Genomic variation and evolution of *Salmonella enterica* serovars Typhi and Paratyphi A

Kathryn Holt

Wolfson College, University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

August 2009

Abstract

Salmonella enterica serovars Typhi and Paratyphi A are bacterial pathogens that cause typhoid fever in humans. Typhi and Paratyphi A are unusual among *S. enterica* serovars, as they are restricted to systemic infection of humans while most serovars cause gastroenteritis in a broad range of animal hosts. Despite their similarities, Typhi and Paratyphi A are thought to have evolved independently, adapting to the human systemic niche via mechanisms which are still poorly understood. There is little genetic variation within each population, making it difficult to study their evolution or population dynamics.

In this thesis, comparative genomic analysis was used to detect variation within the Typhi and Paratyphi A populations, and to compare the evolution of these two pathogens. A total of 19 complete Typhi genome sequences were compared in order to identify genetic variants, including single nucleotide mutations (SNPs), deletions and insertions of novel DNA. A different approach was taken to study the Paratyphi A population, including the comparison of seven complete genome sequences and development of a novel technique to screen for SNPs in a collection of 160 genomes sequenced in pools. Little evidence was found of selection upon Typhi genes, but there was evidence of diversifying selection in genes coding for the biosynthesis of O-antigen in Paratyphi A. There was evidence in both populations of ongoing accumulation of inactivating mutations which result in loss of gene function. Detailed comparison of this functional gene loss in Typhi and Paratyphi A revealed that many of the same genes were inactivated in both serovars, but the mutations occurred independently and were not the result of horizontal transfer of DNA between their genomes. Comparative analysis of variation in the Typhi and Paratyphi A populations suggested that

Paratyphi A is the younger pathogen, with a most recent common ancestor roughly a third as old as that of Typhi.

Bacteria can harbour plasmids (additional strands of circular DNA) that carry genes encoding resistance to drugs. The plasmids are able to spread between bacterial cells, thereby spreading drug resistance within or between pathogen populations. In this thesis, comparative analysis of plasmid sequences from Typhi and Paratyphi A found that the same type of plasmid was present in both serovars, carrying identical DNA sequences encoding resistance to the drugs used to treat typhoid fever. This demonstrates that the evolution of drug resistance in both serovars is tightly linked. Very closely related sequences were also found in other human bacterial pathogens, highlighting how easily drug resistance can spread.

Single nucleotide variants (SNPs) identified in Typhi and in the drug resistance plasmids were used to develop a high-throughput SNP typing assay with which to study Typhi populations. The SNP typing assay was used to interrogate a global collection of Typhi, as well as local Typhi populations from areas where typhoid is endemic, including regions of Vietnam, Nepal, India and Kenya. The analysis linked strain type with plasmid type for the first time, and demonstrated multiple independent acquisitions of distinct drug resistance plasmids over the past 40 years, culminating in the current dominance of a single plasmid type. Analysis of recent Typhi populations circulating in endemic areas showed that the same Typhi clone now dominates all of these regions, although local diversification has resulted in subtle differences between the populations. Importantly, the dominant Typhi clone was closely associated with the dominant plasmid type, suggesting that the success of the clone and plasmid may have been intimately linked.

Declaration

This dissertation is my own work and contains nothing
which is the outcome of work done in collaboration with others,
except as specified in the text and Acknowledgements.

The thesis work was conducted from May 2006 to August 2009
at the Wellcome Trust Sanger Institute, Cambridge, UK
under the supervision of
Gordon Dougan (Wellcome Trust Sanger Institute),
Julian Parkhill (Wellcome Trust Sanger Institute), and
Duncan Maskell (Department of Veterinary Medicine,
University of Cambridge).

To my parents,
who introduced me to the world of science,

and to my husband Mike,
who made it possible to stay.

Acknowledgements

The work presented in this thesis would not have been possible without the efforts of dozens of colleagues, collaborators and friends. My work relies heavily on access to large amounts of sequence data from hundreds of bacterial isolates. For the most part, this data was generated at the Sanger Institute thanks to funding from the Wellcome Trust and the efforts of the sequencing teams, including the Pathogen Sequencing Unit and the many teams involved in the Solexa pipeline. All the Typhi 454 and Solexa sequencing was done prior to the establishment of the pipeline, by Ian Goodhead and Richard Rance. The bacterial isolates themselves were made available by a generous network of international collaborators. I am most indebted to Francois-Xavier Weill at the Pasteur Institute in Paris, who provided not only hundreds of Typhi and Paratyphi A DNA samples for sequencing and SNP typing but also ideas, information and additional experimentation on the isolates in his collection. Additional Paratyphi A DNA samples were provided by Shanta Dutta (National Institute of Cholera and Enteric Diseases, Kolkata), Rajni Gaind (Safdarjung Hospital, Delhi), Rumina Hasan (Aga Khan University Hospital, Karachi) and Sam Kariuki (Kenya Medical Research Institute, Nairobi), thanks to collaborations with John Wain (Sanger Institute and Health Protection Agency, UK). Shanta Dutta and Sam Kariuki also provided extensive collections of Typhi DNA for SNP typing studies, as did Christiane Dolecek (Oxford University Clinical Research Unit, Ho Chi Minh City) and Andrew Pollard (Oxford University, UK). My high throughput SNP typing studies were made possible by the efforts of the Sanger Institute's genotyping teams, in particular Panos Deloukas, Rhian Gwilliam and Ranganath Bangalore Venkatesh. Additional supporting data for the Kathmandu study was provided by Stephen Baker (Sanger Institute and Oxford University Clinical Research Unit, Ho Chi Minh City), Andrew

Pollard (Oxford University, UK), Dominic Kelly (Oxford University, UK) and David Murdoch (University of Otago, Christchurch, NZ). Yajun Song and Mark Achtman (Environmental Research Institute, Cork) kindly performed additional SNP typing of *gyrA* loci in order to complete this study.

The Typhi genome sequencing project was the brainchild of Julian Parkhill, Gordon Dougan, Mark Achtman, Philippe Roumagnac, Stephen Baker and John Wain who, along with Camila Mazzoni (Environmental Research Institute, Cork), assisted in the design and interpretation of the study. Most of the same people were involved in developing the Paratyphi A sequencing study. Nicholas Thomson taught me all aspects of genome annotation and tricks in genome-level sequence analysis, while Zemin Ning was most helpful in teaching me to use his ssaha software for the analysis of new sequence data. Heng Li (Sanger Institute) and Yik Y Teo (Wellcome Trust Centre for Human Genetics, Oxford, UK) were kind enough to modify their wonderful software programmes, Maq and Illuminus, to better suit my analysis needs. Yik Y Teo also helped me immensely whenever I reached the limits of my undergraduate statistics knowledge. John Wain and Minh Duy Phan shared their knowledge and expertise on IncHI1 plasmids and Minh Duy Phan performed several PCR experiments in which we shared an interest. Many of the DNA samples I analysed were prepared at the Sanger Institute by Stephen Baker, Robert Kingsley, Derek Pickard and Satheesh Nair, all of whom also helped me get to grips with the small amount of lab work involved in my project, along with Sally Whitehead and Minh Duy Phan. My PhD supervisors Gordon Dougan, Julian Parkhill and Duncan Maskell each provided extraordinary support and guidance throughout my project, sharing their invaluable knowledge, time and resources. Finally, I am indebted to the Wellcome Trust, who support not only my research, university fees and living expenses, but the research of most of my colleagues and collaborators as well.

Contents

List of Figures **xii**

List of Tables **xvi**

Glossary **xviii**

1 Introduction **1**

1.1 The organisms: *Salmonella enterica* serovars Typhi and Paratyphi A . . 2

1.1.1 The genus *Salmonella* 2

1.1.1.1 Classification and taxonomy 2

1.1.1.2 Host range and pathogenicity 3

1.1.2 *Salmonella* genetics and evolution 4

1.1.2.1 Surface structures and antigens 6

1.1.2.2 *Salmonella* Pathogenicity Islands 9

1.1.2.3 Horizontal gene transfer: 11

1.1.3 Serovar Typhi 16

1.1.4 Serovar Paratyphi A 18

1.2 The disease: enteric fever 20

1.2.1 Pathology and clinical features 20

1.2.2 Asymptomatic carriage 22

1.2.3 Diagnostics 23

1.2.4 Epidemiology 24

1.2.5 Antibiotic treatment and resistance 26

1.2.6 Prevention 28

1.3 The approach: comparative and population genomics 30

1.3.1 Population genetics of bacterial pathogens 30

1.3.1.1 Evolution and variation in pathogen populations 30

1.3.1.2 Methods for studying bacterial pathogen populations . 32

1.3.2 Genome sequencing of bacterial pathogens 34

1.3.2.1 Comparative genomics 36

1.3.2.2 SNP analysis 38

1.3.2.3 New high throughput sequencing technology 42

1.4 Thesis outline 45

2 Genomic sequence variation in Typhi 47

2.1 Introduction 47

2.1.1 Aims 51

2.2 Methods 52

2.2.1 Bacterial strains and DNA 52

2.2.2 DNA sequencing 53

2.2.3 Plasmid identification 53

2.2.4 Phylogenetic analysis 55

2.2.5 $\frac{dN}{dS}$ calculations 55

2.2.6 Transition bias 56

2.2.7 Detection of recombination events 56

2.2.8 Evidence of expression from published microarray data 57

2.2.9 Accession codes 57

2.3 Results 58

2.3.1 Assessment of SNP detection methods 58

2.3.1.1 454 data: comparison of SNP detection from reads or *de novo* assembled contigs 58

2.3.1.2 Solexa data 63

2.3.1.3 Determining quality filters for SNP detection 64

2.3.1.4 Estimating error rates for SNP detection 64

2.3.1.5 Minimisation of potential errors 67

2.3.2 SNP analysis 69

2.3.2.1 $\frac{dN}{dS}$ in the Typhi population 70

2.3.2.2 Potential signals of selection 73

2.3.2.3 Recombination 76

2.3.3 Gene acquisition 78
2.3.3.1 Prophage sequences 78
2.3.3.2 Genomic islands 79
2.3.3.3 Plasmids 80
2.3.4 Loss of gene function 80
2.3.4.1 Genomic deletions 80
2.3.4.2 Accumulation of pseudogenes 82
2.4 Discussion 84
2.4.1 Strengths and limitations of the study 84
2.4.2 Differences between Typhi lineages 88
2.4.2.1 Antibiotic resistance and the H58 lineage 90
2.4.3 Insights into the evolution of Typhi 92
2.4.3.1 Adaptive selection in Typhi genes 92
2.4.3.2 Evolutionary dynamics of the Typhi population 94

3 Genomic sequence variation in Paratyphi A 96

3.1 Introduction 96
3.1.1 Aims 97
3.2 Methods 98
3.2.1 Identification of repetitive and horizontally transferred sequences in the Paratyphi A genome 98
3.2.2 SNP detection 98
3.2.3 Phylogenetic network analysis 98
3.2.4 Detection of insertion/deletion events and plasmid sequences . . 99
3.2.5 Gene ontology analysis 99
3.2.6 Accession codes 100
3.3 Results 100
3.3.1 Comparison of seven Paratyphi A genome sequences 100
3.3.1.1 Whole genome sequencing 100
3.3.1.2 SNP analysis 100
3.3.1.3 Gene acquisition 102
3.3.1.4 Insertion/deletion mutations 105
3.3.1.5 Loss of gene function 107

3.3.2 Optimisation of SNP detection from pooled sequence data 109
3.3.2.1 SNP detection and frequency estimation 109
3.3.2.2 Comparison of potential methods 111
3.3.2.3 Performance of optimised method 116
3.3.2.4 Performance of optimised method over a range of read depths 118
3.3.3 Genomic variation detected in 159 Paratyphi A isolates by pooled sequencing 120
3.3.3.1 SNP detection 122
3.3.3.2 Distribution of SNPs among pools 123
3.3.3.3 Distribution of SNP frequencies 125
3.3.3.4 Distribution of SNPs in the Paratyphi A genome 128
3.3.3.5 Novel pseudogene-forming mutations 133
3.3.3.6 Detection of IncHI1 plasmids 134
3.3.3.7 Detection of plasmid pGY1 137
3.4 Discussion 139
3.4.1 Strengths and limitations of the study 139
3.4.2 Genomic variation and possibilities for typing in the Paratyphi A population 142
3.4.3 Adaptive selection in Paratyphi A genes 144

4 Convergent evolution of Typhi and Paratyphi A 146
4.1 Introduction 146
4.1.1 Aims 150
4.2 Methods 151
4.2.1 Whole genome comparisons 151
4.2.2 Phylogenetic network analysis 151
4.2.3 Bayesian analysis of recombined and non-recombined genes . . . 151
4.2.4 Time estimation using dS 154
4.2.5 Comparison and annotation of pseudogenes 155
4.2.6 Data simulation 156
4.3 Results 156
4.3.1 Evolution of Typhi and Paratyphi A 156

4.3.2 Convergent features of the Typhi and Paratyphi A genomes . . . 162
4.3.2.1 Shared genes 162
4.3.2.2 Comparison of pseudogenes in Typhi and Paratyphi A . 165
4.3.2.3 Genes missing from Typhi and Paratyphi A 166
4.3.2.4 Features shared with Paratyphi C 168
4.3.3 The role of recombination 169
4.3.3.1 Sharing of unique genes and deletions by recombination 169
4.3.3.2 Sharing of pseudogenes by recombination 170
4.3.4 Pseudogene formation in the evolutionary histories of Typhi and Paratyphi A 173
4.3.4.1 Pseudogene formation over time 173
4.3.4.2 Pseudogenes potentially involved in host adaptation . . 176
4.4 Discussion 178
4.4.1 Strengths and limitations of the study 178
4.4.2 Implications for host restriction and adaptation 181
4.4.2.1 Ancestral pseudogenes 183
4.4.2.2 Pseudogenes and novel genes shared by recombination . 184
4.4.2.3 Recent pseudogenes: convergence after recombination . 184
4.4.2.4 Ongoing accumulation of strain-specific pseudogenes . . 186

5 IncHI1 multidrug resistance plasmids in Paratyphi A and Typhi 187
5.1 Introduction 187
5.1.1 Aims 189
5.2 Methods 190
5.2.1 Annotation 190
5.2.2 Sequence comparison and SNP detection 191
5.2.3 Phylogenetic analysis 192
5.2.4 PCR 192
5.2.5 Accession codes 192
5.3 Results 194
5.3.1 Characterisation of IncHI1 plasmid backbone and resistance gene insertions 194
5.3.1.1 The conserved IncHI1 backbone 195

5.3.1.2 Comparison of drug resistance genes in pAKU_1 and pHCM1 197
5.3.1.3 A composite resistance transposon 201
5.3.1.4 Other insertions in pAKU_1 202
5.3.2 Evolution of IncHI1 plasmids and MDR 204
5.3.2.1 Phylogenetic analysis of the IncHI1 plasmid backbone . 204
5.3.2.2 Drug resistance insertions in IncHI1 plasmids 207
5.4 Discussion 209
5.4.1 IncHI1 plasmids in Paratyphi A and Typhi 209
5.4.2 Acquisiton of MDR by IncHI1 plasmids 210
5.4.3 The spread of MDR via IncHI1 plasmids 212

6 Investigating Typhi populations using high throughput SNP typing 214
6.1 Introduction 214
6.1.1 Aims 218
6.2 Methods 219
6.2.1 DNA preparation and quantitation 219
6.2.2 Illumina GoldenGate assay 220
6.2.3 Genotype calling from raw data 221
6.2.3.1 Genotype calling with Illuminus-P 223
6.2.3.2 Heuristic to identify 'no signal' cluster 223
6.2.3.3 Clustering across plates 223
6.2.4 Phylogenetic analysis of genotyping data 224
6.2.5 Visualisation of temporal and spatial data 226
6.2.6 Simpson's diversity index 227
6.3 Results 228
6.3.1 Validation of GoldenGate assay for target loci in Typhi 228
6.3.1.1 Chromosomal loci 228
6.3.1.2 IncHI1 plasmid loci 232
6.3.1.3 Other target loci 236
6.3.2 Validation of GoldenGate SNP typing in a global collection of Typhi isolates, previously typed at 88 loci 236
6.3.2.1 Phylogenetic analysis of chromosomal SNPs 236

6.3.2.2 IncHI1 plasmids and multidrug resistance 239
6.3.2.3 Distribution of other plasmids 243
6.3.3 Endemic typhoid in the Mekong Delta, Vietnam 243
6.3.3.1 Phylogenetic analysis 245
6.3.3.2 Plasmids and drug resistance 247
6.3.3.3 Spatial and temporal distribution 248
6.3.4 Pediatric typhoid in Kathmandu, Nepal 250
6.3.4.1 Phylogenetic analysis 250
6.3.4.2 Drug resistance 252
6.3.4.3 Temporal distribution of haplotypes 253
6.3.5 Endemic typhoid in an urban slum in Kolkata, India 254
6.3.5.1 Phylogenetic analysis 255
6.3.5.2 Spatial and temporal distribution of haplotypes 257
6.3.5.3 Association with the vaccination programme 262
6.3.6 The Typhi population in Nairobi, Kenya over a 21 year period . 263
6.3.7 Typhi H58 and the IncHI1 ST6 plasmid 265
6.3.8 Typhi H58 and mutations in GyrA 272
6.4 Discussion 272
6.4.1 Strengths and limitations of the study 272
6.4.2 Typhi populations in endemic areas 275
6.4.3 The evolution of drug resistance in Typhi 278

7 Final discussion 281

References 287

A Inactivating mutations in Typhi 321

B Paratyphi A isolates sequenced in pools 327

C Genes with >2 SNPs more than expected among Paratyphi A pools 333

D Variable pseudogenes in the Paratyphi A population 338

E Typhi isolates used for SNP typing 342

List of Figures

1.1 Model for the evolution of virulence in the genus *Salmonella* 5

1.2 Structure of a *Salmonella* cell, flagellum and cell wall. 7

1.3 Methods of DNA transfer 12

1.4 Genome rearrangements and phage differences between Typhi CT18 and Ty2 17

1.5 Biology of *Salmonella* infection 21

1.6 Trends in enteric fever incidence in the UK, 1990-2008 25

1.7 Timeline of the use of, and development of resistance to, antibiotics in enteric fever 27

1.8 Complete bacterial genome sequences deposited in public databases . . . 35

1.9 Sequence assembly 36

1.10 Phylogenetic discovery bias 41

2.1 Phylogenetic tree guiding selection of Typhi isolates for sequencing . . . 50

2.2 Error models for 454 reads 60

2.3 Read depth vs genome coverage for real and simulated 454 data 60

2.4 Read depth vs mean contig size for real and simulated 454 data 61

2.5 Distribution of quality parameters for SNP detection 65

2.6 Phylogenetic tree of Typhi based on SNP data 71

2.7 Trajectory of $\frac{dN}{dS}$ over time in Typhi 73

2.8 Distribution of number of SNPs per Typhi gene 74

2.9 Distribution of prophage, IS elements and deletions in the Typhi genome and phylogenetic tree 78

2.10 Coverage of SPI7 in Typhi isolate E98-3139 82

2.11 Accumulation of gene-inactivating mutations in Typhi lineages 84

2.12 Distribution of number of SNPs between pairs of Typhi lineages 89
2.13 Distribution of number of deletions, prophage and pseudogenes between pairs of Typhi lineages 90

3.1 Phylogenetic tree of seven Paratyphi A isolates based on genome-wide SNPs detected by sequencing 103
3.2 Phylogenetic network of seven Paratyphi A isolates including seven serovars as outgroups. 104
3.3 Distribution of SNPs per Paratyphi A gene 104
3.4 Detection of small indels from short read data 109
3.5 Sensitivity and error rates for different weighting measures 112
3.6 Ranges for each accuracy measure 113
3.7 Sensitivity and false positive rates for different assembly parameters . . 114
3.8 Accuracy measures for different pileup parameters 114
3.9 Accuracy measures for different weighting equations 116
3.10 Expected frequencies vs SNP frequencies estimated from Paratyphi A test pool sequence data 117
3.11 Distributions of sample standard deviations calculated among SNPs with correct and incorrect frequency estimates 118
3.12 Error rates expected at different levels of read depth 119
3.13 Numbers of SNPs detected in each Paratyphi A pool 124
3.14 Distribution of SNPs detected uniquely in pools MA6 and MA10. 125
3.15 Distribution of estimated Paratyphi A SNP frequencies 126
3.16 Distribution of frequencies across pools for SNPs originally detected among seven individually-sequenced Paratyphi A isolates 127
3.17 Pool-wide frequencies of SNPs defining different branches of the seven-strain phylogenetic tree of Paratyphi A 128
3.18 Distribution of SNPs within the Paratyphi A genome 129
3.19 $\frac{dN}{dS}$ plotted against SNP frequency in Paratyphi A 130
3.20 Number of SNPs per gene in Paratphi A from pools 131
3.21 Distribution of *wba* cluster SNPs in Paratyphi A pools 133
3.22 IncHI1 SNPs detected in Paratyphi A pools 136
3.23 Phylogenetic network of pGY1 plasmids detected in Paratyphi A pools . 137

3.24 Difference in SNP frequencies given biased and unbiased sampling . . . 140
3.25 Distribution of number of SNPs between two Paratyphi A lineages . . . 142
3.26 Distribution of number of deletions, prophage and pseudogenes between two Paratyphi A lineages 143

4.1 Phylogenetic trees for *Salmonella enterica* 157
4.2 Phylogenetic trees for *Salmonella* and *E. coli* with divergence time estimates 159
4.3 Phylogenetic trees for *S. enterica* with divergence time estimates 160
4.4 Pseudogenes, recombined genes, and unique genes in the Typhi and Paratyphi A genomes 164
4.5 Overlap of pseudogenes in Typhi, Paratyphi A and Paratyphi C 166
4.6 Scenarios of recombination and pseudogene formation in Paratyphi A and Typhi 171
4.7 Pseudogene accumulation in Typhi and Paratyphi A over time 175

5.1 Functions of genes annotated in the IncHI1 plasmid pAKU_1 from Paratyphi A 194
5.2 Comparison of three complete IncHI1 plasmid sequences from *Salmonella* 196
5.3 Transposons identified in pAKU_1 198
5.4 Rearrangements of composite transposons inserted in IncHI1 plasmids . 199
5.5 Distribution of SNPs in the pAKU_1 IncHI1 plasmid 205
5.6 Phylogenetic trees of IncHI1 plasmids based on sequence data 206

6.1 Typhoid incidence around the world and SNP typing study sites 217
6.2 Example cluster plots for genotyping assays 222
6.3 Phylogenetic analysis workflow for SNP typing studies 225
6.4 Effect of assay failure on relative branch lengths for Typhi chromosomal SNPs 230
6.5 Distribution of assayed SNPs in the Typhi CT18 chromosome 231
6.6 Distribution of assayed SNPs in the IncHI1 plasmid 233
6.7 Effect of assay failure on relative branch lengths for IncHI1 plasmid SNPs234
6.8 Phylogenetic trees for a global collection of 180 Typhi isolates (1958-2005)237
6.9 Discrimination within known Typhi haplotypes 238

6.10 Distribution of IncHI1 plasmid SNPs detected in MDR and drug sensitive isolates . . . 239
6.11 Phylogenetic trees of Typhi chromosomes and IncHI1 plasmids in a global collection of Typhi isolates . . . 241
6.12 Geographical sources of isolates from the Mekong Delta . . . 244
6.13 Phylogenetic distribution of Typhi isolates from the Mekong Delta . . . 246
6.14 Distribution of IncHI1 plasmids among Typhi isolates from the Mekong Delta . . . 247
6.15 Distribution of haplotypes among provinces in the Mekong Delta . . . 248
6.16 Distribution of typhoid fever cases over two years in the Mekong Delta . 249
6.17 Phylogenetic distribution of Typhi isolates from Kathmandu . . . 251
6.18 Distribution of patient ages for H58-G vs other haplotypes detected in Kathmandu . . . 252
6.19 Distribution of typhoid cases in Kathmandu by month . . . 253
6.20 Phylogenetic distribution of Typhi isolates from Kolkata . . . 256
6.21 Distribution of typhoid cases during a four year study in Kolkata . . . 258
6.22 Spatial clustering of typhoid cases in Kolkata . . . 259
6.23 Spatial clustering during typhoid peak-incidence periods in Kolkata . . . 260
6.24 Distribution of Typhi cases among households in Kolkata . . . 261
6.25 Frequency distribution of Typhi haplotypes before and after the introduction of a Vi conjugate vaccine in Kolkata . . . 262
6.26 Distribution of Typhi haplotypes in Nairobi, Kenya . . . 264
6.27 Distribution of H58 subtypes among Typhi isolates from four regional collections . . . 266
6.28 Distribution of IncHI1 plasmids in time and space . . . 268
6.29 Distribution of IncHI1 plasmids and *IS*1 among Typhi haplotypes . . . 269
6.30 Distribution of IncHI1 plasmids and *IS*1 among Typhi H58 subtypes . . 270
6.31 GyrA SNPs distributed among Typhi H58 subtypes . . . 271

List of Tables

1.1 *Salmonella* species, subspecies and serovars 3

1.2 Genetic similarity within *Salmonella* and among closely related genera . 5

1.3 Genes unique to pairs of *Salmonella* and *E. coli* genomes 19

1.4 Throughput and accuracy of next-generation sequencing technologies . . 45

2.1 Typhi isolates sequenced in this study 52

2.2 Primers used for PCR and sequencing of deletions and insertion sites in the Typhi gneome 55

2.3 Error rates in SNP detection using simulated sequence data 62

2.4 SNPs detected in Typhi 454 data by analysis of contigs and reads 63

2.5 Thresholds for filters used during SNP calling 64

2.6 Estimated measures of SNP detection accuracy 66

2.7 Repetitive Typhi CT18 sequences excluded from SNP detection anlaysis 68

2.8 Genetic variation detected in 19 Typhi genomes 72

2.9 Genes with potential signals of adaptive selection 75

2.10 Recombination events detected in Typhi isolates 77

2.11 Genomic deletions detected in this study 81

2.12 Drug resistance phenotypes and genetic variants for sequenced Typhi isolates 91

3.1 Paratyphi A isolates with whole genome sequence data available 101

3.2 Repetitive Paratyphi A AKU_12601 sequences excluded from SNP detection anlaysis 102

3.3 Insertion/deletion mutations detected between two Paratyphi A AKU_12601 genomes 106

3.4 Pseudogene-forming mutations detected among seven Paratyphi A genomes 108
3.5 Analysis of variance for factors affecting accuracy of SNP detection and frequency estimation 115
3.6 Solexa sequence data for Paratyphi A pools 121
3.7 SNP clusters detected in Paratyphi A pools 123
3.8 Genes containing at least five more SNPs than expected by chance . . . 132
3.9 IncHI1 plasmids detected in pools 135
3.10 Ratio of read depths for IncHI1 plasmids and *Salmonella* chromosomes . 136
3.11 pGY1 plasmids detected in Paratyphi A pools 138

4.1 *S. enterica* serovar genomes 152
4.2 Genes unique to Typhi and/or Paratyphi A 163
4.3 Pseudogenes shared between Paratyphi A and Typhi 167
4.4 Genes absent from Typhi and Paratyphi A but present in 11 other serovars 168
4.5 Distribution of serovar-specific and shared pseudogenes in recombined regions 172
4.6 Strain-specific pseudogenes shared between Paratyphi A and Typhi . . . 174
4.7 Pseudogenes in Typhi and Paratyphi A associated with secreted effectors, fimbriae or transmembrane domains 176

5.1 Functional categories for genome annotation 191
5.2 PCR primers for analysis of IncHI1 plasmids 193
5.3 Resistance gene insertions in IncHI1 plasmids determined from sequence data and PCR 208

6.1 Study sites for SNP typing of localised Typhi populations 218
6.2 SNPs for detection of resistance genes and IncHI1 plasmid deletions . . 235
6.3 Typhoid case parameters by Typhi haplotype in Kathmandu 251
6.4 Typhoid case parameters by Typhi haplotype in Kolkata 255
6.5 Summary of Typhi populations from endemic areas 277

Glossary

bp	Base pairs
CDS	Protein-coding sequences
contig	Contiguous sequence assembled from overlapping reads
Gb	Gigabase pairs (1 billion bp)
GTR	General time reversible substitution model
homoplasy	Identity by state but not by descent
IncHI1	Plasmid incompatibility type HI1
indel	Insertion/deletion mutation
IS	Insertion sequence
IVI	International Vaccine Instute, Seoul, South Korea
kbp	Kilobase pairs (1 thousand bp)
KEMRI	Kenya Medical Research Institute, Nairobi, Kenya
LPS	Lipopolysaccharide
Mbp	Megabase pairs (1 million bp)
MCMC	Markov chain Monte Carlo
MDR	Multiple drug resistance, defined as resistance to chloramphenicol, ampicillin and co-trimoxazole
MIC	Minimum inhibitory concentration, defined as the minimum concentration of an antimicrobial that can inhibit the visible growth of a microorganism
MLST	Multi-locus sequence typing
mrca	Most recent common ancestor
Mya	Million years ago
Nal	Nalidixic acid
NICED	National Institute for Cholera and Enteric Diseases, Kolkata, India
NTS	Non-typhoidal salmonellosis
OUCRU	Oxford University Clinical Research Unit, Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam
PFGE	Pulsed-field gel electrophoresis
PSU	Pathogen Sequencing Unit at the Wellcome Trust Sanger Institute, Cambridge, UK
SNP	Single nucleotide polymorphism
SPI	*Salmonella* Pathogenicity Island
Tn	Transposon
TTSS	Type III secretion system
VNTR	Variable number tandem repeat