Characterisation of the Transcriptome and Proteome of *Salmonella enterica* subspecies *enterica* serovar Typhi.

Thesis submitted to the University of Cambridge for the degree of Doctor of Philosophy

2nd January 2009

Timothy Trevor Perkins
Jesus College

Declaration

I hereby declare that this thesis consists of work derived entirely of my own work. Due to the nature of this work, and interdisciplinary nature of biological sciences, it was not possible and impractical to perform all of these techniques, however, it was possible to design all experiments associated with this thesis. Work that was done by other persons is clearly stated in the Materials and Methods section.

This thesis is no longer than 300 pages as required by the Department of Biological Sciences.

Abstract

Salmonella enterica subspecies *enterica* serovar Typhi (*S.* Typhi) is the cause of human typhoid. A combination of approaches were utilised to analyse the transcriptome and proteome of this human-restricted and highly clonal pathogen.

A novel method for strand-specific RNA-seq was developed, facilitating a whole genome analysis of the *S.* Typhi Ty2 BRD948 transcriptome. These data were validated using novel analyses methods and compared to RNA derived from an *ompR* mutant derivative. The data was used to reinterpret a previous annotation of the *S.* Typhi Ty2 genome. Mass Spectrometry sequenced peptides were mapped back to the genome to further enhance the quality of the annotated data of the *S.* Typhi genome.

DNA microarray analysis was used to quantify differences in gene expression between wild type and a *S.* Typhi Ty2 *ompR* mutant and identify genes, which were significantly differentially expressed in both transcriptome sequencing and microarray experiments. A subset of genes that were OmpR-regulated, hypothetical and absent from *E. coli* were identified to evaluate their virulence potential using *S.* Typhimurium mutant derivatives and *in vitro* assays. Protein homology searches were used to explore the potential function of several novel OmpR-regulated proteins. Finally, applying new DNA sequencing technology allowed development of a chIP-seq protocol to identify OmpR-binding sites.

“The last thing I'll say to the people that don't believe in cycling – the cynics, the sceptics, I’m sorry for you. I’m sorry you can’t dream big and I’m sorry you don’t believe in miracles.”

Lance Armstrong, 2005.

Acknowledgements

First and foremost I would like to thank my parents, Jenny and Phil, who encouraged me throughout my formative years to believe in myself and to take whichever opportunities arose, no matter how large or small. Their guidance and counsel gave me the confidence to accept this challenge offered by Professor Gordon Dougan (Doog).

To Doog, many thanks for introducing me to molecular microbiology, training me and above all, persisting with me, I hope I have upheld your belief.

Most importantly I would like to thank my best friend and wife, Kura who has stood by me for all these years, always believing. Kura has been a rock and inspiration to me since we met and has always offered unconditional support to all of my pursuits.

During my PhD, I have received substantial guidance from Dr Robert Kingsley (Dr Bob). He has honed my scientific skills over the four years I've been at the Sanger Institute, many thanks Dr Bob. I would also like to thank rest of Doog's team, those who have directly helped with experiments or those who have just put things away I forgetfully left around the lab.

Finally, many thanks to the Wellcome Trust who have underwritten this body of work.

Declaration 1
Abstract 2
Acknowledgements 4
Tables 10
Figures 11
Abbreviations 13
1 Introduction 15
1.1 The genus *Salmonella* 16
1.1.1 Classification and nomenclature 16
1.2 Disease syndromes caused by *Salmonella* 18
1.3 Salmonella pathogenicity 21
1.3.1 The biology of *Salmonella* infection 22
1.3.2 Typhoid vaccines 25
1.4 The genomics of invasive *Salmonella* 27
1.5 Genetics of *Salmonella* disease 33
1.5.1 TTSSs 33
1.5.2 The virulence (Vi) antigen 39
1.5.3 Horizontally acquired DNA 40
1.5.4 Adherence 41
1.5.5 Motility 42
1.5.6 Two-component regulators and virulence 42
1.6 Regulation of bacterial gene expression 44
1.6.1 Transcriptional activation 45
1.6.2 Post-transcriptional control 49
1.7 DNA sequencing technologies 49
1.7.1 DNA microarrays 49
1.7.2 High-capacity DNA sequencing 51
1.8 *In silico* - software and tools 52
1.8.1 Artemis and Artemis Comparison Tool (ACT) 53
1.8.2 BLAST and other packages 53
1.9 Aims of the Thesis 53
2 Materials and Methods 54
2.1 General chemicals, reagents and buffers 55
2.2 Bacterial strains 55
2.3 Plasmids 55
2.4 Microbiological media and techniques 56
2.4.1 Bacterial culture 56
2.4.1.1 Standard methods 56
2.4.1.2 Minimal media 58
2.4.2 Harvesting of bacteria from broth cultures 58
2.4.3 Growth curves 58
2.4.4 Identification of bacteria 59
2.5 Molecular biology techniques 59
2.5.1 Bacterial genomic DNA extraction 59
2.5.2 Plasmid DNA extraction 60
2.5.3 Agarose gel electrophoresis 60
2.5.4 Quantification of DNA and RNA by spectroscopy 61

2.5.5 DNA manipulation techniques 61
2.5.5.1 DNA restriction 61
2.5.5.2 Oligonucleotides 61
2.5.5.3 Polymerase chain reaction (PCR) 64
2.5.5.3.1 PCR supermix (Invitrogen) 64
2.5.5.3.2 PFU polymerase high fidelity PCR 65
2.5.5.3.3 Real time PCR 65
2.5.5.3.4 Overlap extension PCR 66
2.5.5.4 Purification of PCR products 66
2.5.5.5 Cloning 66
2.5.5.5.1 Cloning PCR amplicons 66
2.5.5.5.2 Sub-cloning sequences into pWT12 67
2.5.6 Precipitation of DNA 67
2.5.7 DNA sequencing 67
2.5.7.1 Capillary-based sequencing 67
2.5.7.2 Illumina sequencing 68
2.5.8 DNA transformation 69
2.5.8.1 Transformation of chemically competent cells 69
2.5.8.1.1 Preparation of chemically competent cells 69
2.5.8.1.2 Chemical transformation 69
2.5.8.2 Transformation by electroporation 70
2.5.8.2.1 Preparation of electrocompetent cells 70
2.5.8.2.2 Electroporation 70
2.5.8.3 Screening for transformed DNA 71
2.5.8.3.1 Screening for non recombinant DNA 71
2.5.8.3.2 Screening for recombinant DNA 71
2.5.9 Bacterial conjugation 71
2.5.10 Transduction 72
2.5.10.1 Infection of donor strain with P22 phage 72
2.5.10.2 Isolation of P22 phage containing genomic DNA 72
2.5.10.3 Infection of recipient strain with P22 phage 72
2.5.11 Mutagenesis 73
2.5.11.1 Red recombinase mutagenesis 73
2.5.11.2 FLP recombinase mutagenesis 74
2.5.11.3 Allelic exchange mediated by suicide vector constructs 74
2.5.12 RNA methodologies 75
2.5.12.1 RNA stabilisation 75
2.5.12.2 RNA isolation 75
2.5.12.2.1 RNA isolation for microarray 75
2.5.12.2.2 RNA isolation for Illumina sequencing 76
2.5.12.2.3 rRNA removal 76
2.5.12.3 Removal of DNA from RNA samples 76
2.5.12.4 Reverse transcription of RNA 77
2.5.12.4.1 Reverse transcription of RNA for aminoallyl labelling 77
2.5.12.4.2 Reverse transcription of RNA for Illumina sequencing and real time PCR 77
2.5.12.5 Labelling of RNA for microarray 77
2.5.12.5.1 Purification of cDNA 77
2.5.12.5.2 Coupling of CyDye™ ester to dUTP 78
2.5.12.5.3 Purification of labelled cDNA 78

2.5.13 DNA microarray 79
2.5.13.1 Hybridisation and scanning 79
2.5.13.2 Gridding and feature extraction 79
2.5.14 Protein methodologies 80
2.5.14.1 Preparation of protein 80
2.5.14.1.1 Whole cell lysate 80
2.5.14.1.2 Cellular fractionation 80
2.5.14.1.2.1 Cell lysis 80
2.5.14.1.2.2 Precipitation of the soluble fraction 81
2.5.14.1.2.3 Separation of the membrane fractions 81
2.5.14.1.3 Heat shock 82
2.5.14.1.4 Preparation of proteins for mass spectroscopy analysis 82
2.5.14.2 One-dimensional SDS-PAGE gel electrophoresis of proteins 82
2.5.14.3 Protein visualisation 82
2.5.14.3.1 Coomassie blue protein stain 82
2.5.14.3.2 Western blot 83
2.5.14.4 Immunoprecipitation for chIP-seq 83
2.5.14.5 chIP-seq data mapping and analysis 83
2.6 Murine model of Salmonellosis 84
2.6.1 Competitive infections 84
2.7 LC-MS analysis 85
2.8 Bioinformatics tools 87
2.8.1 Artemis and ACT 87
2.8.2 Mapping software 89
2.8.2.1 Transcriptome mapping software 89
2.8.2.1 Protein mapping scripts 91
2.8.3 Perl scripts 91
2.8.3.1 Pileup to stranded plot - maqpileup2depth.pl 91
2.8.3.2 Local plots 91
2.8.3.3 Extracting expression data from plot 91
2.8.4 R 92
2.8.4.1 Limma analysis 92
2.8.5 Genespring 93
2.8.6 Graph pad prism 5.0 93
3 A Method for RNA-seq in prokaryotes 94
3.1 Aims of the work described in this Chapter 95
3.2 Introduction 95
3.3 Results 96
3.3.1 Directional Sequencing of Single Stranded cDNA 96
3.3.2 Sequencing of the *S.* Typhi transcriptome 100
3.3.3 Mapping and visualising data 103
3.4 Discussion and conclusion 105
4 Deep sequencing of the *S.* Typhi Ty2 transcriptome 107
4.1 Aims of the work described in this Chapter 108
4.2 Introduction 108
4.3 Results 109
4.3.1 Mapping Sequence Reads to the Annotated Ty2 Genome 109
4.3.2 Coding Sequences 112
4.3.2.1 General Features 112
4.3.2.2 Virulence Genes 118

4.3.2.3 Phage and putative cargo genes 121
4.3.2.4 Pseudogenes 122
4.3.2.5 Hypothetical Genes 123
4.3.2.6 A *S.* Typhi Ty2 Specific Insertion 123
4.3.3 Non-Coding Sequences 124
4.3.4 Transcriptome comparison with *ompR*-null mutant 129
4.3.4.1 Quality control 129
4.3.4.2 General features of the BRD948 *ompR* transcriptome 133
4.3.4.3 Quantified differences 134
4.3.4.4 Quantified differences pre-Benjamini Hochberg correction 137
4.4 Discussion 137
4.4.1 General transcriptome results 137
4.4.2 OmpR comparison 139
4.5 Conclusion 141
5 Global transcriptome analysis of an *S.* Typhi *ompR*-null Mutant. 142
5.1 Aims of this chapter 143
5.2 Introduction 143
5.3 Results 145
5.3.1 Analysis of the *S.* Typhi OmpR regulon, preliminary work 145
5.3.2 Microarray analysis 145
5.3.2.1 Biological conditions, quality control and analysis 145
5.3.2.2 The OmpR regulon during exponential growth (OD_{600}=0.6) 148
5.3.2.3 The OmpR regulon during early stationary phase (OD_{600}=1.1) 150
5.3.2.4 The OmpR regulon grown under SPI-2 inducing conditions 151
5.3.3 Real time PCR confirmation of expression profile 152
5.3.4 Succinate dehydrogenase enzyme assay 153
5.3.5 chIP-seq to determine ompR-binding domains 155
5.3.5.1 Rabbit anti-OmpR 155
5.3.5.2 Construction of an OmpR:3xFLAG 155
5.3.5.3 chIP-Seq Quality Control 156
5.3.5.4 chIP-Seq using the Illumina sequencing platform 157
5.3.5.6 Comparison with genes differentially transcribed in microarray data 159
5.4 Discussion 160
5.4.1 Microarray Data 160
5.4.1.1 Regulators 161
5.4.1.2 Membrane transport 163
5.4.1.3 Respiration 165
5.4.1.4 Chemotaxis and motility 167
5.4.1.5 Genes of unknown function 167
5.4.1.6 Indirect comparison with RNA-seq data 168
5.4.2 chIP-seq 170
5.5 Conclusion 171
6 Characterisation of Hypothetical Genes in the OmpR Regulon 172
6.1 Aims of this chapter 173
6.2 Introduction 173
6.3 Results 174
6.3.1 *In silico* analyses 174
6.3.1.1 t1787-t1793 exhibited increased expression in the *ompR* mutant 175
6.3.1.2 t4498-4501 exhibited decreased expression in the *ompR* mutant 177

6.3.2 Identification of proteins using LC-MS 178
6.3.3 Construction of recombinant *S.* Typhimurium mutants in which homologues of t1787-1790, t1791-1793 and t4498-4501 are deleted 180
6.3.4 *in vitro* Phenotyping of Mutants 182
6.3.4.1 LPS silver stained gel 182
6.4.4.2 Growth on different carbon sources 183
6.4.5 *in vivo* phenotyping 184
6.4.5.1 Competitive infection assays 184
6.5 Discussion 187
6.5.1 RAK103 and RAK105 characterisation 187
6.5.2 TT56.1 phenotype 190
6.6 Conclusion 190
7 General Discussion 192
8 References 196
9 Appendices 216
9.1 Maqdeth2pileup.pl 216
9.2 extractLines.pl 218
9.3 tram.pl 220
9.4 LIMMA Analysis for Illumina data 221
9.5 Hypothetical gene sequence coverage 227
9.6 Non-coding RNA details 281
9.7 Differential expression pre-Benjamini-Hochberg correction 288
9.8 Microarray Differences 297
9.9 chIP data 302

Tables

Table 1.1 Salmonella nomenclature in use at CDC, 2000 [4] 17
Table 1.2 Three years cumulative risk of typhoid fever (RR, 95% CI) [66] 27
Table 1.3 Genes unique to pairs of genomes [82] 31
Table 1.4. Effectors of the SPI1- and SPI2-encoded type III secretion systems [108] 36
Table 1.5 Major families of bacterial transcription factors [164] 46
Table 2.1 Strains used in this study 55
Table 2.2 Plasmids used in this study 56
Table 2.3 Antibiotic concentration and solvents 58
Table 2.4 Primers used in this study 62
Table 3.1 Read mapping for different depletion methods 102
Table 4.1 Analysis of sequences mapped to the Ty2 genome 110
Table 4.2. Sequence data mapped for each experiment 130
Table 4.3 Statistically different genes in the *ompR* mutant 135
Table 5.1 Genes differentially expressed in microarray experiments and with chIP-seq enrichment sites mapped upstream 159
Table 5.2 Experimentally verified and published genes in the OmpR regulon of *S.* Typhi or *S.* Typhimurium identified as being differentially regulated in the microarray experiments (2-fold, $p<0.05$) 161
Table 6.1 Differentially expressed hypothetical genes and the published annotation 174
Table 6.2 Differentially expressed hypothetical genes re-annotated according to functional identity derived using pBLAST and Pfam database searches 178
Table 6.3 Growth on specific sole carbon sources 184

Figures

Figure 1.1 Disease syndromes and subspecies [7]. 19
Figure 1.2 WHO estimates of the global burden of typhoid fever and its geographical distribution [8] 20
Figure 1.3 The horizontal transfer of virulence-associated genes within the genus *Salmonella* [78]. 29
Figure 1.4 Analysis of the P4-like bacteriophage ST46 [89]. 32
Figure 1.5 Model for substrate recognition and delivery of proteins by TTSS machines [96]. 34
Figure 1.6 Cellular phenotypes associated with the function of SPI-2 [113]. 38
Figure 1.7 Genetic structure of the viaB locus in *S.* Typhi Ty2. 40
Figure 1.8 Genetic content of horizontally acquired regions in S. Typhi CT18. 41
Figure 1.9 Some of the different combinations of histidine protein kinase (HPK) and aspartate response regulator (RR) domains in histidine–aspartate phosphorelay (HAP) systems [166]. 48
Figure 1.10 Schematic of Illumina sequencing technology (http://www.illumina.com/). 52
Figure 3.1 The hypotheses proposed to account for the attachment of Illumina adapted dimers to sscDNA 98
Figure 3.2 Single stranded sequencing 99
Figure 3.3 Quantitative comparison of terminator exonuclease and oligonucleotide hybridisation rRNA depletion techniques. 103
Figure. 3.4 Representation of transcriptomic sequence read coverage plot relative to genome annotation, as displayed in Artemis 105
Figure 4.1 Circular plot of mapped sequence data. 111
Figure 4.2 Artemis representation of transcriptome plot. 112
Figure 4.3 Defined deletions in BRD948. 112
Figure 4.4 Sense and Antisense outliers. 114
Figure 4.5. Transcripts mapping to well characterised regions. 115
Figure 4.6 Consistency between sequencing experiments. 115
Figure 4.7 “Mosaic coding”. 116
Figure 4.8. Putative error in the annotation. 116
Figure 4.9. Functional classification of sequenced transcripts. 117
Figure 4.10. Attenuation. 118
Figure 4.11 Coverage for SPI-1. 119
Figure 4.12 Coverage across SPI-2. 120
Figure 4.13. Sequenced transcripts identifying cargo genes within *Salmonella* prophage. 122
Figure 4.14. Third Party Annotation of the Ty2-specific Insertion. 124
Figure 4.15 Coverage of putative non-coding RNAs. 126
Figure 4.16. SPI-1 Sequence coverage 128
Figure 4.17. Paralogues of RUF107c. 129
Figure 4.18 Quality control. 132
Figure 4.19 Assigning transcriptome to functional class. 133
Figure 4.20. Start of the *viaB* locus. 134
Figure 4.21 Outer membrane porin C. 134
Figure 5.1 Quality control for expression microarrays. 147

Figure 5.2 Numbers of genes differentially transcribed by functional class..... 149
Figure 5.3 Circular plot of the S. Typhi genome and genes differentially transcribed in the *ompR* mutant. 150
Figure 5.4 Comparison of expression values determined by real-time PCR and microarray..... 153
Figure 5.5 Succinate dehydrogenase assay. 154
Figure 5.6 Schematic of overlap extension PCR to incorporate 3XFLAG peptide into the C-terminal region of OmpR. 156
Figure 5.7 Western blot of immunoprecipitated OmpR:3XFLAG..... 157
Figure 5.8 Transporter transcript changes..... 164
Figure 5.9 Respiration..... 166
Figure 5.10 Genes of unknown function..... 168
Figure 5.11 Expression ratio for genes found to be differentially expressed in microarray and deep-sequencing experiments..... 169
Figure 6.1. Comparative alignment of regions predicted to be OmpR regulated in *S.* Typhi. 176
Figure 6.2 The chamical structure of related acetylated aminosugars..... 177
Figure 6.3 Sequenced peptides mapped to regions predicted to be OmpR-regulated. 180
Figure 6.4 Details of mutagenesis for further phenotypic characterisation. 182
Figure 6.5 LPS silver stained gel..... 183
Figure 6.6 Control competitive infections. 186
Figure 6.7 Comparison of mutants RAK103 and TT56.1 with RAK113 (SL1344 Δ*phoN*)..... 186

Abbreviations

Acronym	Definition
ACT	Artemis comparison tool
ADS	arginine deiminase
AM	arithmetic mean per base-pair
ATP	adenosine tri-phosphate
BLAST	basic local alignment search tool
bp	base-pair
CDC	Centre for Disease Control
cDNA	complementary DNA
CDS	coding sequence
CFU	colony forming units
CGH	comparative genome hybridisation
chIP	chromatin immunoprecipitation
CL3	containment level 3
DEPC	diethylpyrocarbonate
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
DTT	dithiothreitol
E. coli	*Escherichia coli*
EDTA	ethylenediaminetetraacetic acid
EGTA	ethylene glycol tetraacetic acid
FDR	false discovery rate
FNR	false non-discovery rate
gDNA	genomic DNA
HAP	histidine aspartate phosphorelay
HIV	human immunodeficiency virus
HPK	histidine protein kinase
JCICSB	Judicial Commission of the International Committee of Systematic Bacteriology
LASER	light amplification by stimulated emission of radiation
LB	Luria-Bertani
LC-MS	liquid chromatography mass spectroscopy
LPS	lipopolysaccharide
M cells	Microfold cells
MOPS	3-(N-morpholino)propanesulfonic acid
mRNA	messenger RNA
NAG	N-acetyl glucosamine
NAM	N-acetylmuramic acid
NB	Note
nc	nucleotide
NCBI	National Center for Biotechnology Information
ncRNA	non-coding RNA
Nramp-1	natural resistance-associated macrophage protein one
NTS	Non-typoidal Salmonella
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate buffered saline

PCR	polymerase chain reaction
PMN	polymorphonuclear
RES	reticuloendothelial system
RNA	ribonucleic acid
RR	response regulator
rRNA	ribosomal RNA
RUF	RNA of unknown function
S.	*Salmonella enterica* serovar *enterica*
SCV	*Salmonella* containing vacuole
SDS	sodium dodecyl sulfate
SPI	*Salmonella* Pathogenicity Island
STM	signature tagged mutagenesis
TCA	trichloroacetic acid
tRNA	transfer RNA
TTSS	type three secretion system
UDP	uridine diphosphate
USA	United States of America
UTR	untranslated region
UV	ultraviolet
WHO	World Health Organisation
WT	wild-type
WTSI	Wellcome Trust Sanger Institute