RIT predictions will look most reasonable on a passing inspection.

The bar plots in figure 4.5 report the percentage of genes reported to be terminated by a RIT in each genome by TransTermHP and RNIE. In general, the number of predictions made by RNIE is comparable to TransTermHP, particularly when the higher number of predictions by TransTermHP on shuffled sequence is taken in to account. Interestingly, the only genome where RNIE predicts more RITs than TransTermHP is *B. subtilis*, where most of the training data for the RNIE CMs originated. Additionally, there are a number of genomes where few RITs are predicted by either method. Both of these points will be addressed in more detail in the next chapter.

### 4.3.3 A novel termination motif in *Mycobacterium tuberculosis*

In the course of benchmarking RNIE, we noticed that neither our method nor TransTermHP made many RIT predictions in the *Mycobacterium tuberculosis* genome. While some bacterial lineages are hypothesized not to use intrinsic termination, there is a body of prior work suggesting that *M. tuberculosis* does utilize secondary structure in termination (Washio, Sasayama, and Tomita, 1998; Unniraman, Prakash, and Nagaraja, 2001; Unniraman, Prakash, and Nagaraja, 2002; Mitra, Angamuthu, and Nagaraja,



**Figure 4.5: Beta benchmark.** Ideal terminator predictors will generally produce predictions that are immediately 3' to annotated genes on native sequence and no predictions on shuffled controls. For all the test genomes in Table 1 (excluding *E. coli* and *B. subtilis*), we computed the distance to the nearest 3' genic element, including CDSs, ncRNAs and riboswitches. This was done for both native sequences and dinucleotide shuffled control sequences with corresponding gene annotation transferred to the controls. The figure on the left shows the distribution of distances for RNIE genome and gene modes and for the TransTermHP method. Inset is a barplot showing the total number of predictions for each method on native and shuffled genomes. The figures on the right show the percentage of genes that have a predicted RIT in the region  $-50$  to  $+150$  from an annotated 3'-end of a CDS or ncRNA across all the genome sequences described in Table 1. The upper panel illustrates the results for the native genomes, while the lower panel illustrates results for the permuted genomes. Reproduced from Gardner, Barquist, et al. (2011).

2008; Mitra, Angamuthu, Jayashree, et al., 2009). In particular the Nagaraja group has developed a method, GeSTer, which attempts to classify predicted secondary structures from the terminuses of coding regions into one of five categories of structural motifs. More than 90% of terminal motifs in *M. tuberculosis* fall into their “I-shaped” category, or short stem-loop with no poly-U tail. With this in mind, I developed the following procedure to search for a potential structured termination motif in *M. tuberculosis*.

I extracted 100-nucleotide 3' sequences from the *Mycobacterium tuberculosis* CDC1551, starting 20 bases before annotated CDS ends. Predicted MFE folding scores for each sequence were calculated using RNAfold (Hofacker et al., 1994). I performed a pooled



**Figure 4.6: Putative mycobacterial transcription termination motif.** A) The frequency of TRITs and RITs near the terminal regions of *M. tuberculosis* (EMBL accession: AE000516) genic features. B) The distribution of structural stability derived p-values for the most significant *M. tuberculosis* terminal regions coloured by TRIT (red), RIT (black) or unclassified (blue). C) The secondary structure and sequence conservation of the TRIT motif as displayed by R2R (Weinberg and Breaker, 2011). (D&E) Sequence logos generated for the 5' D) and 3' E) halves of an alignment of the 147 copies of TRIT in the *M. tuberculosis* genome. Reproduced from Gardner, Barquist, et al. (2011).

permutation test for lower than expected MFEs using 1000 dinucleotide shuffles from each 3' sequence. I then ran the CMfinder (Yao, Weinberg, and Ruzzo, 2006) RNA motif-finder over sequences with a p-value less than 0.001. The subsequent alignment was manually refined using the RALEE RNA alignment editor (Griffiths-Jones, 2005). The refined alignment was used to construct an Infernal CM (Nawrocki, Kolbe, and Eddy, 2009), as had been done for canonical RITs, which was then searched across all *Mycobacteria* genomes in the EMBL nucleotide sequence database.

This revealed a well-conserved structured sequence motif associated with gene terminal regions in *Mycobacteria* which we named the tuberculosis Rho-independent terminators, or TRITs, in light of the source of the discovery (see Figure 4.6). TRITs are found across the entire genus, ranging in approximate copy-number from 150 to 250 in *M. abscessus*, *M. avium*, *M. bovis*, *M. gilvum*, *M. intracellulare*, *M. kansasii*, *M. leprae*, *M. marinum*, *M.*

*smegmatis*, *M. tuberculosis*, *M. ulcerans* and *M. vanbaalenii*. The TRITs account for 72% (59/82) of terminal sequences with highly significant secondary structure ( $p < 0.001$ ) in *M. tuberculosis*. TRIT predictions made by our model fall overwhelmingly at the terminus of annotated coding regions, tending to start 8 bases before the annotated gene end (Figure 4.6A), distinct from the distribution of RITs. In addition, TRITs appear to be associated with sharp drops in transcription in RNA-seq experiments (data presented in the next chapter). Additionally, since the publication of this study two sRNA screens in *Mycobacteria* have discovered TRITs apparently terminating sRNA transcription (Miotto et al., 2012; Li, Ng, et al., 2013), providing additional evidence for their activity. The high sequence conservation (Figure 4.6D&E) across elements suggests that this element has either arisen relatively recently, or possibly requires a nucleotide-binding co-factor to perform its function. In the next chapter, I describe a study scaling up this approach to discover transcriptional termination motifs across the entire bacterial phylogeny.

# Chapter 5

## Kingdom-wide discovery of bacterial intrinsic termination motifs

### 5.1 Introduction

As discussed in the previous chapter, intrinsic termination of transcription is a fundamental cellular process in many, if not all, bacterial species. As reviewed in the previous chapter, the bulk of work on intrinsic termination has focused on canonical Rho-independent terminators (RITs), consisting of a G/C-rich hairpin structure followed by a poly-U tail. This is due to both their prevalence in model organisms such as *Escherichia coli* and *Bacillus subtilis*, as well as the distinctiveness of this motif making it an easy target for automated classification.

Despite this focus on canonical RITs, a number of intrinsic terminators which do not rely on a poly-U tail for termination activity are known. These include synthetic constructs derived from canonical RITs (Abe and Aiba, 1996), as well as naturally occurring terminators in *Streptomyces* (Deng, Kieser, and Hopwood, 1987; Neal and Chater, 1991; Ingham, Hunter, and Smith, 1995) and *Mycobacteria* (Unniraman, Prakash, and Nagaraja, 2001). Additionally, a number of ncRNA screens in Actinobacteria have described potential non-canonical RITs terminating ncRNA transcription (Swiercz et al., 2008; Miotto et al., 2012; Li, Ng, et al., 2013). However, a more wide-spread effort at characterization of these elements has been hampered by two factors: their occurrence primarily in non-model organisms such as the Actinobacteria, and a lack of a systematic

classification of these elements making it difficult to determine how wide-spread such elements are. The only study surveying potential alternative intrinsic terminators in the bacterial kingdom relied primarily on categorizing elements based on the shape of their predicted secondary structure (Unniraman, Prakash, and Nagaraja, 2002). However, this fails to consider the large number of very different sequences that can give rise to any particular secondary structure (Schuster et al., 1994). It is well understood from studies of synthetic perturbations of canonical RITs that the sequence of both the hairpin structure and flanking sequence can have large, and often unexpected, effects on termination efficiency (Reynolds and Chamberlin, 1992; Abe and Aiba, 1996; Cambray et al., 2013; Chen, Liu, et al., 2013); there is no reason to think that non-canonical RITs would not exhibit a similar pattern of sequence specificity. As a result, there is a need for a robust classification of potential non-canonical RITs which considers both the sequence and structural features of these elements so that they can be systematically investigated.

In the previous chapter I showed that covariance models (CMs) are able to capture sequence as well as structural features of canonical as well as putative non-canonical RITs. In this chapter I describe a method for the discovery of potential structured termination motifs across the bacterial kingdom, present an initial analysis of the elements discovered, and provide evidence for their activity through the analysis of a large collection of publicly-available RNA-seq data.

## 5.2 Methods

*James Hadfield (University of Canterbury) ran the MCL clustering under my supervision. Paul P. Gardner (University of Canterbury) developed and ran the analysis of expression data, and assisted in manual curation of cluster alignments. Stinus Lindgreen (University of Canterbury/University of Copenhagen) processed RNA-seq data and performed mapping. I performed all other work described here.*

### 5.2.1 Genome-wise motif discovery

1853 EMBL format files containing the genomic sequence and annotations for 1639 bacteria were obtained from the EMBL European Nucleotide Archive completed bacterial genomes pages, see supplementary information for organisms and accession numbers.

Each EMBL file was screened independently for putative multi-copy termination motifs. For each EMBL file, I extracted sequences from -20 to +80 around annotated ORF stop site. Each extracted sequence was screened for a lower than expected predicted MFE using RNAfold in order to screen out locally GC-rich but unstructured sequences. The sequence under consideration was shuffled 1000 times preserving dinucleotide frequencies, and a Gumbel distribution was fitted to the resulting empirical null MFE distribution using the R MASS package (Venables and Ripley, 1994). Sequences with a native MFE below the 95th percentile of the null distribution were discarded. The resulting set of sequences was then given as input to CMfinder (Yao, Weinberg, and Ruzzo, 2006), which produces collections of locally-aligned structurally conserved motifs. I built a CM for each motif using Infernal 1.0.2 (Nawrocki, Kolbe, and Eddy, 2009). The resulting CMs were searched against the EMBL file the motif was discovered in, and were then screened on the following criteria for the collection of search hits with an E-value of less than 1: a copy number of between 100 and 3000, and a median distance of <10 to the nearest annotated ORF stop site. This resulted in a collection of 4359 putative termination motif CMs, each derived from a single EMBL file.

### 5.2.2 Clustering covariance models

In order to cluster CMs, I developed an extension of MCL-based clustering (Enright, Van Dongen, and Ouzounis, 2002) to generative models of sequence variation. I call the measure of CM similarity I developed for this purpose the reciprocal similarity score (RSS), defined as:

$$\left[ \frac{\sum_{i=1}^n -\ln(E_{x,y,i}) + \sum_{j=1}^n -\ln(E_{x,y,j})}{2n} \right] + \ln(n)$$

where  $E_{x,y,i}$  is the E-value of the  $i$ th sequence emitted by model  $x$  scored by model  $y$  and for the purposes of this study  $n = 1000$ . Briefly, for each pair of CMs 1000 sequences were emitted from each CM and reciprocally scored with the other CM. The average of the negative log-transformed E-values was calculated, then shifted to be strictly positive by adding  $\ln(1000)$  to generate the RSS appropriate for use with MCL. MCL was run over the resulting RSS matrix, and the 100 largest clusters, ranging in size from 332 to 6 CMs, were taken forward for further analysis.

### **5.2.3 Building consensus covariance models**

To build covariance models which captured the diversity of sequences represented by each cluster, I searched the ten CMs with the highest sum of RSS scores in each cluster against the set of genomes which contributed motifs to the cluster. Sequences on which at least four CMs agreed on with an E-value of  $< 1$  were collected. The redundancy of the collected sequences was iteratively reduced in an alignment-free fashion using cd-hit (Li and Godzik, 2006) with the parameters -G 0 -aL 0.1 -aS 0.3 until there were less than 2000 sequences remaining or there were no remaining sequences with  $> 85\%$  nucleotide identity. Sequences were extended by 20 bases on each side to capture features which may not have been in the CMfinder-derived motifs, e.g. poorly conserved poly-U tracts. The resulting set of sequences was aligned using MAFFT Q-INS-i (Katoh and Toh, 2008) using McCaskill base-pairing probabilities (McCaskill, 1990), and secondary structures were predicted using CentroidAlifold (Hamada et al., 2009), again with McCaskill base-pairing probabilities. CMs were built from the resulting cluster alignments, and sequences which did not match the CM with a bitscore of at least 20 were iteratively discarded. The resulting alignments were then manually curated using RALEE (Griffiths-Jones, 2005), trimming non-conserved flanking sequence and extending the predicted secondary structure where possible. Conserved stop codons were specifically trimmed, so as not to bias subsequent searches.

### **5.2.4 Genome annotation**

The resulting cluster CMs were searched over the initial 1853 EMBL files. Bitscore thresholds were set for hit significance for each cluster CM using shuffled sequence. Specifically, each cluster model was also used to search a dinucleotide shuffled database of these same 1853 EMBL file. For each model, a Gumbel distribution was fitted to the distribution of bitscores over the shuffled database, and this null Gumbel distribution was used to compute P-values for hit significance in the native sequences. P-values were corrected for multiple hypothesis testing using the method of Benjamini and Hochberg (1995), and these were used to set bitscore thresholds at specific FDRs.

### 5.2.5 Analysis of expression data

Data sets were downloaded from the SRA (Leinonen et al., 2011), preferring whenever possible to start our own analyses with the raw fastq input instead of relying on previous mapping results. This was done to make the data sets comparable. After retrieving the data sets, we extracted fastq reads for further analysis. Most data sets were downloaded in SRA format. Fastq files were extracted using the command fastqdump –split-3 from the SRA toolkit version 2.3.2-4. This creates two fastq files in the case of paired end data, and one fastq file in case of single end data. When BED files were used as the primary input, the BAM file was extracted directly using bedToBam from the bedtools package, version 2.17.0 (Quinlan and Hall, 2010). Data sets in SOLiD format was translated to fastq using solid2fastq from bfast version 0.7.0a (Homer, Merriman, and Nelson, 2009). All extracted fastq files were cleaned using AdapterRemoval version 1.4 (Lindgreen, 2012) with the flags –trimns –trimqualities to remove residual adapters from the reads and to remove low quality segments and stretches of Ns in the 5 and/or 3 ends.

Most data sets were mapped using bowtie2 version 2.1 (Langmead and Salzberg, 2012), and the output was saved in BAM format using samtools version 0.1.18 (Li, Handsaker, et al., 2009). In the single end case, the following command was used:

```
bowtie2 -x <INDEX> -U <READS> | samtools view -bS - \  
| samtools sort - <OUTPUT>.sorted
```

In the paired end case, a similar command was used, but the number of input files was larger because 1) there are two files containing the paired reads, and 2) additional single end reads might have been produced by AdapterRemoval because some pairs were collapsed due to overlaps, or one mate pair was discarded due to e.g. low quality. For 454 data, using the above command produced few mappings to the reference genome. We therefore used bowtie2 but with relaxed parameters to accommodate the longer reads by adding the flags –local –very-sensitive-local. For SOLiD data, we used bfast version 0.7.0a for mapping with the following commands:

```
bfast match -f <INDEX> -A 1 -r <READS> > <OUTPUT>.bmf  
bfast localalign -f <INDEX> -A 1 -m <OUTPUT>.bmf > <OUTPUT>.baf  
bfast postprocess -f <INDEX> -i <OUTPUT>.baf -A 1 | samtools view -bS - \  
| samtools sort - <OUTPUT>.sorted
```

For each BAM-file, we generated a PLOT file containing two tab separated columns (reverse strand, forward strand) and a line per position in the genome. Each line gives information on the number of mapped reads on each strand for that particular position in the reference genome. The PLOT files were generated using the following commands from samtools version 0.1.18:

```
samtools view -F 0x10 -b <INPUT> (for reads mapped to the forward strand)  
samtools view -f 0x10 -b <INPUT> (for reads mapped to the reverse strand)
```

Then, the samtools depth command was used to get the actual depths and save them in a WIG format file, which was then transformed to PLOT file by filling out the 0-depth positions based on the length of the reference genome.

(info on plots here)

## 5.3 Results

### 5.3.1 Kingdom-wide motif discovery

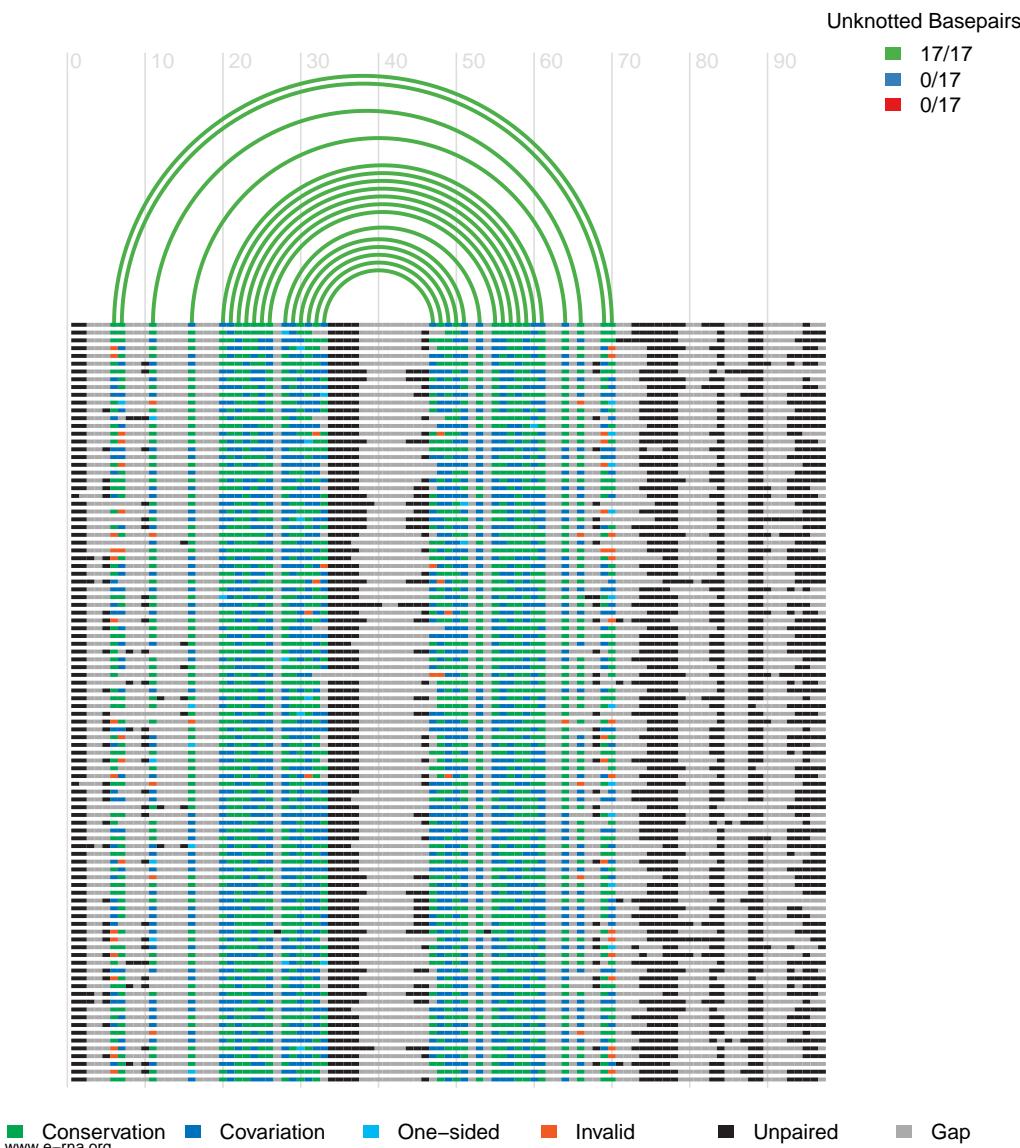
The pipeline I developed for discovering putative termination motifs consisted of 3 major stages (see figure X): genome-wise motif discovery with CMfinder (Yao, Weinberg, and Ruzzo, 2006), clustering of motifs using a novel similarity measure and the MCL algorithm (Enright, Van Dongen, and Ouzounis, 2002), and manual curation of the resulting motif clusters.

In the first stage I extracted sequence from -20 to +80 with respect to annotated stop sites, which were then filtered on predicted structural potential to screen for sequences with stronger structures than predicted by their dinucleotide content alone (see Methods). For each genome, I used the resulting set of sequences as input for the CMfinder algorithm (Yao, Weinberg, and Ruzzo, 2006). Briefly, CMfinder uses heuristic sequence search, thermodynamic and mutual information-based predictions of secondary structure, and CM-based searches within an expectation-maximization (EM) framework to iteratively discover and refine potential structured RNA motifs, returning a multiple sequence alignment and corresponding CM. CMfinder has previously been successfully used as part of pipelines for the discovery of non-coding RNAs in bacteria (Weinberg, Barrick, et al., 2007; Weinberg, Wang, et al., 2010) and eukaryotes (Torarinsson, Yao, et al., 2008), as

well as in our previous discovery of the TRIT element (Gardner, Barquist, et al., 2011). Applying this algorithm to the filtered sequences for each genome resulted in a total of 22310 motif predictions. I searched these CMs back over the genome they were predicted from and removed from consideration motifs with very low ( $<100$ ) or very high ( $> 3000$ ) copy number, or were not enriched with respect to gene terminal regions, leaving a set of 4359 putative termination motifs, approximately 2.5 per organism.

To reduce the complexity of this data set, I developed a method for clustering CMs. Two previous approaches to comparing CMs have been described in the literature. The first, known as CMcompare (Höner zu Siederdissen and Hofacker, 2010), computes the score of a so-called ‘link sequence’, that is a sequence with the highest value of  $\min(S_1(s), S_2(s))$ , where  $S_x(s)$  is the score of sequence  $s$  with respect to model  $S_x$ . While this has been proposed as a measure of CM specificity in the context of the Rfam database, it is unclear how accurately this single link sequence captures the overlap between the sequence spaces described by two CMs, let alone the reality of overlaps in actual biological sequence databases. A second method, proposed as part of the Evofam pipeline for automated ncRNA family discovery in eukaryotic genome alignments (Parker et al., 2011), approximates the Kullback-Leibler divergence between two CMs, that is the similarity of the probability distributions over sequences emitted by the two CMs, using the difference in Infernal CM E-value calculations on a human reference sequence from each model’s training set. In the context of the Evofam pipeline, the use of the human reference sequence is justifiable, as the study was primarily concerned with the discovery of ncRNA families present in the human genome. However, in the present case of clustering motifs across an entire domain of life, there is no obvious single sequence to use as a reference for the purposes of a comparison between every pair of CMs.

I have developed a sampling based approach to measuring CM similarity, inspired by discussions of using summed bitscores as a measure of remote homology between CMs (personal communication, Paul P. Gardner and Sean Eddy) and the reciprocal BLAST measure used by TRIBE-MCL (Enright, Van Dongen, and Ouzounis, 2002). Rather than using a single reference sequence for the purpose of comparison, I use the fact that CMs are generative models to measure the average similarity of of their respective sequence spaces. Infernal reports bitscores and E-value for each match between a CM and a given sequence region. The bitscore, ignoring the specifics of algorithm used (either CYK or



**Figure 5.1: Example alignment of cluster consensus sequences.** Partial alignment of the consensus sequences for cluster 16, visualized using the R-CHIE webserver (Lai et al., 2012). Green arcs represent base-pairing interactions. Nucleotides are visualized as blocks below, and are colored to highlight conservation and covariation in base-pairing relationships within the stem-loop structure.

Inside), is

$$S = \log_2 \left( \frac{P(x | H)}{P(x | R)} \right)$$

where  $P(x | H)$  is the probability of sequence  $x$  under model  $H$ , and  $P(x | R)$  is the probability of  $x$  under a null model  $R$ , generally an iid sites model with a geometric length distribution. This score is expected to follow a Type 1 Extreme Value (or Gumbel) distribution (Karlin and Altschul, 1990; Eddy, 2008), and this empirically appears to be the case for Infernal scores (Nawrocki and Eddy, 2007). Hence the E-value can be calculated as

$$e^{-\lambda(S-\mu)}$$

where  $\lambda$  and  $\mu$  are fitted parameters depending on the size of the database searched and the model architecture, and normalize for these factors. So the reciprocal similarity score (RSS) I have defined:

$$RSS_{x,y} = \left[ \frac{\sum_{i=1}^n -\ln(E_{x,y,i}) + \sum_{j=1}^n -\ln(E_{y,x,j})}{2n} \right] + \ln(n)$$

where  $E_{x,y,i}$  is the E-value of the  $i$ th sequence emitted by model  $x$  scored by model  $y$ , can be understood as the average normalized bitscore of each model over the other's sequence space, and is similar in spirit to Monte Carlo approximations to the Kullback-Leibler divergence (Parker et al., 2011; Juang and Rabiner, 1985).

This measure appeared robust to the number of samples used, but this may depend in part on model complexity. As the maximal E-value in this case is  $n$ ,  $-\ln(n)$  is a theoretical lower bound on the average  $-\ln(E)$ , and the subtraction of this factor ensures that the RSS is strictly positive. It is worth noting that this measure is symmetric ignoring sampling error. Asymmetric variants may have some applications. For instance, by taking the minimum of the average bitscore under either model, one would give preference to full-length model matches in comparisons between models of various sizes due to the glocal nature of Infernal search (global with respect to the model, local with respect to sequence), and this may be preferable for determining similarity between ncRNA families. Conversely, taking the maximum may have some utility in searching for shorter motifs. In the current application, I expect all CMs to be of roughly similar sizes and symmetric measures simplify clustering. This measure should be applicable to any

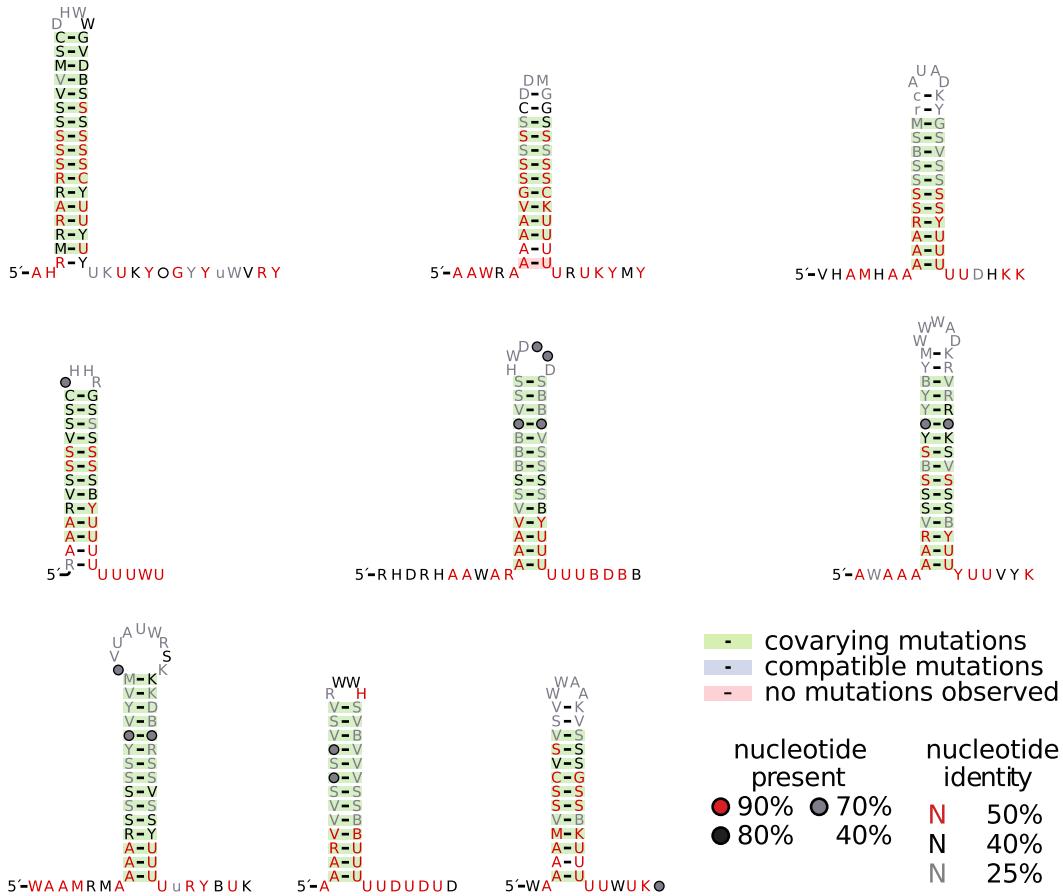
generative model, and so could be similarly used to cluster e.g. HMMs.

A related measure was previously used by the TRIBE-MCL algorithm to cluster protein families based on reciprocal  $\log_{10}$  BLAST E-values (Enright, Van Dongen, and Ouzounis, 2002). The MCL algorithm is described in detail elsewhere (Van Dongen, 2008), but briefly uses simulations of random walks on a weighted graph to define clusters through an unsupervised, iterative process. Unsurprisingly, many of the clusters that were generated using MCL with RSSes appeared to be composed of CMs representing canonical RITs on visual inspection with some notable exceptions, described below. However, despite a complete lack of phylogenetic assumptions in our pipeline, we found that the majority of clusters were dominated by one or two orders, generally within the same phyla, and sometimes even a single genera. This both validates our clustering procedure and indicates that RITs, despite their small size and stereotypical composition, carry a phylogenetic signal when considered in aggregate.

To further study lineage-specific biases in terminator composition, I took the top 100 clusters, ranging in size from 332 to 6 CMs, and constructed consensus models through a semi-automated process. First, for each cluster I selected the 10 CMs with the highest sum of RSS scores with other cluster CMs (or all CMs in the case of clusters with < 10 members), and searched these across all of the genomes the cluster CMs were derived from. Regions that these CMs agreed were likely to be terminator sequences were collected and aligned using MAFFT Q-INS-i (Katoh and Toh, 2008), a heuristic Sankoff alignment algorithm which considers both sequence and secondary structure in alignment, and secondary structure was predicted using CentroidAlifold (Hamada et al., 2009), and manually refined using RALEE (Griffiths-Jones, 2005) (see figure 5.1; see also Methods for detailed alignment protocol). I annotated the 1853 EMBL files we started with, and iteratively removed from consideration any model with at least 85% of its sequence hits covered by another model. This left 16 putative terminator models, on which all further analysis was done.

### 5.3.2 Lineage-specific differences in RIT utilization

Of the 16 resulting cluster, 9 appeared to be canonical RITs on visual inspection (see figure 5.2). All shared known features of canonical RITs, including a 5' poly-A region, a G/C-rich hairpin, and a poly-U tail, but differ in stem length, hairpin loop length,

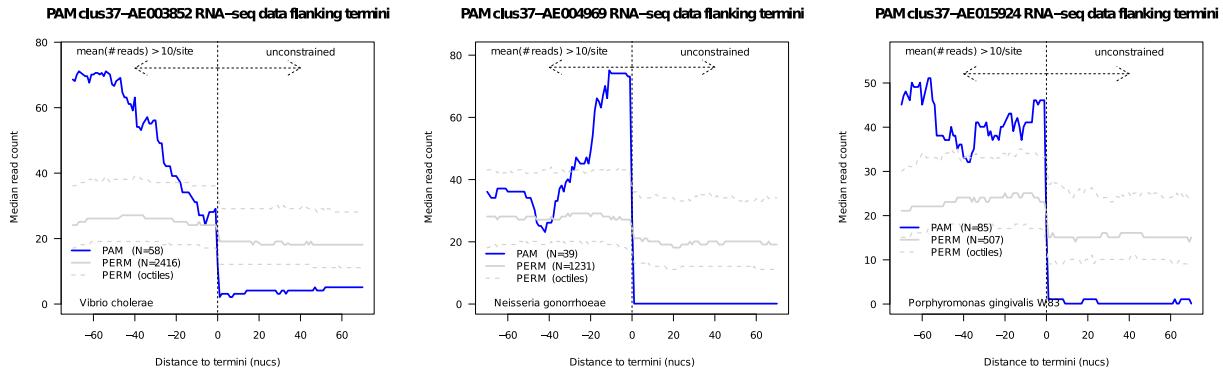


**Figure 5.2: Most informative sequence for nine canonical RIT clusters.** Each cluster consensus model was searched across all genomes and sequence hits with an FDR of 0.01 were aligned to the model. Duplicate sequences were removed and 5000 randomly sampled sequences were used to calculate the most informative sequence (MIS), a projection of any bases with frequencies above .25 onto IUPAC characters (Freyhult, Moulton, and Gardner, 2005). Structures were drawn using R2R (Weinberg and Breaker, 2011). From left to right, images shown represent consensus alignments for clusters 16, 25, 29 (top row); 37, 88, 89 (middle row); 95, and 96 (bottom row).

and base composition. An interesting feature of these models is the universal presence of base-pairing interactions between the poly-A and poly-U regions. Though it has been widely assumed that the poly-A region's function is primarily to contribute to bidirectional activity of RITs, some studies have shown that complementarity between the poly-A and poly-U region increase termination efficiency (Abe and Aiba, 1996; Chen, Liu, et al., 2013), presumably by contributing to the ratcheting effect of hairpin formation on the poly-U tail. In fact, a recent study showed that strong terminators with clear poly-A regions generally do not posses strong bidirectional activity, suggesting that the primary function of the poly-A region is to contribute to this ratcheting (Chen, Liu, et al., 2013). I have observed covariation within many of the A-U pairs in our terminator models, supporting this observation.

To validate RIT predictions, publicly available RNA-seq datasets were collected and plots summarizing the behavior of transcription across the predictions were created (see Methods for details). There are some difficulties in using RNA-seq data to validate terminator activity. In perfect digital transcriptomic data, we would expect to observe the majority of transcripts terminating precisely within the poly-U tail of annotated RITs. Unfortunately, modern high-throughput sequencing technologies do not sequence complete RNA molecules, rather sequencing short stretches of size-selected fragmented RNA libraries. These fragments in these libraries are incidentally selected for sequence composition during both library amplification through PCR and sequencing, often with poorly understood biases, giving rise to the characteristically hilly appearance of these data sets when visualized. Additonally, protocols for the sequencing of RNA retaining strand information generally sequence all fragments of a particular RNA molecule in the same direction (for example, see Croucher, Fookes, et al. (2009)). As a result, if we assume that the fragmentation proceeds roughly by a Poisson process, this will naturally lead to an exponential decay in apparent expression along the 3' region of each transcript. Newer data sets with longer read lengths tend to give cleaner indications of termination activity. Finally, in some data sets we observed patterns of reported transcription that are suggestive of degradation of the RNA by 3' exonucleases or high levels of genomic DNA contamination.

Despite these potential problems, when taken in aggregate, we can observe a clear signal from the termination activity of canonical RITs (see figure 5.3).



**Figure 5.3: Analysis of diverse RNA-seq datasets confirm canonical terminator activity.** These plots present representative analysis for putative attenuation motifs (PAM) by the cluster 37 canonical terminator consensus model. The median expression over PAMs with an upstream mean expression of at least 10 reads per position is plotted in blue. Random positions meeting this same constraint are plotted in grey, and the dashed grey lines provide a 75% confidence interval for this estimate. RNA-seq data (from left to right) drawn from experiments in the  $\gamma$ -proteobacterium *Vibrio cholerae* (Mandlik et al., 2011), the  $\beta$ -proteobacterium *Neisseria gonorrhoeae* (Isabella and Clark, 2011), and the Bacteroidetes *Porphyromonas gingivalis* (Høvik et al., 2012).

### 5.3.3 Non-canonical putative attenuators of transcription

## 5.4 Discussion



# Publications

Publications arising in the course of this thesis:

- Read H., Johnson S., Barquist L., Mills G., Gardner P.P., Patrick W.M., Wiles S. **The effect of constitutive bioluminescence expression on the in vitro and in vivo fitness of the mouse enteropathogen *Citrobacter rodentium*.** Manuscript in preparation.
- Okoro C.K., Barquist L., Kingsley R.A., Connor T.R., Harris S.R., Arends M., Stevens M., Parry C.M., Al-Mashhadani M.N., Kariuki S., Msefula C.L., Gordon M.A., de Pinna E., Wain J., Heyderman R.S., Obaro S., Alonso P.L., Mandomando I., MacLennan C.A., Tapia M.D., Levine M.M., Tennant S.M., Parkhill J., Dougan G. **Signatures of adaptation in human invasive *S. Typhimurium* populations.** Manuscript in preparation.
- Wilf N.M., Reid A.J., Ramsay J.P., Williamson N.R., Croucher N.J., Gatto L., Hester S.S., Goulding D., Barquist L., Lilley K.S., Kingsley R.A., Dougan G., Salmond G.P.C.. **RNA-seq reveals the RNA binding proteins, Hfq and RsmA, play various roles in virulence, antibiotic production and genomic flux in *Serratia* sp. 39006.** Manuscript under review.
- Pettit L.J., Browne H.P., Yu L., Smits W.K., Fagan R.P., Barquist L., Martin M.J., Goulding D., Duncan S.H., Flint H.J., Dougan G., Choudhary J.S., Lawley T.D. **Functional genomics reveals that *Clostridium difficile* Spo0A coordinates sproutulation, virulence and metabolism.** Manuscript under review.

- Reuter S., Connor T.R., Barquist L., Walker D., Feltwell T., Harris S.R., Fookes M., Hall M.E., Fuchs T.M., Corander J., Dufour M., Ringwood T., Savin C., Bouchier C., Martin L., Miettinen M., Shubin M., Laukkanen-Ninios R., Sihvonen L.M., Siitonens A., Skurnik M., Falcão J.P., Fukushima H., Scholz H.C., Prentice M., Wren B.W., Parkhill J., Carniel E., Achtman M., McNally A., Thomson N.R. **Parallel independent evolution of pathogenicity within the genus *Yersinia*.** Manuscript under review.
- Barquist L., Burge S.W., Gardner P.P. **Building non-coding RNA families.** *Methods in Molecular Biology*, in press.
- Hoeppner M.P., Barquist L., Gardner P.P. **An introduction to RNA databases.** *Methods in Molecular Biology*, in press.
- Croucher N.J., Mitchell A.M., Gould K.A., Inverarity D., Barquist L., Feltwell T., Fookes M.C., Harris S.R., Dordel J., Salter S.J., Browall S., Zemlickova H., Parkhill J., Normark S., Henriques-Normark B., Hinds J., Mitchell T.J., Bentley S.D. **Dominant role of nucleotide substitution in the diversification of serotype 3 pneumococci over decades and during a single infection.** *PLoS Genetics*, 2013.
- Kingsley R.A., Whitehead S., Connor T., Barquist L., Sait L., Holt K., Sivaraman K., Wileman T., Goulding D., Clare S., Hale C., Seshasayee A., Harris S., Thomson N., Gardner P., Rabsch W., Wigley P., Humphrey T., Parkhill J., Dougan G. **Genome and transcriptome adaptation accompanying emergence of the DT2 host-restricted *Salmonella* Typhimurium pathovar.** *mBio*, 2013.
- Martin M.J., Clare S., Goulding D., Faulds-Pain A., Barquist L., Browne H., Pettit L., Dougan G., Lawley T.D., Wren B.W. **The *agr* locus regulates virulence and colonization genes in *Clostridium difficile* 027.** *Journal of Bacteriology*, 2013.
- Barquist L., Boinett C.J., Cain A.K.. **Approaches to querying bacterial genomes with transposon-insertion sequencing.** *RNA Biology*, 10(7), 2013.

## *Published Works*

---

- Barquist L., Langridge G.C., Turner D.J., Phan M.D., Turner A.K., Bateman A., Parkhill J., Wain J., Gardner P.P. **A comparison of dense transposon insertion libraries in the *Salmonella* serovars Typhi and Typhimurium.** *Nucleic Acids Research*, 41(8):4549-4564, 2013.
- Burge S.W., Daub J., Eberhardt R., Tate J., Barquist L., Nawrocki E.P., Eddy S.R., Gardner P.P., Bateman A. **Rfam 11.0: 10 years of RNA families.** *Nucleic Acids Research*. 41(D1):D226-D232, 2013
- Croucher N.J., Harris S.R., Barquist L., Parkhill J., Bentley S.D. **A high-resolution view of genome-wide pneumococcal transformation.** *PLoS Pathogens*, 8(6), 2012
- Westesson O., Barquist L., Holmes I. **HandAlign: Bayesian multiple sequence alignment, phylogeny, and ancestral reconstruction.** *Bioinformatics*, 28(8):1170-1171, 2012
- Gardner P.P., Barquist L., Bateman A., Nawrocki E.P., Weinberg Z. **RNIE: genome-wide prediction of bacterial intrinsic terminators.** *Nucleic Acids Research*, 39(14):5845-5852, 2011



# References

- Abe, H. and H. Aiba (1996). “Differential contributions of two elements of rho-independent terminator to transcription termination and mRNA stabilization”. *Biochimie* 78.11-12 (1996), pp. 1035–42 (see pp. 99, 100, 110).
- Abrahams, G. L. and M. Hensel (2006). “Manipulating cellular transport and immune responses: dynamic interactions between intracellular *Salmonella enterica* and its host cells”. *Cell Microbiol* 8.5 (May 2006), pp. 728–37. DOI: 10.1111/j.1462-5822.2006.00706.x (see pp. 22, 25).
- AbuOun, M., P. F. Suthers, G. I. Jones, B. R. Carter, M. P. Saunders, C. D. Maranas, M. J. Woodward, and M. F. Anjum (2009). “Genome scale reconstruction of a *Salmonella* metabolic model: comparison of similarity and differences with a commensal *Escherichia coli* strain”. *J Biol Chem* 284.43 (Oct. 2009), pp. 29480–8. DOI: 10.1074/jbc.M109.005868 (see p. 20).
- Achtman, M., K. Zurth, G. Morelli, G. Torrea, A. Guiyoule, and E. Carniel (1999). “*Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*”. *Proceedings of the National Academy of Sciences of the United States of America* 96.24 (1999), pp. 14043–8 (see p. xix).
- Achtman, M. and M. Wagner (2008). “Microbial diversity and the genetic nature of microbial species”. *Nat Rev Microbiol* 6.6 (June 2008), pp. 431–40. DOI: 10.1038/nrmicro1872 (see p. xix).
- Adey, A., H. G. Morrison, Asan, X. Xun, J. O. Kitzman, E. H. Turner, B. Stackhouse, A. P. MacKenzie, N. C. Caruccio, X. Zhang, and J. Shendure (2010). “Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition”. *Genome biology* 11.12 (2010), R119 (see p. 6).
- Akerley, B. J., E. J. Rubin, V. L. Novick, K. Amaya, N. Judson, and J. J. Mekalanos (2002). “A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*”. *Proc Natl Acad Sci U S A* 99.2 (Jan. 2002), pp. 966–71. DOI: 10.1073/pnas.012602299 (see p. 2).
- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990). “Basic local alignment search tool”. *J Mol Biol* 215.3 (Oct. 1990), pp. 403–10. DOI: 10.1016/S0022-2836(05)80360-2 (see p. 61).

- Anders, S. and W. Huber (2010). “Differential expression analysis for sequence count data”. *Genome biology* 11.10 (2010), R106 (see p. 67).
- Andersson, S. G., A. Zomorodipour, J. O. Andersson, T. Sicheritz-Pontén, U. C. Alsmark, R. M. Podowski, A. K. Näslund, A. S. Eriksson, H. H. Winkler, and C. G. Kurland (1998). “The genome sequence of Rickettsia prowazekii and the origin of mitochondria”. *Nature* 396.6707 (Nov. 1998), pp. 133–40. DOI: 10.1038/24094 (see p. 23).
- Antal, M., V. Bordeau, V. Douchin, and B. Felden (2005). “A small bacterial RNA regulates a putative ABC transporter”. *The Journal of biological chemistry* 280.9 (2005), pp. 7901–8 (see p. 49).
- Arnvig, K. and D. Young (2012). “Non-coding RNA and its potential role in *Mycobacterium tuberculosis* pathogenesis”. *RNA biology* 9.4 (2012), pp. 427–36 (see p. 14).
- Artsimovitch, I. and R. Landick (2002). “The transcriptional regulator RfaH stimulates RNA chain synthesis after recruitment to elongation complexes by the exposed nontemplate DNA strand”. *Cell* 109.2 (Apr. 2002), pp. 193–203 (see p. 79).
- Aseev, L. V., A. A. Levandovskaya, L. S. Tchufistova, N. V. Scaptsova, and I. V. Boni (2008). “A new regulatory circuit in ribosomal protein operons: S2-mediated control of the rpsB-tsf expression in vivo”. *RNA* 14.9 (2008), pp. 1882–94 (see p. 49).
- Baba, T., T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori (2006). “Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection”. *Molecular systems biology* 2 (2006), p. 2006 0008 (see pp. 2, 11, 32, 33, 50).
- Barquist, L., C. J. Boinett, and A. K. Cain (2013). “Approaches to querying bacterial genomes with high-throughput transposon insertion sequencing”. *RNA Biology* (2013) (see pp. xix, 1, 26).
- Barquist, L., G. C. Langridge, D. J. Turner, M. D. Phan, A. K. Turner, A. Bateman, J. Parkhill, J. Wain, and P. P. Gardner (2013). “A comparison of dense transposon insertion libraries in the *Salmonella* serovars Typhi and Typhimurium”. *Nucleic acids research* (2013) (see pp. xx, 4, 6, 11, 17, 19).
- Barrick, J. E. and R. R. Breaker (2007). “The distributions, mechanisms, and structures of metabolite-binding riboswitches”. *Genome biology* 8.11 (2007), R239 (see pp. 84, 85, 94).
- Bäumler, A. J. (1997). “The record of horizontal gene transfer in *Salmonella*”. *Trends Microbiol* 5.8 (Aug. 1997), pp. 318–22. DOI: 10.1016/S0966-842X(97)01082-2 (see p. 22).

## References

---

- Bäumler, A. J., J. G. Kusters, I. Stojiljkovic, and F. Heffron (1994). “Salmonella typhimurium loci involved in survival within macrophages”. *Infect Immun* 62.5 (May 1994), pp. 1623–30 (see p. 75).
- Bäumler, A. J., R. M. Tsolis, T. A. Ficht, and L. G. Adams (1998). “Evolution of host adaptation in *Salmonella enterica*”. *Infect Immun* 66.10 (Oct. 1998), pp. 4579–87 (see pp. 22, 23).
- Baumler, A. J., R. M. Tsolis, and F. Heffron (1996). “Contribution of fimbrial operons to attachment to and invasion of epithelial cell lines by *Salmonella typhimurium*”. *Infection and immunity* 64.5 (1996), pp. 1862–5 (see p. 56).
- Bauwens, L., F. Vercammen, S. Bertrand, J.-M. Collard, and S. De Ceuster (2006). “Isolation of *Salmonella* from environmental samples collected in the reptile department of Antwerp Zoo using different selective methods”. *J Appl Microbiol* 101.2 (Aug. 2006), pp. 284–9. DOI: [10.1111/j.1365-2672.2006.02977.x](https://doi.org/10.1111/j.1365-2672.2006.02977.x) (see p. 20).
- Benard, L., C. Philippe, B. Ehresmann, C. Ehresmann, and C. Portier (1996). “Pseudoknot and translational control in the expression of the S15 ribosomal protein”. *Biochimie* 78.7 (1996), pp. 568–76 (see p. 49).
- Bengtsson, M., A. Ståhlberg, P. Rorsman, and M. Kubista (2005). “Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels”. *Genome Res* 15.10 (Oct. 2005), pp. 1388–92. DOI: [10.1101/gr.3820805](https://doi.org/10.1101/gr.3820805) (see p. 67).
- Benjamini, Y. and Y. Hochberg (1995). “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. *Journal of the Royal Statistical Society. Series B (Methodological)* (1995), pp. 289–300 (see pp. 69, 102).
- Bentley, D. R. et al. (2008). “Accurate whole human genome sequencing using reversible terminator chemistry”. *Nature* 456.7218 (Nov. 2008), pp. 53–9. DOI: [10.1038/nature07517](https://doi.org/10.1038/nature07517) (see p. 9).
- Berardinis, V. de et al. (2008). “A complete collection of single-gene deletion mutants of *Acinetobacter baylyi ADP1*”. *Mol Syst Biol* 4 (2008), p. 174. DOI: [10.1038/msb.2008.10](https://doi.org/10.1038/msb.2008.10) (see p. 2).
- Bhutta, Z. A. and J. Threlfall (2009). “Addressing the global disease burden of typhoid fever”. *JAMA : the journal of the American Medical Association* 302.8 (2009), pp. 898–9 (see p. 24).
- Bochner, B. R. (2009). “Global phenotypic characterization of bacteria”. *FEMS microbiology reviews* 33.1 (2009), pp. 191–205 (see p. 18).
- Bossi, L. and N. Figueroa-Bossi (2007). “A small RNA downregulates LamB maltoporin in *Salmonella*”. *Molecular microbiology* 65.3 (2007), pp. 799–810 (see p. 52).

- Brégeon, D., V. Colot, M. Radman, and F. Taddei (2001). “Translational misreading: a tRNA modification counteracts a +2 ribosomal frameshift”. *Genes Dev* 15.17 (Sept. 2001), pp. 2295–306. DOI: 10.1101/gad.207701 (see p. 79).
- Brenner, F. W., R. G. Villar, F. J. Angulo, R. Tauxe, and B. Swaminathan (2000). “Salmonella nomenclature”. *J Clin Microbiol* 38.7 (July 2000), pp. 2465–7 (see p. 19).
- Briani, F., G. Deho, F. Forti, and D. Ghisotti (2001). “The plasmid status of satellite bacteriophage P4”. *Plasmid* 45.1 (2001), pp. 1–17 (see p. 36).
- Brutinel, E. D. and J. A. Gralnick (2012). “Preferential utilization of D-lactate by *Shewanella oneidensis*”. *Applied and environmental microbiology* 78.23 (2012), pp. 8474–6 (see p. 4).
- Buchmeier, N. A. and F. Heffron (1989). “Intracellular survival of wild-type *Salmonella typhimurium* and macrophage-sensitive mutants in diverse populations of macrophages”. *Infect Immun* 57.1 (Jan. 1989), pp. 1–7 (see p. 57).
- Burge, S. W., J. Daub, R. Eberhardt, J. Tate, L. Barquist, E. P. Nawrocki, S. R. Eddy, P. P. Gardner, and A. Bateman (2013). “Rfam 11.0: 10 years of RNA families”. *Nucleic acids research* 41.D1 (2013), pp. D226–32 (see pp. 27, 89).
- Burrows, M. and D. J. Wheeler (1994). “A block-sorting lossless data compression algorithm”. (1994) (see p. 61).
- Cain, A. K. and R. M. Hall (2012). “Evolution of a multiple antibiotic resistance region in IncH1 plasmids: reshaping resistance regions in situ”. *The Journal of antimicrobial chemotherapy* 67.12 (2012), pp. 2848–53 (see p. 25).
- Cambray, G., J. C. Guimaraes, V. K. Mutalik, C. Lam, Q.-A. Mai, T. Thimmaiah, J. M. Carothers, A. P. Arkin, and D. Endy (2013). “Measurement and modeling of intrinsic transcription terminators”. *Nucleic Acids Res* 41.9 (May 2013), pp. 5139–48. DOI: 10.1093/nar/gkt163 (see pp. 84, 85, 93, 100).
- Carafa, Y. d’Aubenton, E. Brody, and C. Thermes (1990). “Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop structures”. *J Mol Biol* 216.4 (Dec. 1990), pp. 835–58 (see pp. 85, 86).
- Caspi, R., T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, A. Pujar, A. G. Shearer, M. Travers, D. Weerasinghe, P. Zhang, and P. D. Karp (2012). “The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases”. *Nucleic Acids Res* 40.Database issue (Jan. 2012), pp. D742–53. DOI: 10.1093/nar/gkr1014 (see p. 72).
- CDC (2009). “Salmonella surveillance: Annual summary”. (2009) (see p. 24).

## References

---

- Chao, Y., K. Papenfort, R. Reinhardt, C. M. Sharma, and J. Vogel (2012). “An atlas of Hfq-bound transcripts reveals 3' UTRs as a genomic reservoir of regulatory small RNAs”. *The EMBO journal* 31.20 (2012), pp. 4005–19 (see pp. 49, 50).
- Chen, S., E. A. Lesnik, T. A. Hall, R. Sampath, R. H. Griffey, D. J. Ecker, and L. B. Blyn (2002). “A bioinformatics based approach to discover small RNA genes in the Escherichia coli genome”. *Bio Systems* 65.2-3 (2002), pp. 157–77 (see p. 49).
- Chen, Y.-J., P. Liu, A. A. K. Nielsen, J. A. N. Brophy, K. Clancy, T. Peterson, and C. A. Voigt (2013). “Characterization of 582 natural and synthetic terminators and quantification of their design constraints”. *Nat Methods* 10.7 (July 2013), pp. 659–64. DOI: 10.1038/nmeth.2515 (see pp. 84, 85, 93, 100, 110).
- Chinni, S. V., C. A. Raabe, R. Zakaria, G. Randau, C. H. Hoe, A. Zemann, J. Brosius, T. H. Tang, and T. S. Rozhdestvensky (2010). “Experimental identification and characterization of 97 novel ncRNA candidates in *Salmonella enterica* serovar Typhi”. *Nucleic acids research* 38.17 (2010), pp. 5893–908 (see pp. 27, 49, 50).
- Chomsky, N. (1959). “On certain formal properties of grammars”. *Information and control* 2.2 (1959), pp. 137–167 (see p. 89).
- Christen, B., E. Abeliuk, J. M. Collier, V. S. Kalogeraki, B. Passarelli, J. A. Coller, M. J. Fero, H. H. McAdams, and L. Shapiro (2011). “The essential genome of a bacterium”. *Molecular systems biology* 7 (2011), p. 528 (see pp. 4, 6, 11, 14, 26, 34, 54).
- Citron, M. and H. Schuster (1990). “The c4 repressors of bacteriophages P1 and P7 are antisense RNAs”. *Cell* 62.3 (1990), pp. 591–8 (see p. 53).
- Cole, S. T. et al. (2001). “Massive gene decay in the leprosy bacillus”. *Nature* 409.6823 (Feb. 2001), pp. 1007–11. DOI: 10.1038/35059006 (see p. 23).
- Coornaert, A., A. Lu, P. Mandin, M. Springer, S. Gottesman, and M. Guillier (2010). “MicA sRNA links the PhoP regulon to cell envelope stress”. *Molecular microbiology* 76.2 (2010), pp. 467–79 (see p. 52).
- Crick, F. H. (1966). “The genetic code. 3”. *Scientific American* 215.4 (1966), 55–60 passim (see p. 46).
- Cromie, M. J., Y. Shi, T. Latifi, and E. A. Groisman (2006). “An RNA sensor for intracellular Mg(2+)”. *Cell* 125.1 (2006), pp. 71–84 (see p. 25).
- Croucher, N. J., M. C. Fookes, T. T. Perkins, D. J. Turner, S. B. Marguerat, T. Keane, M. A. Quail, M. He, S. Assefa, J. Bähler, R. A. Kingsley, J. Parkhill, S. D. Bentley, G. Dougan, and N. R. Thomson (2009). “A simple method for directional transcriptome sequencing using Illumina technology”. *Nucleic Acids Res* 37.22 (Dec. 2009), e148. DOI: 10.1093/nar/gkp811 (see p. 110).
- Croucher, N. J., S. R. Harris, L. Barquist, J. Parkhill, and S. D. Bentley (2012). “A high-resolution view of genome-wide pneumococcal transformation”. *PLoS Pathog* 8.6 (2012), e1002745. DOI: 10.1371/journal.ppat.1002745 (see p. 73).

- Crump, J. A., S. P. Luby, and E. D. Mintz (2004). “The global burden of typhoid fever”. *Bulletin of the World Health Organization* 82.5 (2004), pp. 346–53 (see p. 24).
- Darwin, K. H. and V. L. Miller (1999). “InvF is required for expression of genes encoding proteins secreted by the SPI1 type III secretion apparatus in *Salmonella typhimurium*”. *Journal of bacteriology* 181.16 (1999), pp. 4949–54 (see p. 25).
- Datsenko, K. A. and B. L. Wanner (2000). “One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products”. *Proceedings of the National Academy of Sciences of the United States of America* 97.12 (2000), pp. 6640–5 (see p. 2).
- Datta, N. (1962). “Transmissible drug resistance in an epidemic strain of *Salmonella typhimurium*”. *The Journal of hygiene* 60 (1962), pp. 301–10 (see p. 25).
- De Las Penas, A., L. Connolly, and C. A. Gross (1997). “SigmaE is an essential sigma factor in *Escherichia coli*”. *Journal of bacteriology* 179.21 (1997), pp. 6862–4 (see p. 50).
- DeJesus, M., Y. J. Zhang, C. M. Sasetti, E. J. Rubin, J. C. Sacchettini, and T. R. Ioerger (2013). “Bayesian Analysis of Gene Essentiality based on Sequencing of Transposon Insertion Libraries”. *Bioinformatics* (2013) (see p. 11).
- Deng, W., S. R. Liou, 3. Plunkett G., G. F. Mayhew, D. J. Rose, V. Burland, V. Kodoyianni, D. C. Schwartz, and F. R. Blattner (2003). “Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18”. *Journal of bacteriology* 185.7 (2003), pp. 2330–7 (see p. 24).
- Deng, Z., T. Kieser, and D. A. Hopwood (1987). “Activity of a *Streptomyces* transcriptional terminator in *Escherichia coli*”. *Nucleic acids research* 15.6 (1987), pp. 2665–2675 (see p. 99).
- Dillies, M.-A., A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaeffer, S. Le Crom, M. Guedj, F. Jaffrézic, and on behalf of The French StatOmique Consortium (2012). “A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis”. *Brief Bioinform* (Sept. 2012). DOI: 10.1093/bib/bbs046 (see p. 66).
- Dillon, S. C., A. D. Cameron, K. Hokamp, S. Lucchini, J. C. Hinton, and C. J. Dorman (2010). “Genome-wide analysis of the H-NS and Sfh regulatory networks in *Salmonella Typhimurium* identifies a plasmid-encoded transcription silencing mechanism”. *Molecular microbiology* 76.5 (2010), pp. 1250–65 (see p. 43).
- Diwa, A., A. L. Bricker, C. Jain, and J. G. Belasco (2000). “An evolutionarily conserved RNA stem-loop functions as a sensor that directs feedback regulation of RNase E gene expression”. *Genes & development* 14.10 (2000), pp. 1249–60 (see p. 49).

## References

---

- Dongen, S. van, C. Abreu-Goodger, and A. J. Enright (2008). “Detecting microRNA binding and siRNA off-target effects from expression data”. *Nat Methods* 5.12 (Dec. 2008), pp. 1023–5. doi: 10.1038/nmeth.1267 (see p. 73).
- Doolittle, R. F., D. F. Feng, S. Tsang, G. Cho, and E. Little (1996). “Determining divergence times of the major kingdoms of living organisms with a protein clock”. *Science* 271.5248 (Jan. 1996), pp. 470–7 (see pp. xix, 20).
- Doolittle, W. F. and O. Zhaxybayeva (2009). “On the origin of prokaryotic species”. *Genome Res* 19.5 (May 2009), pp. 744–56. doi: 10.1101/gr.086645.108 (see p. xix).
- Douchin, V., C. Bohn, and P. Bouloc (2006). “Down-regulation of porins by a small RNA bypasses the essentiality of the regulated intramembrane proteolysis protease RseP in Escherichia coli”. *The Journal of biological chemistry* 281.18 (2006), pp. 12253–9 (see p. 49).
- Doyle, M., M. Fookes, A. Ivens, M. W. Mangan, J. Wain, and C. J. Dorman (2007). “An H-NS-like stealth protein aids horizontal DNA transmission in bacteria”. *Science* 315.5809 (2007), pp. 251–2 (see p. 43).
- Dziva, F., P. M. van Diemen, M. P. Stevens, A. J. Smith, and T. S. Wallis (2004). “Identification of Escherichia coli O157 : H7 genes influencing colonization of the bovine gastrointestinal tract using signature-tagged mutagenesis”. *Microbiology* 150.Pt 11 (2004), pp. 3631–45 (see p. 10).
- Echols, H. (1971). “Lysogeny: viral repression and site-specific recombination”. *Annual review of biochemistry* 40 (1971), pp. 827–54 (see p. 35).
- Eckert, S. E., F. Dziva, R. R. Chaudhuri, G. C. Langridge, D. J. Turner, D. J. Pickard, D. J. Maskell, N. R. Thomson, and M. P. Stevens (2011). “Retrospective application of transposon-directed insertion site sequencing to a library of signature-tagged mini-Tn5Km2 mutants of Escherichia coli O157:H7 screened in cattle”. *Journal of bacteriology* 193.7 (2011), pp. 1771–6 (see pp. 4, 10).
- Eddy, S. R. (1998). “Profile hidden Markov models”. *Bioinformatics* 14.9 (1998), pp. 755–63 (see p. 87).
- Eddy, S. R. and R. Durbin (1994). “RNA sequence analysis using covariance models”. *Nucleic Acids Res* 22.11 (June 1994), pp. 2079–88 (see p. 87).
- Eddy, S. R. (2008). “A probabilistic model of local sequence alignment that simplifies statistical significance estimation”. *PLoS Comput Biol* 4.5 (May 2008), e1000069. doi: 10.1371/journal.pcbi.1000069 (see p. 107).
- Eddy, S. R. (2011). “Accelerated Profile HMM Searches”. *PLoS Comput Biol* 7.10 (Oct. 2011), e1002195. doi: 10.1371/journal.pcbi.1002195 (see p. 89).
- Edwards, M. F. and B. A. Stocker (1988). “Construction of delta aroA his delta pur strains of Salmonella typhi”. *Journal of bacteriology* 170.9 (1988), pp. 3991–5 (see p. 44).

- Enright, A. J., S. Van Dongen, and C. A. Ouzounis (2002). “An efficient algorithm for large-scale detection of protein families”. *Nucleic Acids Res* 30.7 (Apr. 2002), pp. 1575–84 (see pp. 101, 104, 105, 108).
- Eriksson, S., S. Lucchini, A. Thompson, M. Rhen, and J. C. D. Hinton (2003). “Unravelling the biology of macrophage infection by gene expression profiling of intracellular *Salmonella enterica*”. *Mol Microbiol* 47.1 (Jan. 2003), pp. 103–18 (see pp. 56, 74).
- Ermolaeva, M. D., H. G. Khalak, O. White, H. O. Smith, and S. L. Salzberg (2000). “Prediction of transcription terminators in bacterial genomes”. *J Mol Biol* 301.1 (Aug. 2000), pp. 27–33. DOI: 10.1006/jmbi.2000.3836 (see pp. 85, 86).
- Faucher, S. P., S. Porwollik, C. M. Dozois, M. McClelland, and F. Daigle (2006). “Transcriptome of *Salmonella enterica* serovar Typhi within macrophages revealed through the selective capture of transcribed sequences”. *Proc Natl Acad Sci U S A* 103.6 (Feb. 2006), pp. 1906–11. DOI: 10.1073/pnas.0509183103 (see p. 57).
- Feesey, N. A., G. Dougan, R. A. Kingsley, R. S. Heyderman, and M. A. Gordon (2012). “Invasive non-typhoidal salmonella disease: an emerging and neglected tropical disease in Africa”. *Lancet* 379.9835 (June 2012), pp. 2489–99. DOI: 10.1016/S0140-6736(11)61752-2 (see pp. xix, 23).
- Fedurco, M., A. Romieu, S. Williams, I. Lawrence, and G. Turcatti (2006). “BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies”. *Nucleic Acids Res* 34.3 (2006), e22. DOI: 10.1093/nar/gnj023 (see p. 9).
- Fields, P. I., R. V. Swanson, C. G. Haidaris, and F. Heffron (1986). “Mutants of *Salmonella typhimurium* that cannot survive within the macrophage are avirulent”. *Proc Natl Acad Sci U S A* 83.14 (July 1986), pp. 5189–93 (see p. 56).
- Figueira, R. and D. W. Holden (2012). “Functions of the *Salmonella* pathogenicity island 2 (SPI-2) type III secretion system effectors”. *Microbiology* 158.Pt 5 (May 2012), pp. 1147–61. DOI: 10.1099/mic.0.058115-0 (see pp. 56, 58).
- Finlay, B. B. and P. Cossart (1997). “Exploitation of mammalian host cell functions by bacterial pathogens”. *Science* 276.5313 (May 1997), pp. 718–25 (see p. xix).
- Finn, R. D., J. Clements, and S. R. Eddy (2011). “HMMER web server: interactive sequence similarity searching”. *Nucleic acids research* 39.Web Server issue (2011), W29–37 (see pp. 27, 36, 37).
- Fisher, R. A. (1941). “The negative binomial distribution”. *Annals of Eugenics* 11.1 (1941), pp. 182–187 (see p. 67).
- Fookes, M. et al. (2011). “*Salmonella bongori* provides insights into the evolution of the *Salmonellae*”. *PLoS Pathog* 7.8 (Aug. 2011), e1002191. DOI: 10.1371/journal.ppat.1002191 (see pp. 21, 22).

## References

---

- Forest, C. G., E. Ferraro, S. C. Sabbagh, and F. Daigle (2010). “Intracellular survival of *Salmonella enterica* serovar Typhi in human macrophages is independent of *Salmonella* pathogenicity island (SPI)-2”. *Microbiology* 156.Pt 12 (Dec. 2010), pp. 3689–98. DOI: 10.1099/mic.0.041624-0 (see pp. 57, 76).
- Forti, F., I. Dragoni, F. Briani, G. Deho, and D. Ghisotti (2002). “Characterization of the small antisense CI RNA that regulates bacteriophage P4 immunity”. *Journal of molecular biology* 315.4 (2002), pp. 541–9 (see p. 36).
- Frank, D. N. and N. R. Pace (1998). “Ribonuclease P: unity and diversity in a tRNA processing ribozyme”. *Annual review of biochemistry* 67 (1998), pp. 153–80 (see p. 49).
- Fraser, C. M. et al. (2001). “Massive gene decay in the leprosy bacillus”. *Nature* 409.6823 (Feb. 2001), pp. 1007–11. DOI: 10.1038/35059006 (see p. 23).
- Freeman, J. A., M. E. Ohl, and S. I. Miller (2003). “The *Salmonella enterica* serovar typhimurium translocated effectors SseJ and SifB are targeted to the *Salmonella*-containing vacuole”. *Infection and immunity* 71.1 (2003), pp. 418–27 (see pp. 42, 43).
- Freyhult, E., V. Moulton, and P. Gardner (2005). “Predicting RNA structure using mutual information”. *Appl Bioinformatics* 4.1 (2005), pp. 53–9 (see p. 109).
- Gallagher, L. A., J. Shendure, and C. Manoil (2011). “Genome-scale identification of resistance functions in *Pseudomonas aeruginosa* using Tn-seq”. *mBio* 2.1 (2011), e00315–10 (see pp. 4, 5, 7–9, 12, 53).
- Gardner, P. P., L. Barquist, A. Bateman, E. P. Nawrocki, and Z. Weinberg (2011). “RNIE: genome-wide prediction of bacterial intrinsic terminators”. *Nucleic acids research* 39.14 (2011), pp. 5845–5852 (see pp. xx, 83, 84, 94, 96, 97, 105).
- Gardner, P. P., J. Daub, J. G. Tate, E. P. Nawrocki, D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn, S. Griffiths-Jones, S. R. Eddy, and A. Bateman (2009). “Rfam: updates to the RNA families database”. *Nucleic Acids Res* 37.Database issue (Jan. 2009), pp. D136–40. DOI: 10.1093/nar/gkn766 (see p. 90).
- Gawronski, J. D., S. M. Wong, G. Giannoukos, D. V. Ward, and B. J. Akerley (2009). “Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung”. *Proceedings of the National Academy of Sciences of the United States of America* 106.38 (2009), pp. 16422–7 (see pp. 4–7, 13, 53).
- Gene Ontology Consortium (2013). “Gene Ontology annotations and resources”. *Nucleic Acids Res* 41.Database issue (Jan. 2013), pp. D530–5. DOI: 10.1093/nar/gks1050 (see p. 72).
- Gerdes, S. Y. et al. (2003). “Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655”. *Journal of bacteriology* 185.19 (2003), pp. 5673–84 (see p. 33).

- Gibson, D. G. et al. (2010). "Creation of a bacterial cell controlled by a chemically synthesized genome". *Science* 329.5987 (July 2010), pp. 52–6. DOI: 10.1126/science.1190719 (see p. 2).
- Gilbreath, J. J., J. Colvocoresses Dodds, P. D. Rick, M. J. Soloski, D. S. Merrell, and E. S. Metcalf (2012). "Enterobacterial common antigen mutants of *Salmonella enterica* serovar Typhimurium establish a persistent infection and provide protection against subsequent lethal challenge". *Infect Immun* 80.1 (Jan. 2012), pp. 441–50. DOI: 10.1128/IAI.05559-11 (see p. 79).
- Glass, J. I., N. Assad-Garcia, N. Alperovich, S. Yooseph, M. R. Lewis, M. Maruf, 3. Hutchison C. A., H. O. Smith, and J. C. Venter (2006). "Essential genes of a minimal bacterium". *Proceedings of the National Academy of Sciences of the United States of America* 103.2 (2006), pp. 425–30 (see p. 11).
- Goodman, A. L., N. P. McNulty, Y. Zhao, D. Leip, R. D. Mitra, C. A. Lozupone, R. Knight, and J. I. Gordon (2009). "Identifying genetic determinants needed to establish a human gut symbiont in its habitat". *Cell host & microbe* 6.3 (2009), pp. 279–89 (see pp. 4, 5, 7, 9, 11, 13, 17).
- Goodman, A. L., M. Wu, and J. I. Gordon (2011). "Identifying microbial fitness determinants by insertion sequencing using genome-wide transposon mutant libraries". *Nature protocols* 6.12 (2011), pp. 1969–80 (see pp. 5, 7).
- Goryshin, I. Y. and W. S. Reznikoff (1998). "Tn5 in vitro transposition". *The Journal of biological chemistry* 273.13 (1998), pp. 7367–74 (see p. 6).
- Green, B., C. Bouchier, C. Fairhead, N. L. Craig, and B. P. Cormack (2012). "Insertion site preference of Mu, Tn5, and Tn7 transposons". *Mobile DNA* 3.1 (2012), p. 3 (see p. 6).
- Griffin, J. E., J. D. Gawronski, M. A. Dejesus, T. R. Ioerger, B. J. Akerley, and C. M. Sassetti (2011). "High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism". *PLoS pathogens* 7.9 (2011), e1002251 (see pp. 4, 10–12, 53).
- Griffiths-Jones, S. (2005). "RALEE–RNA ALignment editor in Emacs". *Bioinformatics* 21.2 (Jan. 2005), pp. 257–9. DOI: 10.1093/bioinformatics/bth489 (see pp. 90, 97, 102, 108).
- Grimont, P. A. and F.-X. Weill (2007). "Antigenic formulae of the *Salmonella* serovars". *WHO Collaborating Centre for Reference and Research on Salmonella, Institut Pasteur, Paris, France* (2007) (see p. 22).
- Gulig, P. A. and 3. Curtiss R. (1987). "Plasmid-associated virulence of *Salmonella typhimurium*". *Infection and immunity* 55.12 (1987), pp. 2891–901 (see p. 25).
- Gulig, P. A., H. Danbara, D. G. Guiney, A. J. Lax, F. Norel, and M. Rhen (1993). "Molecular analysis of spv virulence genes of the *Salmonella* virulence plasmids". *Molecular microbiology* 7.6 (1993), pp. 825–30 (see p. 25).

## References

---

- Haft, D. H., J. D. Selengut, R. A. Richter, D. Harkins, M. K. Basu, and E. Beck (2013). “TIGRFAMs and Genome Properties in 2013”. *Nucleic Acids Res* 41.Database issue (Jan. 2013), pp. D387–95. DOI: 10.1093/nar/gks1234 (see p. 72).
- Hamada, M., K. Sato, H. Kiryu, T. Mituyama, and K. Asai (2009). “CentroidAlign: fast and accurate aligner for structured RNAs by maximizing expected sum-of-pairs score”. *Bioinformatics* 25.24 (Dec. 2009), pp. 3236–43. DOI: 10.1093/bioinformatics/btp580 (see pp. 102, 108).
- Hashimoto, M., T. Ichimura, H. Mizoguchi, K. Tanaka, K. Fujimitsu, K. Keyamura, T. Ote, T. Yamakawa, Y. Yamazaki, H. Mori, T. Katayama, and J. Kato (2005). “Cell size and nucleoid organization of engineered Escherichia coli cells with a reduced genome”. *Molecular microbiology* 55.1 (2005), pp. 137–49 (see p. 33).
- Hautefort, I., A. Thompson, S. Eriksson-Ygberg, M. L. Parker, S. Lucchini, V. Danino, R. J. M. Bongaerts, N. Ahmad, M. Rhen, and J. C. D. Hinton (2008). “During infection of epithelial cells Salmonella enterica serovar Typhimurium undergoes a time-dependent transcriptional adaptation that results in simultaneous expression of three type 3 secretion systems”. *Cell Microbiol* 10.4 (Apr. 2008), pp. 958–84. DOI: 10.1111/j.1462-5822.2007.01099.x (see pp. 56, 74).
- Hebrard, M., C. Kroger, S. Srikumar, A. Colgan, K. Handler, and J. Hinton (2012). “sRNAs and the virulence of Salmonella enterica serovar Typhimurium”. *RNA biology* 9.4 (2012) (see pp. 26, 75).
- Henkin, T. M. and C. Yanofsky (2002). “Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions”. *Bioessays* 24.8 (Aug. 2002), pp. 700–7. DOI: 10.1002/bies.10125 (see pp. 84, 85).
- Hensel, M., J. E. Shea, A. J. Baumler, C. Gleeson, F. Blattner, and D. W. Holden (1997). “Analysis of the boundaries of Salmonella pathogenicity island 2 and the corresponding chromosomal region of Escherichia coli K-12”. *Journal of bacteriology* 179.4 (1997), pp. 1105–11 (see p. 43).
- Hensel, M., J. E. Shea, C. Gleeson, M. D. Jones, E. Dalton, and D. W. Holden (1995). “Simultaneous identification of bacterial virulence genes by negative selection”. *Science* 269.5222 (1995), pp. 400–3 (see p. 3).
- Hensel, M., J. E. Shea, S. R. Waterman, R. Mundy, T. Nikolaus, G. Banks, A. Vazquez-Torres, C. Gleeson, F. C. Fang, and D. W. Holden (1998). “Genes encoding putative effector proteins of the type III secretion system of Salmonella pathogenicity island 2 are required for bacterial virulence and proliferation in macrophages”. *Molecular microbiology* 30.1 (1998), pp. 163–74 (see p. 43).
- Hensel, M. (2004). “Evolution of pathogenicity islands of Salmonella enterica”. *Int J Med Microbiol* 294.2-3 (Sept. 2004), pp. 95–102. DOI: 10.1016/j.ijmm.2004.06.025 (see p. 20).

- Heras, B., S. R. Shouldice, M. Totsika, M. J. Scanlon, M. A. Schembri, and J. L. Martin (2009). “DSB proteins and bacterial pathogenicity”. *Nat Rev Microbiol* 7.3 (Mar. 2009), pp. 215–25. DOI: 10.1038/nrmicro2087 (see p. 79).
- Hilbert, F., F. Smulders, R. Chopra-Dewasthaly, and P. Paulsen (2012). “*ij* in the wildlife-human interface”. *Food Research International* 45.2 (2012), pp. 603–608 (see p. 22).
- Hirose, K., T. Ezaki, M. Miyake, T. Li, A. Q. Khan, Y. Kawamura, H. Yokoyama, and T. Takami (1997). “Survival of Vi-capsulated and Vi-deleted *Salmonella typhi* strains in cultured macrophage expressing different levels of CD14 antigen”. *FEMS Microbiol Lett* 147.2 (Feb. 1997), pp. 259–65 (see pp. 70, 80).
- Hobbs, E. C., J. L. Astarita, and G. Storz (2010). “Small RNAs and small proteins involved in resistance to cell envelope stress and acid shock in *Escherichia coli*: analysis of a bar-coded mutant collection”. *Journal of bacteriology* 192.1 (2010), pp. 59–67 (see pp. 2, 54).
- Hofacker, I. L., W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster (1994). “Fast folding and comparison of RNA secondary structures”. *Monatshefte für Chemie/Chemical Monthly* 125.2 (1994), pp. 167–188 (see p. 96).
- Holt, K. E., J. Parkhill, C. J. Mazzoni, P. Roumagnac, F. X. Weill, I. Goodhead, R. Rance, S. Baker, D. J. Maskell, J. Wain, C. Dolecek, M. Achtman, and G. Dougan (2008). “High-throughput sequencing provides insights into genome variation and evolution in *Salmonella Typhi*”. *Nature genetics* 40.8 (2008), pp. 987–93 (see p. 24).
- Holt, K. E., Y. Y. Teo, H. Li, S. Nair, G. Dougan, J. Wain, and J. Parkhill (2009). “Detecting SNPs and estimating allele frequencies in clonal bacterial populations by sequencing pooled DNA”. *Bioinformatics* 25.16 (2009), pp. 2074–5 (see pp. 23, 25, 35).
- Holt, K. E., N. R. Thomson, J. Wain, M. D. Phan, S. Nair, R. Hasan, Z. A. Bhutta, M. A. Quail, H. Norbertczak, D. Walker, G. Dougan, and J. Parkhill (2007). “Multidrug-resistant *Salmonella enterica* serovar paratyphi A harbors IncHI1 plasmids similar to those found in serovar typhi”. *Journal of bacteriology* 189.11 (2007), pp. 4257–64 (see p. 25).
- Homer, N., B. Merriman, and S. F. Nelson (2009). “BFAST: an alignment tool for large scale genome resequencing”. *PLoS One* 4.11 (2009), e7767. DOI: 10.1371/journal.pone.0007767 (see p. 103).
- Höner zu Siederdissen, C. and I. L. Hofacker (2010). “Discriminatory power of RNA family models”. *Bioinformatics* 26.18 (Sept. 2010), pp. i453–9. DOI: 10.1093/bioinformatics/btq370 (see p. 105).
- Hoon, M. J. L. de, Y. Makita, K. Nakai, and S. Miyano (2005). “Prediction of transcriptional terminators in *Bacillus subtilis* and related species”. *PLoS*

## References

---

- Comput Biol* 1.3 (Aug. 2005), e25. DOI: 10.1371/journal.pcbi.0010025 (see pp. 85, 89, 93).
- Høvik, H., W.-H. Yu, I. Olsen, and T. Chen (2012). “Comprehensive transcriptome analysis of the periodontopathogenic bacterium *Porphyromonas gingivalis* W83”. *J Bacteriol* 194.1 (Jan. 2012), pp. 100–14. DOI: 10.1128/JB.06385-11 (see p. 111).
- Hutchison, C. A., S. N. Peterson, S. R. Gill, R. T. Cline, O. White, C. M. Fraser, H. O. Smith, and J. C. Venter (1999). “Global transposon mutagenesis and a minimal *Mycoplasma* genome”. *Science* 286.5447 (1999), pp. 2165–9 (see pp. 3, 4, 10).
- Ingham, C. J., I. S. Hunter, and M. C. Smith (1995). “Rho-independent terminators without 3 poly-U tails from the early region of actinophage  $\phi$ C31”. *Nucleic acids research* 23.3 (1995), pp. 370–376 (see p. 99).
- Isabella, V. M. and V. L. Clark (2011). “Deep sequencing-based analysis of the anaerobic stimulon in *Neisseria gonorrhoeae*”. *BMC Genomics* 12 (2011), p. 51. DOI: 10.1186/1471-2164-12-51 (see p. 111).
- Iyer, L. M., M. M. Babu, and L. Aravind (2006). “The HIRAN domain and recruitment of chromatin remodeling and repair activities to damaged DNA”. *Cell cycle* 5.7 (2006), pp. 775–82 (see p. 37).
- Jacobs, M. A., A. Alwood, I. Thaipisuttikul, D. Spencer, E. Haugen, S. Ernst, O. Will, R. Kaul, C. Raymond, R. Levy, L. Chun-Rong, D. Guenthner, D. Bovee, M. V. Olson, and C. Manoil (2003). “Comprehensive transposon mutant library of *Pseudomonas aeruginosa*”. *Proceedings of the National Academy of Sciences of the United States of America* 100.24 (2003), pp. 14339–44 (see pp. 2, 11).
- Johnsborg, O., V. Eldholm, and L. S. Havarstein (2007). “Natural genetic transformation: prevalence, mechanisms and function”. *Research in microbiology* 158.10 (2007), pp. 767–78 (see p. 6).
- Jones, A. M., A. Goodwill, and T. Elliott (2006). “Limited role for the DsrA and RprA regulatory RNAs in rpoS regulation in *Salmonella enterica*”. *Journal of bacteriology* 188.14 (2006), pp. 5077–88 (see p. 52).
- Jørgensen, B. B., M. F. Isaksen, and H. W. Jannasch (1992). “Bacterial Sulfate Reduction Above 100degreesC in Deep-Sea Hydrothermal Vent Sediments”. *Science* 258.5089 (Dec. 1992), pp. 1756–7. DOI: 10.1126/science.258.5089.1756 (see p. xix).
- Juang, B.-H. and L. Rabiner (1985). “A Probabilistic Distance Measure for Hidden Markov Models”. *AT&T Technical Journal* 64.2 (1985), pp. 391–408 (see p. 107).
- Kanehisa, M., S. Goto, Y. Sato, M. Furumichi, and M. Tanabe (2012). “KEGG for integration and interpretation of large-scale molecular data sets”. *Nucleic acids research* 40.Database issue (2012), pp. D109–14 (see pp. 72, 76).

- Kang, Y., T. Durfee, J. D. Glasner, Y. Qiu, D. Frisch, K. M. Winterberg, and F. R. Blattner (2004). “Systematic mutagenesis of the *Escherichia coli* genome”. *Journal of bacteriology* 186.15 (2004), pp. 4921–30 (see p. 33).
- Kaper, J. B., J. P. Nataro, and H. L. Mobley (2004). “Pathogenic *Escherichia coli*”. *Nat Rev Microbiol* 2.2 (Feb. 2004), pp. 123–40. DOI: 10.1038/nrmicro818 (see p. 20).
- Karlin, S. and S. F. Altschul (1990). “Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes”. *Proceedings of the National Academy of Sciences of the United States of America* 87.6 (1990), pp. 2264–8 (see p. 107).
- Katoh, K. and H. Toh (2008). “Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework”. *BMC Bioinformatics* 9 (2008), p. 212. DOI: 10.1186/1471-2105-9-212 (see pp. 102, 108).
- Kawano, M., S. Kanaya, T. Oshima, Y. Masuda, T. Ara, and H. Mori (2002). “Distribution of repetitive sequences on the leading and lagging strands of the *Escherichia coli* genome: comparative study of Long Direct Repeat (LDR) sequences”. *DNA research : an international journal for rapid publication of reports on genes and genomes* 9.1 (2002), pp. 1–10 (see p. 50).
- Khan, S. A., R. Stratford, T. Wu, N. McKelvie, T. Bellaby, Z. Hindle, K. A. Sinha, S. Eltze, P. Mastroeni, D. Pickard, G. Dougan, S. N. Chatfield, and F. R. Brennan (2003). “*Salmonella typhi* and *S typhimurium* derivatives harbouring deletions in aromatic biosynthesis and *Salmonella Pathogenicity Island-2* (SPI-2) genes as vaccines and vectors”. *Vaccine* 21.5-6 (2003), pp. 538–48 (see p. 57).
- Khatiwara, A., T. Jiang, S. S. Sung, T. Dawoud, J. N. Kim, D. Bhattacharya, H. B. Kim, S. C. Ricke, and Y. M. Kwon (2012). “Genome scanning for conditionally essential genes in *Salmonella enterica* Serotype Typhimurium”. *Applied and environmental microbiology* 78.9 (2012), pp. 3098–107 (see pp. 4, 12).
- Khatri, P., M. Sirota, and A. J. Butte (2012). “Ten years of pathway analysis: current approaches and outstanding challenges”. *PLoS Comput Biol* 8.2 (2012), e1002375. DOI: 10.1371/journal.pcbi.1002375 (see p. 73).
- Kidgell, C., U. Reichard, J. Wain, B. Linz, M. Torpdahl, G. Dougan, and M. Achtman (2002). “*Salmonella typhi*, the causative agent of typhoid fever, is approximately 50,000 years old”. *Infect Genet Evol* 2.1 (Oct. 2002), pp. 39–45 (see pp. xx, 24).
- Kingsford, C. L., K. Ayanbule, and S. L. Salzberg (2007). “Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake”. *Genome Biol* 8.2 (2007), R22. DOI: 10.1186/gb-2007-8-2-r22 (see pp. 85–87, 93).

## References

---

- Kingsley, R. A., A. D. Humphries, E. H. Weening, M. R. De Zoete, S. Winter, A. Papaconstantinopoulou, G. Dougan, and A. J. Baumler (2003). “Molecular and phenotypic analysis of the CS54 island of *Salmonella enterica* serotype typhimurium: identification of intestinal colonization and persistence determinants”. *Infection and immunity* 71.2 (2003), pp. 629–40 (see p. 25).
- Kinscherf, T. G. and D. K. Willis (2002). “Global regulation by gidA in *Pseudomonas syringae*”. *J Bacteriol* 184.8 (Apr. 2002), pp. 2281–6 (see p. 79).
- Klein, B. A., E. L. Tenorio, D. W. Lazinski, A. Camilli, M. J. Duncan, and L. T. Hu (2012). “Identification of essential genes of the periodontal pathogen *Porphyromonas gingivalis*”. *BMC genomics* 13 (2012), p. 578 (see p. 4).
- Knuth, K., H. Niesalla, C. J. Hueck, and T. M. Fuchs (2004). “Large-scale identification of essential *Salmonella* genes by trapping lethal insertions”. *Molecular microbiology* 51.6 (2004), pp. 1729–44 (see pp. 2, 32).
- Kobayashi, K. et al. (2003). “Essential *Bacillus subtilis* genes”. *Proc Natl Acad Sci U S A* 100.8 (Apr. 2003), pp. 4678–83. DOI: [10.1073/pnas.0730515100](https://doi.org/10.1073/pnas.0730515100) (see p. 2).
- Koonin, E. V. (2003). “Comparative genomics, minimal gene-sets and the last universal common ancestor”. *Nat Rev Microbiol* 1.2 (Nov. 2003), pp. 127–36. DOI: [10.1038/nrmicro751](https://doi.org/10.1038/nrmicro751) (see pp. 1, 2).
- Kothari, A., A. Pruthi, and T. D. Chugh (2008). “The burden of enteric fever”. *Journal of infection in developing countries* 2.4 (2008), pp. 253–9 (see p. 24).
- Kröger, C. et al. (2012). “The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium”. *Proceedings of the National Academy of Sciences of the United States of America* 109.20 (2012), E1277–86 (see pp. 26, 27, 50).
- Kröger, M. and R. Wahl (1998). “Compilation of DNA sequences of *Escherichia coli* K12: description of the interactive databases ECD and ECDC”. *Nucleic Acids Res* 26.1 (Jan. 1998), pp. 46–9 (see p. 85).
- Kropinski, A. M., A. Sulakvelidze, P. Konczy, and C. Poppe (2007). “*Salmonella* phages and prophages—genomics and practical aspects”. *Methods in molecular biology* 394 (2007), pp. 133–75 (see pp. 35–37).
- Kuhle, V. and M. Hensel (2004). “Cellular microbiology of intracellular *Salmonella enterica*: functions of the type III secretion system encoded by *Salmonella* pathogenicity island 2”. *Cell Mol Life Sci* 61.22 (Nov. 2004), pp. 2812–26. DOI: [10.1007/s00018-004-4248-z](https://doi.org/10.1007/s00018-004-4248-z) (see pp. 22, 25, 42).
- Kvam, V. M., P. Liu, and Y. Si (2012). “A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data”. *Am J Bot* 99.2 (Feb. 2012), pp. 248–56. DOI: [10.3732/ajb.1100340](https://doi.org/10.3732/ajb.1100340) (see p. 68).

- Lai, D., J. R. Proctor, J. Y. A. Zhu, and I. M. Meyer (2012). “R-CHIE: a web server and R package for visualizing RNA secondary structures”. *Nucleic Acids Res* 40.12 (July 2012), e95. DOI: 10.1093/nar/gks241 (see p. 106).
- Lampe, D. J., M. E. Churchill, and H. M. Robertson (1996). “A purified mariner transposase is sufficient to mediate transposition in vitro”. *The EMBO journal* 15.19 (1996), pp. 5470–9 (see p. 6).
- Lampe, D. J., T. E. Grant, and H. M. Robertson (1998). “Factors affecting transposition of the Himar1 mariner transposon in vitro”. *Genetics* 149.1 (1998), pp. 179–87 (see p. 6).
- Landt, S. G., E. Abeliuk, P. T. McGrath, J. A. Lesley, H. H. McAdams, and L. Shapiro (2008). “Small non-coding RNAs in Caulobacter crescentus”. *Molecular microbiology* 68.3 (2008), pp. 600–14 (see p. 14).
- Langmead, B. and S. L. Salzberg (2012). “Fast gapped-read alignment with Bowtie 2”. *Nat Methods* 9.4 (Apr. 2012), pp. 357–9. DOI: 10.1038/nmeth.1923 (see p. 103).
- Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg (2009). “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”. *Genome Biol* 10.3 (2009), R25. DOI: 10.1186/gb-2009-10-3-r25 (see p. 61).
- Langridge, G. C. (2010). “Metabolic capability in host-restricted serovars of *Salmonella enterica*”. PhD thesis. University of Cambridge, 2010 (see pp. 19, 27, 59, 63).
- Langridge, G. C., M. D. Phan, D. J. Turner, T. T. Perkins, L. Parts, J. Haase, I. Charles, D. J. Maskell, S. E. Peters, G. Dougan, J. Wain, J. Parkhill, and A. K. Turner (2009). “Simultaneous assay of every *Salmonella Typhi* gene using one million transposon mutants”. *Genome research* 19.12 (2009), pp. 2308–16 (see pp. 4–7, 11, 12, 26–28, 30, 32, 45, 53).
- Larson, M. H., W. J. Greenleaf, R. Landick, and S. M. Block (2008). “Applied force reveals mechanistic and energetic details of transcription termination”. *Cell* 132.6 (Mar. 2008), pp. 971–82. DOI: 10.1016/j.cell.2008.01.027 (see p. 84).
- Lease, R. A., M. E. Cusick, and M. Belfort (1998). “Riboregulation in *Escherichia coli*: DsrA RNA acts by RNA:RNA interactions at multiple loci”. *Proceedings of the National Academy of Sciences of the United States of America* 95.21 (1998), pp. 12456–61 (see pp. 49, 52).
- Leinonen, R., H. Sugawara, M. Shumway, and International Nucleotide Sequence Database Collaboration (2011). “The sequence read archive”. *Nucleic Acids Res* 39.Database issue (Jan. 2011), pp. D19–21. DOI: 10.1093/nar/gkq1019 (see p. 103).

## References

---

- Lemire, S., N. Figueroa-Bossi, and L. Bossi (2011). “Bacteriophage crosstalk: coordination of prophage induction by trans-acting antirepressors”. *PLoS genetics* 7.6 (2011), e1002149 (see p. 36).
- Lesnik, E. A., R. Sampath, H. B. Levene, T. J. Henderson, J. A. McNeil, and D. J. Ecker (2001). “Prediction of rho-independent transcriptional terminators in Escherichia coli”. *Nucleic Acids Res* 29.17 (Sept. 2001), pp. 3583–94 (see pp. 85, 93).
- Li, H., J. Ruan, and R. Durbin (2008). “Mapping short DNA sequencing reads and calling variants using mapping quality scores”. *Genome research* 18.11 (2008), pp. 1851–8 (see pp. 28, 61).
- Li, H. and R. Durbin (2010). “Fast and accurate long-read alignment with Burrows-Wheeler transform”. *Bioinformatics* 26.5 (Mar. 2010), pp. 589–95. DOI: 10.1093/bioinformatics/btp698 (see p. 61).
- Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup (2009). “The Sequence Alignment/Map format and SAMtools”. *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–9. DOI: 10.1093/bioinformatics/btp352 (see p. 103).
- Li, S.-K., P. K.-S. Ng, H. Qin, J. K.-Y. Lau, J. P.-Y. Lau, S. K.-W. Tsui, T.-F. Chan, and T. C.-K. Lau (2013). “Identification of small RNAs in Mycobacterium smegmatis using heterologous Hfq”. *RNA* 19.1 (Jan. 2013), pp. 74–84. DOI: 10.1261/rna.034116.112 (see pp. 98, 99).
- Li, W. and A. Godzik (2006). “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences”. *Bioinformatics* 22.13 (July 2006), pp. 1658–9. DOI: 10.1093/bioinformatics/btl158 (see p. 102).
- Lindgreen, S. (2012). “AdapterRemoval: easy cleaning of next-generation sequencing reads”. *BMC Res Notes* 5 (2012), p. 337. DOI: 10.1186/1756-0500-5-337 (see p. 103).
- Lipman, D. J. and W. R. Pearson (1985). “Rapid and sensitive protein similarity searches”. *Science* 227.4693 (Mar. 1985), pp. 1435–41 (see p. 61).
- Lowe, T. M. and S. R. Eddy (1997). “tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence”. *Nucleic acids research* 25.5 (1997), pp. 955–64 (see pp. 46, 89).
- Lu, J., J. K. Tomfohr, and T. B. Kepler (2005). “Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach”. *BMC Bioinformatics* 6 (2005), p. 165. DOI: 10.1186/1471-2105-6-165 (see p. 67).
- Lucchini, S., G. Rowley, M. D. Goldberg, D. Hurd, M. Harrison, and J. C. Hinton (2006). “H-NS mediates the silencing of laterally acquired genes in bacteria”. *PLoS pathogens* 2.8 (2006), e81 (see pp. 12, 42, 79).

- Lucchini, S., P. McDermott, A. Thompson, and J. C. D. Hinton (2009). “The H-NS-like protein StpA represses the RpoS (sigma 38) regulon during exponential growth of *Salmonella Typhimurium*”. *Mol Microbiol* 74.5 (Dec. 2009), pp. 1169–86. DOI: 10.1111/j.1365-2958.2009.06929.x (see p. 79).
- Lund, S. P., D. Nettleton, D. J. McCarthy, G. K. Smyth, et al. (2012). “Detecting Differential Expression in RNA-sequence Data Using Quasi-likelihood with Shrunken Dispersion Estimates”. *Statistical applications in genetics and molecular biology* 11.5 (2012) (see p. 68).
- Majowicz, S. E., J. Musto, E. Scallan, F. J. Angulo, M. Kirk, S. J. O’Brien, T. F. Jones, A. Fazil, and R. M. Hoekstra (2010). “The global burden of nontyphoidal *Salmonella* gastroenteritis”. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 50.6 (2010), pp. 882–9 (see p. 24).
- Mandlik, A., J. Livny, W. P. Robins, J. M. Ritchie, J. J. Mekalanos, and M. K. Waldor (2011). “RNA-Seq-based monitoring of infection-linked changes in *Vibrio cholerae* gene expression”. *Cell Host Microbe* 10.2 (Aug. 2011), pp. 165–74. DOI: 10.1016/j.chom.2011.07.007 (see p. 111).
- Mann, B., T. van Opijnen, J. Wang, C. Obert, Y. D. Wang, R. Carter, D. J. McGoldrick, G. Ridout, A. Camilli, E. I. Tuomanen, and J. W. Rosch (2012). “Control of virulence by small RNAs in *Streptococcus pneumoniae*”. *PLoS pathogens* 8.7 (2012), e1002788 (see pp. 4, 15, 16).
- Manna, D., S. Porwollik, M. McClelland, R. Tan, and N. P. Higgins (2007). “Microarray analysis of Mu transposition in *Salmonella enterica*, serovar Typhimurium: transposon exclusion by high-density DNA binding proteins”. *Molecular microbiology* 66.2 (2007), pp. 315–28 (see pp. 12, 42).
- Marck, C. and H. Grosjean (2002). “tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features”. *RNA* 8.10 (2002), pp. 1189–232 (see p. 47).
- Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad (2008). “RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays”. *Genome Res* 18.9 (Sept. 2008), pp. 1509–17. DOI: 10.1101/gr.079558.108 (see p. 67).
- Mazurkiewicz, P., C. M. Tang, C. Boone, and D. W. Holden (2006). “Signature-tagged mutagenesis: barcoding mutants for genome-wide screens”. *Nature reviews. Genetics* 7.12 (2006), pp. 929–39 (see p. 3).
- McCarthy, D. J., Y. Chen, and G. K. Smyth (2012). “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation”. *Nucleic acids research* 40.10 (2012), pp. 4288–97 (see p. 68).
- McCaskill, J. S. (1990). “The equilibrium partition function and base pair binding probabilities for RNA secondary structure”. *Biopolymers* 29.6-7 (1990), pp. 1105–19. DOI: 10.1002/bip.360290621 (see p. 102).

## References

---

- McClelland, M., K. E. Sanderson, S. W. Clifton, et al. (2004). “Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid”. *Nature genetics* 36.12 (2004), pp. 1268–74 (see pp. 23, 25, 35, 44).
- McClelland, M., K. E. Sanderson, J. Spieth, et al. (2001). “Complete genome sequence of *Salmonella enterica* serovar *Typhimurium* LT2”. *Nature* 413.6858 (2001), pp. 852–6 (see p. 24).
- McGourty, K., T. L. Thurston, S. A. Matthews, L. Pinaud, L. J. Mota, and D. W. Holden (2012). “*Salmonella* inhibits retrograde trafficking of mannose-6-phosphate receptors and lysosome function”. *Science* 338.6109 (Nov. 2012), pp. 963–7. DOI: [10.1126/science.1227037](https://doi.org/10.1126/science.1227037) (see p. 56).
- Mermin, J., L. Hutwagner, D. Vugia, S. Shallow, P. Daily, J. Bender, J. Koehler, R. Marcus, F. J. Angulo, and Emerging Infections Program FoodNet Working Group (2004). “Reptiles, amphibians, and human *Salmonella* infection: a population-based, case-control study”. *Clin Infect Dis* 38 Suppl 3 (Apr. 2004), S253–61. DOI: [10.1086/381594](https://doi.org/10.1086/381594) (see pp. 20, 22).
- Meyer, M. M., T. D. Ames, D. P. Smith, Z. Weinberg, M. S. Schwalbach, S. J. Giovannoni, and R. R. Breaker (2009). “Identification of candidate structured RNAs in the marine organism ‘*Candidatus Pelagibacter ubique*’”. *BMC genomics* 10 (2009), p. 268 (see p. 49).
- Miao, E. A. and S. I. Miller (2000). “A conserved amino acid sequence directing intracellular type III secretion by *Salmonella typhimurium*”. *Proceedings of the National Academy of Sciences of the United States of America* 97.13 (2000), pp. 7539–44 (see p. 42).
- Miki, T., N. Okada, and H. Danbara (2004). “Two periplasmic disulfide oxidoreduc-tases, DsbA and SrgA, target outer membrane protein SpiA, a component of the *Salmonella* pathogenicity island 2 type III secretion system”. *J Biol Chem* 279.33 (Aug. 2004), pp. 34631–42. DOI: [10.1074/jbc.M402760200](https://doi.org/10.1074/jbc.M402760200) (see p. 80).
- Miotto, P., F. Forti, A. Ambrosi, D. Pellin, D. F. Veiga, G. Balazsi, M. L. Gennaro, C. Di Serio, D. Ghisotti, and D. M. Cirillo (2012). “Genome-wide discovery of small RNAs in *Mycobacterium tuberculosis*”. *PLoS One* 7.12 (2012), e51950. DOI: [10.1371/journal.pone.0051950](https://doi.org/10.1371/journal.pone.0051950) (see pp. 98, 99).
- Mitra, A., K. Angamuthu, and V. Nagaraja (2008). “Genome-wide analysis of the intrinsic terminators of transcription across the genus *Mycobacterium*”. *Tuberculosis (Edinb)* 88.6 (Nov. 2008), pp. 566–75. DOI: [10.1016/j.tube.2008.06.004](https://doi.org/10.1016/j.tube.2008.06.004) (see p. 95).
- Mitra, A., K. Angamuthu, H. V. Jayashree, and V. Nagaraja (2009). “Occurrence, divergence and evolution of intrinsic terminators across eubacteria”. *Genomics* 94.2 (Aug. 2009), pp. 110–6. DOI: [10.1016/j.ygeno.2009.04.004](https://doi.org/10.1016/j.ygeno.2009.04.004) (see p. 96).

- Monack, D. M., B. Raupach, A. E. Hromockyj, and S. Falkow (1996). “Salmonella typhimurium invasion induces apoptosis in infected macrophages”. *Proc Natl Acad Sci U S A* 93.18 (Sept. 1996), pp. 9833–8 (see pp. 56, 63).
- Moran, N. A. (2004). “Cellular microbiology of intracellular *Salmonella enterica*: functions of the type III secretion system encoded by *Salmonella* pathogenicity island 2”. *Cell Mol Life Sci* 61.22 (Nov. 2004), pp. 2812–26. DOI: 10.1007/s00018-004-4248-z (see p. 23).
- Morozova, O. and M. A. Marra (2008). “Applications of next-generation sequencing technologies in functional genomics”. *Genomics* 92.5 (Nov. 2008), pp. 255–64. DOI: 10.1016/j.ygeno.2008.07.001 (see p. 8).
- Mueller, C. A., P. Broz, and G. R. Cornelis (2008). “The type III secretion system tip complex and translocon”. *Mol Microbiol* 68.5 (June 2008), pp. 1085–95. DOI: 10.1111/j.1365-2958.2008.06237.x (see p. 56).
- Murray, A. E., F. Kenig, C. H. Fritsen, C. P. McKay, K. M. Cawley, R. Edwards, E. Kuhn, D. M. McKnight, N. E. Ostrom, V. Peng, A. Ponce, J. C. Priscu, V. Samarkin, A. T. Townsend, P. Wagh, S. A. Young, P. T. Yung, and P. T. Doran (2012). “Microbial life at -13 °C in the brine of an ice-sealed Antarctic lake”. *Proc Natl Acad Sci U S A* 109.50 (Dec. 2012), pp. 20626–31. DOI: 10.1073/pnas.1208607109 (see p. xix).
- Mushegian, A. R. and E. V. Koonin (1996). “A minimal gene set for cellular life derived by comparison of complete bacterial genomes”. *Proc Natl Acad Sci U S A* 93.19 (Sept. 1996), pp. 10268–73 (see p. 1).
- Näsvall, S. J., P. Chen, and G. R. Björk (2004). “The modified wobble nucleoside uridine-5-oxyacetic acid in tRNAPro(cmo5UGG) promotes reading of all four proline codons in vivo”. *RNA* 10.10 (2004), pp. 1662–73 (see p. 46).
- Navarre, W. W., S. Porwollik, Y. Wang, M. McClelland, H. Rosen, S. J. Libby, and F. C. Fang (2006). “Selective silencing of foreign DNA with low GC content by the H-NS protein in *Salmonella*”. *Science* 313.5784 (2006), pp. 236–8 (see pp. 12, 39, 42, 43, 52, 79).
- Naville, M. and D. Gautheret (2010). “Premature terminator analysis sheds light on a hidden world of bacterial transcriptional attenuation”. *Genome biology* 11.9 (2010), R97 (see pp. 50, 84, 85, 87).
- Naville, M., A. Ghuillot-Gaudeffroy, A. Marchais, and D. Gautheret (2011). “ARNold: a web tool for the prediction of Rho-independent transcription terminators”. *RNA Biol* 8.1 (2011), pp. 11–3 (see p. 85).
- Nawrocki, E. P., D. L. Kolbe, and S. R. Eddy (2009). “Infernal 1.0: inference of RNA alignments”. *Bioinformatics* 25.10 (2009), pp. 1335–7 (see pp. 89, 90, 97, 101).

## References

---

- Nawrocki, E. P. and S. R. Eddy (2007). “Query-dependent banding (QDB) for faster RNA similarity searches”. *PLoS Comput Biol* 3.3 (Mar. 2007), e56. DOI: 10.1371/journal.pcbi.0030056 (see pp. 88, 89, 107).
- Neal, R. J. and K. F. Chater (1991). “Bidirectional promoter and terminator regions bracket mmr, a resistance gene embedded in the Streptomyces coelicolor A3(2) gene cluster encoding methylenomycin production”. *Gene* 100 (Apr. 1991), pp. 75–83 (see p. 99).
- Needleman, S. B. and C. D. Wunsch (1970). “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. *J Mol Biol* 48.3 (Mar. 1970), pp. 443–53 (see p. 61).
- Nichols, R. J., S. Sen, Y. J. Choo, P. Beltrao, M. Zietek, R. Chaba, S. Lee, K. M. Kazmierczak, K. J. Lee, A. Wong, M. Shales, S. Lovett, M. E. Winkler, N. J. Krogan, A. Typas, and C. A. Gross (2011). “Phenotypic landscape of a bacterial cell”. *Cell* 144.1 (2011), pp. 143–56 (see p. 18).
- Ochman, H. and E. A. Groisman (1996). “Distribution of pathogenicity islands in *Salmonella* spp”. *Infection and immunity* 64.12 (1996), pp. 5410–2 (see p. 25).
- Ochman, H. and A. C. Wilson (1987). “Evolution in bacteria: evidence for a universal substitution rate in cellular genomes”. *J Mol Evol* 26.1-2 (1987), pp. 74–86 (see p. 20).
- Ohlson, M. B., K. Fluhr, C. L. Birmingham, J. H. Brumell, and S. I. Miller (2005). “SseJ deacylase activity by *Salmonella enterica* serovar Typhimurium promotes virulence in mice”. *Infection and immunity* 73.10 (2005), pp. 6249–59 (see p. 43).
- Okamura, M., H. S. Lillehoj, R. B. Raybourne, U. S. Babu, R. A. Heckert, H. Tani, K. Sasai, E. Baba, and E. P. Lillehoj (2005). “Differential responses of macrophages to *Salmonella enterica* serovars Enteritidis and Typhimurium”. *Vet Immunol Immunopathol* 107.3-4 (Sept. 2005), pp. 327–35. DOI: 10.1016/j.vetimm.2005.05.009 (see p. 57).
- Okoro, C. K. et al. (2012). “Intracontinental spread of human invasive *Salmonella* Typhimurium pathovariants in sub-Saharan Africa”. *Nat Genet* 44.11 (Nov. 2012), pp. 1215–21. DOI: 10.1038/ng.2423 (see pp. xix, 23).
- Opijnen, T. van, K. L. Bodi, and A. Camilli (2009). “Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms”. *Nature methods* 6.10 (2009), pp. 767–72 (see pp. 4–7, 9, 11, 13).
- Opijnen, T. van and A. Camilli (2012). “A fine scale phenotype-genotype virulence map of a bacterial pathogen”. *Genome research* 22.12 (2012), pp. 2541–51 (see pp. 4, 9, 12, 13, 18).
- Osawa, S., T. H. Jukes, K. Watanabe, and A. Muto (1992). “Recent evidence for evolution of the genetic code”. *Microbiological reviews* 56.1 (1992), pp. 229–64 (see p. 47).

- Padalon-Brauch, G., R. Hershberg, M. Elgrably-Weiss, K. Baruch, I. Rosenshine, H. Margalit, and S. Altuvia (2008). “Small RNAs encoded within genetic islands of *Salmonella typhimurium* show host-induced expression and role in virulence”. *Nucleic acids research* 36.6 (2008), pp. 1913–27 (see pp. 36, 50).
- Pallen, M. J. and B. W. Wren (2007). “Bacterial pathogenomics”. *Nature* 449.7164 (Oct. 2007), pp. 835–42. DOI: 10.1038/nature06248 (see pp. 22, 23).
- Papenfort, K., V. Pfeiffer, F. Mika, S. Lucchini, J. C. Hinton, and J. Vogel (2006). “SigmaE-dependent small RNAs of *Salmonella* respond to membrane stress by accelerating global *omp* mRNA decay”. *Molecular microbiology* 62.6 (2006), pp. 1674–88 (see p. 51).
- Paradis, S., M. Boissinot, N. Paquette, S. D. Belanger, E. A. Martel, D. K. Boudreau, F. J. Picard, M. Ouellette, P. H. Roy, and M. G. Bergeron (2005). “Phylogeny of the Enterobacteriaceae based on genes encoding elongation factor Tu and F-ATPase beta-subunit”. *International journal of systematic and evolutionary microbiology* 55.Pt 5 (2005), pp. 2013–25 (see p. 20).
- Parker, B. J., I. Moltke, A. Roth, S. Washietl, J. Wen, M. Kellis, R. Breaker, and J. S. Pedersen (2011). “New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes”. *Genome Res* 21.11 (Nov. 2011), pp. 1929–43. DOI: 10.1101/gr.112516.110 (see pp. 105, 107).
- Parkhill, J. et al. (2001). “Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18”. *Nature* 413.6858 (2001), pp. 848–52 (see pp. 23–25).
- Peters, J. M., A. D. Vangeloff, and R. Landick (2011). “Bacterial transcription terminators: the RNA 3'-end chronicles”. *J Mol Biol* 412.5 (Oct. 2011), pp. 793–813. DOI: 10.1016/j.jmb.2011.03.036 (see pp. 83, 84).
- Pham, H. N., K. Ohkusu, N. Mishima, M. Noda, M. Monir Shah, X. Sun, M. Hayashi, and T. Ezaki (2007). “Phylogeny and species identification of the family Enterobacteriaceae based on dnaJ sequences”. *Diagn Microbiol Infect Dis* 58.2 (June 2007), pp. 153–61. DOI: 10.1016/j.diagmicrobio.2006.12.019 (see p. 20).
- Phan, M. D., C. Kidgell, S. Nair, K. E. Holt, A. K. Turner, J. Hinds, P. Butcher, F. J. Cooke, N. R. Thomson, R. Titball, Z. A. Bhutta, R. Hasan, G. Dougan, and J. Wain (2009). “Variation in *Salmonella enterica* serovar typhi IncHI1 plasmids during the global spread of resistant typhoid fever”. *Antimicrobial agents and chemotherapy* 53.2 (2009), pp. 716–27 (see p. 25).
- Pickard, D., R. A. Kingsley, C. Hale, K. Turner, K. Sivaraman, M. Wetter, G. Langridge, and G. Dougan (2013). “A genome-wide mutagenesis screen identifies multiple genes contributing to Vi capsular expression in *Salmonella Typhi*”. *Journal of bacteriology* (2013) (see pp. 4, 12).

## References

---

- Pickard, D., J. Wain, S. Baker, A. Line, S. Chohan, M. Fookes, A. Barron, P. O. Gaora, J. A. Chabalgoity, N. Thanky, C. Scholes, N. Thomson, M. Quail, J. Parkhill, and G. Dougan (2003). “Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding *Salmonella enterica* pathogenicity island SPI-7”. *Journal of bacteriology* 185.17 (2003), pp. 5055–65 (see p. 25).
- Prosseda, G., M. L. Di Martino, R. Campilongo, R. Fioravanti, G. Micheli, M. Casalino, and B. Colonna (2012). “Shedding of genes that interfere with the pathogenic lifestyle: the *Shigella* model”. *Res Microbiol* 163.6-7 (July 2012), pp. 399–406. DOI: 10.1016/j.resmic.2012.07.004 (see pp. 20, 22, 23).
- Punta, M., P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn (2012). “The Pfam protein families database”. *Nucleic acids research* 40.Database issue (2012), pp. D290–301 (see pp. 27, 36, 37).
- Quinlan, A. R. and I. M. Hall (2010). “BEDTools: a flexible suite of utilities for comparing genomic features”. *Bioinformatics* 26.6 (Mar. 2010), pp. 841–2. DOI: 10.1093/bioinformatics/btq033 (see p. 103).
- Quirk, P. G., E. Dunkley, P. Lee, and T. A. Krulwich (1993). “Identification of a putative *Bacillus subtilis* rho gene.” *Journal of bacteriology* 175.3 (1993), pp. 647–654 (see p. 85).
- Rabsch, W., W. Voigt, R. Reissbrodt, R. M. Tsolis, and A. J. Baumler (1999). “*Salmonella typhimurium* IroN and FepA proteins mediate uptake of enterobactin but differ in their specificity for other siderophores”. *Journal of bacteriology* 181.11 (1999), pp. 3610–2 (see p. 44).
- Raghavan, R., E. A. Groisman, and H. Ochman (2011). “Genome-wide detection of novel regulatory RNAs in *E. coli*”. *Genome research* 21.9 (2011), pp. 1487–97 (see pp. 27, 48–50).
- Rathman, M., L. P. Barker, and S. Falkow (1997). “The unique trafficking pattern of *Salmonella typhimurium*-containing phagosomes in murine macrophages is independent of the mechanism of bacterial entry”. *Infect Immun* 65.4 (Apr. 1997), pp. 1475–85 (see p. 56).
- Ravin, N. V., A. N. Svarchevsky, and G. Deho (1999). “The anti-immunity system of phage-plasmid N15: identification of the antirepressor gene and its control by a small processed RNA”. *Molecular microbiology* 34.5 (1999), pp. 980–94 (see p. 53).
- Retchless, A. C. and J. G. Lawrence (2007). “Temporal fragmentation of speciation in bacteria”. *Science* 317.5841 (Aug. 2007), pp. 1093–6. DOI: 10.1126/science.1144876 (see p. 20).

- Reynolds, R. and M. J. Chamberlin (1992). “Parameters affecting transcription termination by *Escherichia coli* RNA. II. Construction and analysis of hybrid terminators”. *J Mol Biol* 224.1 (Mar. 1992), pp. 53–63 (see p. 100).
- Richter, A. S. and R. Backofen (2012). “Accessibility and conservation: General features of bacterial small RNA-mRNA interactions?”: *RNA biology* 9.7 (2012) (see p. 26).
- Rivas, E. and S. R. Eddy (2001). “Noncoding RNA gene detection using comparative sequence analysis”. *BMC bioinformatics* 2 (2001), p. 8 (see p. 50).
- Robinson, M. D., D. J. McCarthy, and G. K. Smyth (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. *Bioinformatics* 26.1 (2010), pp. 139–40 (see pp. 65, 67, 68).
- Robinson, M. D. and A. Oshlack (2010). “A scaling normalization method for differential expression analysis of RNA-seq data”. *Genome Biol* 11.3 (2010), R25. DOI: 10.1186/gb-2010-11-3-r25 (see pp. 65, 66, 68).
- Robinson, M. D. and G. K. Smyth (2007). “Moderated statistical tests for assessing differences in tag abundance”. *Bioinformatics* 23.21 (Nov. 2007), pp. 2881–7. DOI: 10.1093/bioinformatics/btm453 (see pp. 67, 68).
- Robinson, M. D. and G. K. Smyth (2008). “Small-sample estimation of negative binomial dispersion, with applications to SAGE data”. *Biostatistics* 9.2 (Apr. 2008), pp. 321–32. DOI: 10.1093/biostatistics/kxm030 (see p. 68).
- Robinson, M. D., D. Strbenac, C. Stirzaker, A. L. Statham, J. Song, T. P. Speed, and S. J. Clark (2012). “Copy-number-aware differential analysis of quantitative DNA sequencing data”. *Genome Res* 22.12 (Dec. 2012), pp. 2489–96. DOI: 10.1101/gr.139055.112 (see p. 68).
- Rosenblad, M. A., N. Larsen, T. Samuelsson, and C. Zwieb (2009). “Kinship in the SRP RNA family”. *RNA biology* 6.5 (2009), pp. 508–16 (see p. 49).
- Ross-Macdonald, P., P. S. Coelho, T. Roemer, S. Agarwal, A. Kumar, R. Jansen, K. H. Cheung, A. Sheehan, D. Symoniatis, L. Umansky, M. Heidtman, F. K. Nelson, H. Iwasaki, K. Hager, M. Gerstein, P. Miller, G. S. Roeder, and M. Snyder (1999). “Large-scale analysis of the yeast genome by transposon tagging and gene disruption”. *Nature* 402.6760 (1999), pp. 413–8 (see p. 17).
- Rubin, E. J., B. J. Akerley, V. N. Novik, D. J. Lampe, R. N. Husson, and J. J. Mekalanos (1999). “In vivo transposition of mariner-based elements in enteric bacteria and mycobacteria”. *Proceedings of the National Academy of Sciences of the United States of America* 96.4 (1999), pp. 1645–50 (see p. 6).
- Sabbattini, P., F. Forti, D. Ghisotti, and G. Deho (1995). “Control of transcription termination by an RNA factor in bacteriophage P4 immunity: identification of the target sites”. *Journal of bacteriology* 177.6 (1995), pp. 1425–34 (see pp. 37, 53).

## References

---

- Sagan, L. (1967). “On the origin of mitosing cells”. *J Theor Biol* 14.3 (Mar. 1967), pp. 255–74 (see pp. xix, 23).
- Sakakibara, Y., M. Brown, R. C. Underwood, I. S. Mian, and D. Haussler (1994). “Stochastic context-free grammars for modeling RNA”. *System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on*. Vol. 5. IEEE. 1994, pp. 284–293 (see p. 87).
- Salgado, H. et al. (2013). “RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more”. *Nucleic Acids Res* 41.Database issue (Jan. 2013), pp. D203–13. DOI: [10.1093/nar/gks1201](https://doi.org/10.1093/nar/gks1201) (see p. 85).
- Sanger, F., S. Nicklen, and A. R. Coulson (1977). “DNA sequencing with chain-terminating inhibitors”. *Proc Natl Acad Sci U S A* 74.12 (Dec. 1977), pp. 5463–7 (see p. 8).
- Santangelo, T. J. and J. W. Roberts (2002). “RfaH, a bacterial transcription antiterminator”. *Mol Cell* 9.4 (Apr. 2002), pp. 698–700 (see p. 79).
- Santiviago, C. A., M. M. Reynolds, S. Porwollik, S. H. Choi, F. Long, H. L. Andrews-Polymenis, and M. McClelland (2009). “Analysis of pools of targeted *Salmonella* deletion mutants identifies novel genes affecting fitness during competitive infection in mice”. *PLoS pathogens* 5.7 (2009), e1000477 (see pp. 32, 54).
- Santiviago, C. A., C. S. Toro, A. A. Hidalgo, P. Youderian, and G. C. Mora (2003). “Global regulation of the *Salmonella enterica* serovar typhimurium major porin, OmpD”. *Journal of bacteriology* 185.19 (2003), pp. 5901–5 (see p. 52).
- Santos, R. L., S. Zhang, R. M. Tsolis, R. A. Kingsley, L. G. Adams, and A. J. Baumler (2001). “Animal models of *Salmonella* infections: enteritis versus typhoid fever”. *Microbes and infection / Institut Pasteur* 3.14-15 (2001), pp. 1335–44 (see pp. 24, 56).
- Santos, R. L., M. Raffatellu, C. L. Bevins, L. G. Adams, C. Tükel, R. M. Tsolis, and A. J. Bäumler (2009). “Life in the inflamed intestine, *Salmonella* style”. *Trends Microbiol* 17.11 (Nov. 2009), pp. 498–506. DOI: [10.1016/j.tim.2009.08.008](https://doi.org/10.1016/j.tim.2009.08.008) (see p. 22).
- Sassetti, C. M., D. H. Boyd, and E. J. Rubin (2003). “Genes required for mycobacterial growth defined by high density mutagenesis”. *Molecular microbiology* 48.1 (2003), pp. 77–84 (see p. 10).
- Schmitt, C. K., J. S. Ikeda, S. C. Darnell, P. R. Watson, J. Bispham, T. S. Wallis, D. L. Weinstein, E. S. Metcalf, and A. D. O’Brien (2001). “Absence of all components of the flagellar export and synthesis machinery differentially alters virulence of *Salmonella enterica* serovar Typhimurium in models of typhoid fever, survival in macrophages, tissue culture invasiveness, and calf enterocolitis”. *Infect Immun* 69.9 (Sept. 2001), pp. 5619–25 (see p. 75).

- Schuster, P., W. Fontana, P. F. Stadler, and I. L. Hofacker (1994). "From sequences to shapes and back: a case study in RNA secondary structures". *Proc Biol Sci* 255.1344 (Mar. 1994), pp. 279–84. DOI: 10.1098/rspb.1994.0040 (see p. 100).
- Schwan, W. R., X. Z. Huang, L. Hu, and D. J. Kopecko (2000). "Differential bacterial survival, replication, and apoptosis-inducing ability of *Salmonella* serovars within human and murine macrophages". *Infect Immun* 68.3 (Mar. 2000), pp. 1005–13 (see p. 57).
- Seth-Smith, H. M. (2008). "SPI-7: *Salmonella*'s Vi-encoding Pathogenicity Island". *Journal of infection in developing countries* 2.4 (2008), pp. 267–71 (see p. 25).
- Shea, J. E., M. Hensel, C. Gleeson, and D. W. Holden (1996). "Identification of a virulence locus encoding a second type III secretion system in *Salmonella typhimurium*". *Proceedings of the National Academy of Sciences of the United States of America* 93.6 (1996), pp. 2593–7 (see p. 25).
- Shelobolina, E. S., S. A. Sullivan, K. R. O'Neill, K. P. Nevin, and D. R. Lovley (2004). "Isolation, characterization, and U(VI)-reducing potential of a facultatively anaerobic, acid-resistant Bacterium from Low-pH, nitrate- and U(VI)-contaminated subsurface sediment and description of *Salmonella subterranea* sp. nov". *Appl Environ Microbiol* 70.5 (May 2004), pp. 2959–65 (see p. 20).
- Shevchenko, Y., G. G. Bouffard, Y. S. Butterfield, R. W. Blakesley, J. L. Hartley, A. C. Young, M. A. Marra, S. J. Jones, J. W. Touchman, and E. D. Green (2002). "Systematic sequencing of cDNA clones using the transposon Tn5". *Nucleic acids research* 30.11 (2002), pp. 2469–77 (see p. 6).
- Shippy, D. C., N. M. Eakley, C. T. Lauhon, P. N. Bochsler, and A. A. Fadl (2013). "Virulence characteristics of *Salmonella* following deletion of genes encoding the tRNA modification enzymes GidA and MnmE". *Microb Pathog* 57 (Apr. 2013), pp. 1–9. DOI: 10.1016/j.micpath.2013.01.004 (see p. 79).
- Sittka, A., S. Lucchini, K. Papenfort, C. M. Sharma, K. Rolle, T. T. Binnewies, J. C. Hinton, and J. Vogel (2008). "Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq". *PLoS genetics* 4.8 (2008), e1000163 (see pp. 27, 49, 50).
- Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences". *J Mol Biol* 147.1 (Mar. 1981), pp. 195–7 (see p. 61).
- Soneson, C. and M. Delorenzi (2013). "A comparison of methods for differential expression analysis of RNA-seq data". *BMC Bioinformatics* 14 (2013), p. 91. DOI: 10.1186/1471-2105-14-91 (see p. 68).
- Soper, G. A. (1939). "The Curious Career of Typhoid Mary". *Bulletin of the New York Academy of Medicine* 15.10 (1939), pp. 698–712 (see p. 24).
- Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov (2005). "Gene set enrichment analysis: a knowledge-based approach

## References

---

- for interpreting genome-wide expression profiles". *Proc Natl Acad Sci U S A* 102.43 (Oct. 2005), pp. 15545–50. DOI: 10.1073/pnas.0506580102 (see p. 73).
- Swiercz, J. P., Hindra, J. Bobek, J. Bobek, H. J. Haiser, C. Di Berardo, B. Tjaden, and M. A. Elliot (2008). "Small non-coding RNAs in *Streptomyces coelicolor*". *Nucleic Acids Res* 36.22 (Dec. 2008), pp. 7240–51. DOI: 10.1093/nar/gkn898 (see p. 99).
- Thomson, N. R., D. J. Clayton, et al. (2008). "Comparative genome analysis of *Salmonella Enteritidis* PT4 and *Salmonella Gallinarum* 287/91 provides insights into evolutionary and host adaptation pathways". *Genome research* 18.10 (2008), pp. 1624–37 (see pp. 23, 25).
- Thomson, N., S. Baker, D. Pickard, M. Fookes, M. Anjum, N. Hamlin, J. Wain, D. House, Z. Bhutta, K. Chan, S. Falkow, J. Parkhill, M. Woodward, A. Ivens, and G. Dougan (2004). "The role of prophage-like elements in the diversity of *Salmonella enterica* serovars". *Journal of molecular biology* 339.2 (2004), pp. 279–300 (see pp. 35–37).
- Tjaden, B., R. M. Saxena, S. Stolyar, D. R. Haynor, E. Kolker, and C. Rosenow (2002). "Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays". *Nucleic acids research* 30.17 (2002), pp. 3732–8 (see p. 49).
- Toft, C. and S. G. Andersson (2010). "Evolutionary microbial genomics: insights into bacterial host adaptation". *Nature reviews. Genetics* 11.7 (2010), pp. 465–75 (see p. 13).
- Tong, X., J. W. Campbell, G. Balazsi, K. A. Kay, B. L. Wanner, S. Y. Gerdes, and Z. N. Oltvai (2004). "Genome-scale identification of conditionally essential genes in *E. coli* by DNA microarrays". *Biochemical and biophysical research communications* 322.1 (2004), pp. 347–54 (see p. 33).
- Torarinsson, E. and S. Lindgreen (2008). "WAR: Webserver for aligning structural RNAs". *Nucleic Acids Res* 36.Web Server issue (July 2008), W79–84. DOI: 10.1093/nar/gkn275 (see p. 90).
- Torarinsson, E., Z. Yao, E. D. Wiklund, J. B. Bramsen, C. Hansen, J. Kjems, N. Tommerup, W. L. Ruzzo, and J. Gorodkin (2008). "Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions". *Genome Res* 18.2 (Feb. 2008), pp. 242–51. DOI: 10.1101/gr.6887408 (see p. 104).
- Townsend, S. M., N. E. Kramer, R. Edwards, S. Baker, N. Hamlin, M. Simmonds, K. Stevens, S. Maloy, J. Parkhill, G. Dougan, and A. J. Baumler (2001). "*Salmonella enterica* serovar Typhi possesses a unique repertoire of fimbrial gene sequences". *Infection and immunity* 69.5 (2001), pp. 2894–901 (see p. 25).

- Tsolis, R. M., A. J. Baumler, F. Heffron, and I. Stojiljkovic (1996). “Contribution of TonB- and Feo-mediated iron uptake to growth of *Salmonella typhimurium* in the mouse”. *Infection and immunity* 64.11 (1996), pp. 4549–56 (see p. 44).
- Unniraman, S., R. Prakash, and V. Nagaraja (2001). “Alternate paradigm for intrinsic transcription termination in eubacteria”. *J Biol Chem* 276.45 (Nov. 2001), pp. 41850–5. DOI: 10.1074/jbc.M106252200 (see pp. 95, 99).
- Unniraman, S., R. Prakash, and V. Nagaraja (2002). “Conserved economics of transcription termination in eubacteria”. *Nucleic Acids Res* 30.3 (Feb. 2002), pp. 675–84 (see pp. 95, 100).
- Urban, J. H. and J. Vogel (2008). “Two seemingly homologous noncoding RNAs act hierarchically to activate glmS mRNA translation”. *PLoS biology* 6.3 (2008), e64 (see p. 50).
- Van Dongen, S. (2008). “Graph clustering via a discrete uncoupling process”. *SIAM Journal on Matrix Analysis and Applications* 30.1 (2008), pp. 121–141 (see p. 108).
- Velden, A. W. van der, A. J. Bäumler, R. M. Tsolis, and F. Heffron (1998). “Multiple fimbrial adhesins are required for full virulence of *Salmonella typhimurium* in mice”. *Infect Immun* 66.6 (June 1998), pp. 2803–8 (see p. 56).
- Venables, W. N. and B. D. Ripley (1994). *Modern applied statistics with S-PLUS*. Vol. 250. Springer-verlag New York, 1994 (see p. 101).
- Vladoianu, I. R., H. R. Chang, and J. C. Pechère (1990). “Expression of host resistance to *Salmonella typhi* and *Salmonella typhimurium*: bacterial survival within macrophages of murine and human origin”. *Microb Pathog* 8.2 (Feb. 1990), pp. 83–90 (see p. 57).
- Vogel, J. (2009a). “A rough guide to the non-coding RNA world of *Salmonella*”. *Molecular microbiology* 71.1 (2009), pp. 1–11 (see pp. 25, 50, 52).
- Vogel, J. (2009b). “An RNA trap helps bacteria get the most out of chitosugars”. *Molecular microbiology* 73.5 (2009), pp. 737–41 (see pp. 49, 50).
- Vogel, J., V. Bartels, T. H. Tang, G. Churakov, J. G. Slagter-Jager, A. Huttenhofer, and E. G. Wagner (2003). “RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria”. *Nucleic acids research* 31.22 (2003), pp. 6435–43 (see pp. 47, 49).
- Wahl, R. and M. Kröger (1995). “ECDC—a totally integrated and interactively usable genetic map of *Escherichia coli* K12”. *Microbiol Res* 150.1 (Mar. 1995), pp. 7–61. DOI: 10.1016/S0944-5013(11)80034-0 (see pp. 89, 93).
- Wain, J., T. S. Diep, V. A. Ho, A. M. Walsh, T. T. Nguyen, C. M. Parry, and N. J. White (1998). “Quantitation of bacteria in blood of typhoid fever patients and relationship between counts and clinical features, transmissibility, and antibiotic resistance”. *Journal of clinical microbiology* 36.6 (1998), pp. 1683–7 (see pp. 25, 44).

## References

---

- Waldminghaus, T., N. Heidrich, S. Brantl, and F. Narberhaus (2007). “FourU: a novel type of RNA thermometer in *Salmonella*”. *Molecular microbiology* 65.2 (2007), pp. 413–24 (see p. 25).
- Wan, X.-F., G. Lin, and D. Xu (2006). “Rnall: an efficient algorithm for predicting RNA local secondary structural landscape in genomes”. *J Bioinform Comput Biol* 4.5 (Oct. 2006), pp. 1015–31 (see pp. 85, 93).
- Wan, X.-F. and D. Xu (2005). “Intrinsic terminator prediction and its application in *Synechococcus* sp. WH8102”. *Journal of Computer Science and Technology* 20.4 (2005), pp. 465–482 (see pp. 85, 93).
- Washio, T., J. Sasayama, and M. Tomita (1998). “Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination”. *Nucleic acids research* 26.23 (1998), pp. 5456–5463 (see pp. 85, 95).
- Wassarman, K. M., F. Repoila, C. Rosenow, G. Storz, and S. Gottesman (2001). “Identification of novel small RNAs using comparative genomics and microarrays”. *Genes & development* 15.13 (2001), pp. 1637–51 (see p. 49).
- Weinberg, Z., J. E. Barrick, Z. Yao, A. Roth, J. N. Kim, J. Gore, J. X. Wang, E. R. Lee, K. F. Block, N. Sudarsan, S. Neph, M. Tompa, W. L. Ruzzo, and R. R. Breaker (2007). “Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline”. *Nucleic Acids Res* 35.14 (2007), pp. 4809–19. DOI: [10.1093/nar/gkm487](https://doi.org/10.1093/nar/gkm487) (see pp. 94, 104).
- Weinberg, Z. and R. R. Breaker (2011). “R2R—software to speed the depiction of aesthetic consensus RNA secondary structures”. *BMC Bioinformatics* 12 (2011), p. 3. DOI: [10.1186/1471-2105-12-3](https://doi.org/10.1186/1471-2105-12-3) (see pp. 97, 109).
- Weinberg, Z., J. X. Wang, J. Bogue, J. Yang, K. Corbino, R. H. Moy, and R. R. Breaker (2010). “Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes”. *Genome Biol* 11.3 (2010), R31. DOI: [10.1186/gb-2010-11-3-r31](https://doi.org/10.1186/gb-2010-11-3-r31) (see p. 104).
- Weinstein, D. L., M. Carsiotis, C. R. Lissner, and A. D. O’Brien (1984). “Flagella help *Salmonella typhimurium* survive within murine macrophages”. *Infect Immun* 46.3 (Dec. 1984), pp. 819–25 (see p. 75).
- Will, S., K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen (2007). “Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering”. *PLoS Comput Biol* 3.4 (Apr. 2007), e65. DOI: [10.1371/journal.pcbi.0030065](https://doi.org/10.1371/journal.pcbi.0030065) (see p. 90).
- Winkler, W., A. Nahvi, and R. R. Breaker (2002). “Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression”. *Nature* 419.6910 (2002), pp. 952–6 (see pp. 47, 49).

- Workman, C. and A. Krogh (1999). "No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution". *Nucleic acids research* 27.24 (1999), pp. 4816–22 (see p. 87).
- Wu, D. et al. (2009). "A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea". *Nature* 462.7276 (2009), pp. 1056–60 (see pp. 17, 20).
- Yao, Z., Z. Weinberg, and W. L. Ruzzo (2006). "CMfinder—a covariance model based RNA motif finding algorithm". *Bioinformatics* 22.4 (Feb. 2006), pp. 445–52. DOI: 10.1093/bioinformatics/btk008 (see pp. 90, 97, 101, 104).
- Yim, L., I. Moukadiri, G. R. Björk, and M.-E. Armengod (2006). "Further insights into the tRNA modification process controlled by proteins MnmE and GidA of *Escherichia coli*". *Nucleic Acids Res* 34.20 (2006), pp. 5892–905. DOI: 10.1093/nar/gkl752 (see p. 79).
- Yokoseki, T., K. Kutsukake, K. Ohnishi, and T. Iino (1995). "Functional analysis of the flagellar genes in the fliD operon of *Salmonella typhimurium*". *Microbiology* 141 ( Pt 7) (July 1995), pp. 1715–22 (see p. 75).
- Yu, J. and J. S. Kroll (1999). "DsbA: a protein-folding catalyst contributing to bacterial virulence". *Microbes Infect* 1.14 (Dec. 1999), pp. 1221–8 (see p. 80).
- Yu, J., E. E. Oragui, A. Stephens, J. S. Kroll, and M. M. Venkatesan (2001). "Inactivation of DsbA alters the behaviour of *Shigella flexneri* towards murine and human-derived macrophage-like cells". *FEMS Microbiol Lett* 204.1 (Oct. 2001), pp. 81–8 (see p. 80).
- Zhang, Y. J., T. R. Ioerger, C. Huttenhower, J. E. Long, C. M. Sassetti, J. C. Sacchettini, and E. J. Rubin (2012). "Global assessment of genomic regions required for growth in *Mycobacterium tuberculosis*". *PLoS pathogens* 8.9 (2012), e1002946 (see pp. 4, 9–11, 14, 15).
- Zhou, D. and J. Galán (2001). "Salmonella entry into host cells: the work in concert of type III secreted effector proteins". *Microbes Infect* 3.14-15 (2001), pp. 1293–8 (see p. 56).
- Zomer, A., P. Burghout, H. J. Bootsma, P. W. Hermans, and S. A. van Hijum (2012). "ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data". *Plos one* 7.8 (2012), e43012 (see p. 11).
- Zuker, M. and P. Stiegler (1981). "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information". *Nucleic Acids Res* 9.1 (Jan. 1981), pp. 133–48 (see p. 85).
- Zurawski, G., K. Brown, D. Killingly, and C. Yanofsky (1978). "Nucleotide sequence of the leader region of the phenylalanine operon of *Escherichia coli*". *Proceedings of the National Academy of Sciences of the United States of America* 75.9 (1978), pp. 4271–5 (see p. 50).