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

List of Tables

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

Chapter 1

Introduction

Salmonella enterica serovars Typhi and Paratyphi A are closely related bacteria that cause typhoid and paratyphoid fever. They were first described in the late 19th century, although they have probably been causing disease in humans for thousands of years (1, 2). Transmitted by fecal contamination of water or food, these bacteria have been responsible for epidemics all over the world. In addition to causing typhoid fever, infection occasionally results in long-term carriage of the bacteria in the human gall bladder (3). These carriers remain healthy themselves, but can unwittingly spread typhoid to those around them, some famous examples being 'Typhoid Mary', who infected at least 50 people (4), and 'Mr N The Milker', who spread typhoid to more than 200 people over 16 years (5). Tracing the sources of typhoid outbreaks - usually human carriers or contaminated water sources - is a sleuthing exercise that has kept doctors and scientists busy from the 19th century (5) to the present day (6, 7). However direct transmission is hard to prove, as epidemiologically unrelated Typhi isolates are often so similar as to look identical using most typing techniques (2, 6, 8). The incidence of typhoid fever decreased dramatically in the developed world during the twentieth century as sanitation improved (9), but remains high in developing countries where access to clean water is poor (10). Still, thousands of typhoid cases are reported in developed countries each year, often associated with travel to areas where the disease is more common, including India, South Asia, South America and parts of Africa (11, 12). Vaccines against typhoid were developed by the British army in the late 19th century and remained in use until the 1980s (13). Safer and more effective vaccines were developed in the 1980s and currently two are licensed for use (14), however they are almost exclusively used by travellers (15) and are not appropriate for immunising small children (14). The introduction of antibiotics proved effective in the treatment of typhoid fever and is the mainstay of disease control in areas where the disease is endemic (3). However, over the past 40 years, an increasing number of typhoid cases have become resistant to an increasing number of drugs (16, 17). The evolution of resistance to new drugs can be rapid, and poses a major problem for disease control (17). Recent advances in sequence analysis provide new opportunities to study the evolution of Typhi and Paratyphi A at the DNA level - the finest resolution possible - and it is this opportunity that will be explored in this thesis.

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

1.1.1 The genus *Salmonella*

1.1.1.1 Classification and taxonomy

Salmonella is a genus of bacteria belonging to the family *Enterobacteriaceae*, and includes many pathogens responsible for disease in humans and other animals. The genus *Salmonella* is divided into two species, *bongori* and *enterica* (18, 19). *Salmonella enterica* is further divided into six subspecies *enterica, salamae, arizonae, diarizonae, houtenae* and *indica*, which contain over 2,500 serovars or serotypes (see Table 1.1) (18, 19, 20). Subspecies divisions were initially based on biochemical properties and nucleotide similarity (18, 21, 22) and are supported by more recent sequence data (23). Serovars are defined by their O (somatic) antigen and H (flagellar) antigens, with antigenic formulae written as: O antigens; H antigens (phase 1, phase 2) (18). The official list of serovars, known as the Kauffmann-White scheme (19), is maintained and regularly updated by the WHO Collaborating Centre for Reference and Research on *Salmonella*. The majority of disease-associated salmonellae are serovars of *S. enterica* subspecies *enterica* (19). Most serovars have been given names, usually referring to the geographic location from which they were first isolated, which are correctly written unitalicised and beginning with a capital letter (18). While their formal names are of the form *S. enterica* subspecies *enterica* serovar Typhi, they are often shortened to the form *S. enterica* serovar Typhi, *S.* Typhi or simply referred to by the serovar name, e.g. Typhi. For brevity in this thesis, serovars of *S. enterica* subspecies *enterica* will be introduced as serovars of *S. enterica* and thereafter referred to using only the serovar name.

Species	**Subspecies**	**Serovars**
S. enterica		
	subsp. enterica	1504
	subsp. salamae	502
	subsp. arizonae	95
	subsp. diarizonae	333
	subsp. houtenae	72
	subsp. indica	13
S. bongori		22
Total		**2541**

Table 1.1: *Salmonella* species, subspecies and serovars - Serovars defined under each of seven subspecies of *Salmonella*, taken from the last update to the Kauffman-White scheme in 2002 (19, 20).

1.1.1.2 Host range and pathogenicity

Although over 1,500 serovars of *S. enterica* subspecies *enterica* have been defined (Table 1.1, (19)), the pathogenicity of most remains uncharacterised. The majority of *Salmonella*-associated disease in humans and domestic animals is caused by a relatively small number of serovars (19, 24), which vary in their host ranges and disease syndromes. Some serovars cause gastroenteritis in a broad range of host species, for example Typhimurium and Enteritidis are responsible for 40-90% of foodborne salmonellosis in humans in many parts of the world (24, 25, 26, 27, 28, 29, 30), as well as the majority of infections in domestic animals (25). Other serovars are host-adapted, primarily associated with systemic disease in a small range of host species but also associated with relatively infrequent disease in other animals. For example, serovars Dublin and Choleraesuis are generally associated with systemic disease in cattle and pigs respectively, but can also cause infections in humans and other animals (24, 31, 32). Similarly Typhimurium, a frequent cause of gastroenteritis in humans (25), can cause systemic infection in mice. Finally, serovars can be host-restricted, causing systemic disease in a narrow range of closely related species. Serovars Typhi and Paratyphi A are restricted to humans and and other simians (higher primates) (33), causing systemic disease in the form of enteric fever (detailed in 1.2). Occasionally, host-generalist serovars or those adapted to non-human hosts are also able to cause invasive or systemic disease in humans (known collectively as invasive non-typhoidal *Salmonella* or invasive NTS) (24, 34). This may be attributed to bacterial virulence traits, immune deficiencies in the human host, or a combination of such factors (35, 36, 37) and can vary between geographic locations (24, 34). For example Typhimurium and Enteritidis are associated with high rates of invasive NTS in children and HIV-infected adults in parts of Africa (35, 38, 39), while Choleraesuis is a major cause of invasive NTS in Taiwan (40).

1.1.2 *Salmonella* genetics and evolution

Salmonella diverged from *Escherichia coli* approximately 100 million years ago (41, 42, 43, 44, 45). The circular chromosomes of *Salmonella* and *E. coli* are generally 4.5-5 Mbp in size, encode $\sim$4,500 genes (46, 47, 48, 49, 50, 51) and are genetically very similar, sharing $\sim$70% of their genes with 80% identity at the nucleotide level and 90% identity at the amino acid level (46, 50, 52), see Table 1.2. The acquisition of genomic islands, known as *Salmonella* Pathogenicity Islands (SPIs), are considered to be key steps in the evolution of *Salmonella* (53) (see Figure 1.1). SPI1 is present in both species of *Salmonella* (*S. bongori* and *S. enterica*) but is absent from *E. coli*, consistent with a single acquisition by the common ancestor of all extant *Salmonella* (54). SPI2 is present in *S. enterica* but is absent from *S. bongori* (55), consistent with acquisition by the common ancestor of *S. enterica* after divergence from *S. bongori* (see Figure 1.1). Variation in genes required for antigen biosynthesis has led to the differentiation of at least 1,500 serovars (19). Each serovar has accumulated additional chromosomal diversity via point mutations as well as gain and loss of genes (on average, serovars share $\sim$90% of their genes at >98% nucleotide identity, see Table 1.2 (46, 47, 48, 49, 50, 52, 56, 57), leading to diversity in host range and pathogenicity. Horizontal transfer between serovars and even subspecies, via homologous recombination, phage integration and plasmid transfer (detailed in 1.1.2.3) blurs the lines between serovars (determined by variation in the antigen synthesis genes), pathogenicity (determined by gene content and allelic variation) and taxonomy (intended to represent vertical patterns of descent).

Organism	Data source	Homologs of Typhimurium*	Median DNA similarity	Median amino acid similarity
S. enterica serovar				
- Typhimurium LT2	Sequence	100%	100%	100%
- Typhi CT18	Sequence	89%	98%	99%
- Paratyphi A	Sequence[a,b]	87-89%	98%	99%
- Paratyphi B	Microarray	92%	-	-
S. arizonae	Microarray	83%	-	-
S. bongori	Microarray	85%	-	-
E. coli K-12	Sequence	71%	80%	90%
E. coli O157:H7	Sequence	73%	80%	90%
K. pneumoniae	Sequence[a]	73%	76%	88%

Table 1.2: Genetic similarity within *Salmonella* and among closely related genera - Reproduced from (50). Original legend: "*For sequenced genomes, reciprocal best hits, excluding unsampled regions; for microarrays, signal ratio of Typhimurium LT2 with genome is 3:1 or greater and based on roughly 4,330 Typhimurium LT2 coding sequences." [a]97% complete sequence. [b]Microarray.

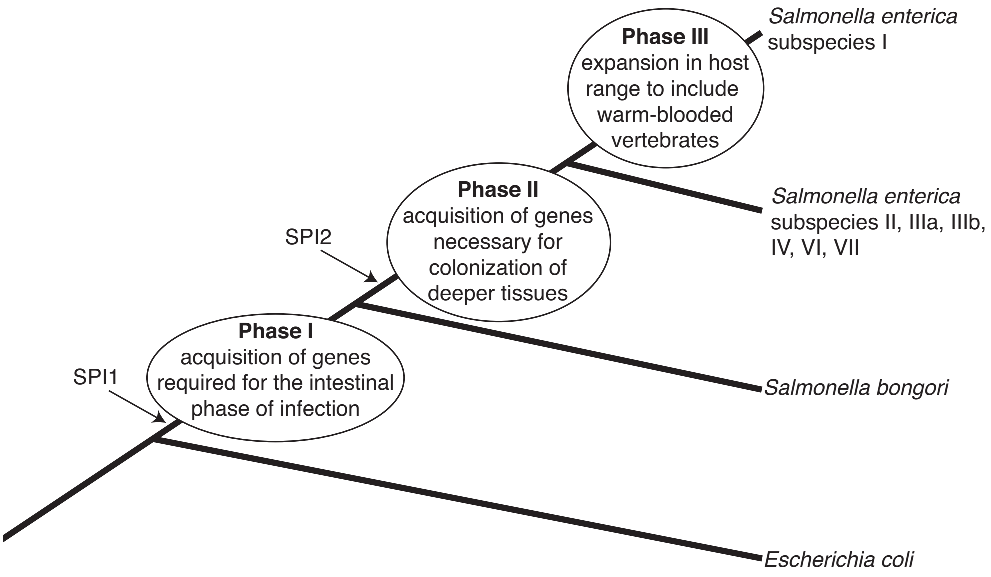

Phase III
expansion in host range to include warm-blooded vertebrates
Salmonella enterica subspecies I
Phase II
acquisition of genes necessary for colonization of deeper tissues
Salmonella enterica subspecies II, IIIa, IIIb, IV, VI, VII
SPI2
Phase I
acquisition of genes required for the intestinal phase of infection
SPI1
Salmonella bongori
Escherichia coli

Figure 1.1: Model for the evolution of virulence in the genus *Salmonella* - Reproduced from (53), with SPI1 and SPI2 insertion points added. Original legend: "The three phases in which virulence evolved in the genus *Salmonella* since its divergence from the *E. coli* lineage have been proposed previously (58). The phylogenetic tree is not drawn to scale." Note *Salmonella enterica* subspecies I is an alternative designation for *S. enterica* subspecies *enterica*

1.1.2.1 Surface structures and antigens

Salmonella, like all bacteria, synthesise a variety of surface structures (see Figure 1.2). These protect the cell, have roles in transport, adhesion, chemotaxis and motility, act as receptors, are important for host immune responses and form the basis for identification by serotyping in the laboratory.

Surface lipopolysaccharide (LPS) or O-antigen forms the outer leaflet of the outer membrane. It comprises a membrane-embedded lipid component, an oligosaccharide core and a long chain polysaccharide consisting of 10-30 repeats of polysaccharide units comprising 2-6 sugars (O-units), see Figure 1.2c (59). Biosynthesis of the O-antigen is encoded in a cluster of genes known as the *wba* cluster (previously known as the *rfb* cluster) (60, 61). The *wba* cluster includes genes for the biosynthesis of sugars (*wba* genes), glycosyl transferases which add sugars sequentially to generate the O-unit (*wba* genes) and O-antigen processing genes which translocate the O-units across the inner membrane and polymerise them into a long chain O-antigen (*wzx, wzy, wzz* genes) (59, 62). Additional modifications of the O-unit can be made by acetyl transferases and glycosyl transferases encoded outside the *wba* cluster (62). Over 50 O-antigens have been identified in *Salmonella* (19), which vary in the nature of the sugars that make up the O-unit, their order and linkages (62). These variations in structure reflect genetic variation in the *wba* cluster (59, 61, 63, 64), which has a mosaic structure indicative of evolution by horizontal transfer between bacterial species (65). The oligosaccharide core is encoded in another cluster known as the *waa* locus, variation in which is associated with structural variation in the oligosaccharide core (66).

Salmonella also express an O-antigen exopolysaccharide (extracellular polysaccharide), made up of O-units similar to those present in O-antigen LPS (67). Biosynthesis of the O-antigen capsule is dependent on genes outside the *wba* cluster (*yihU-yshA* and *yihV-yihW*), which are required for the formation of biofilm on gallstones (important for establishing long-term asymptomatic carriage in mammalian hosts) (68) and for attachment to plants (associated with foodborne transmission) (69).

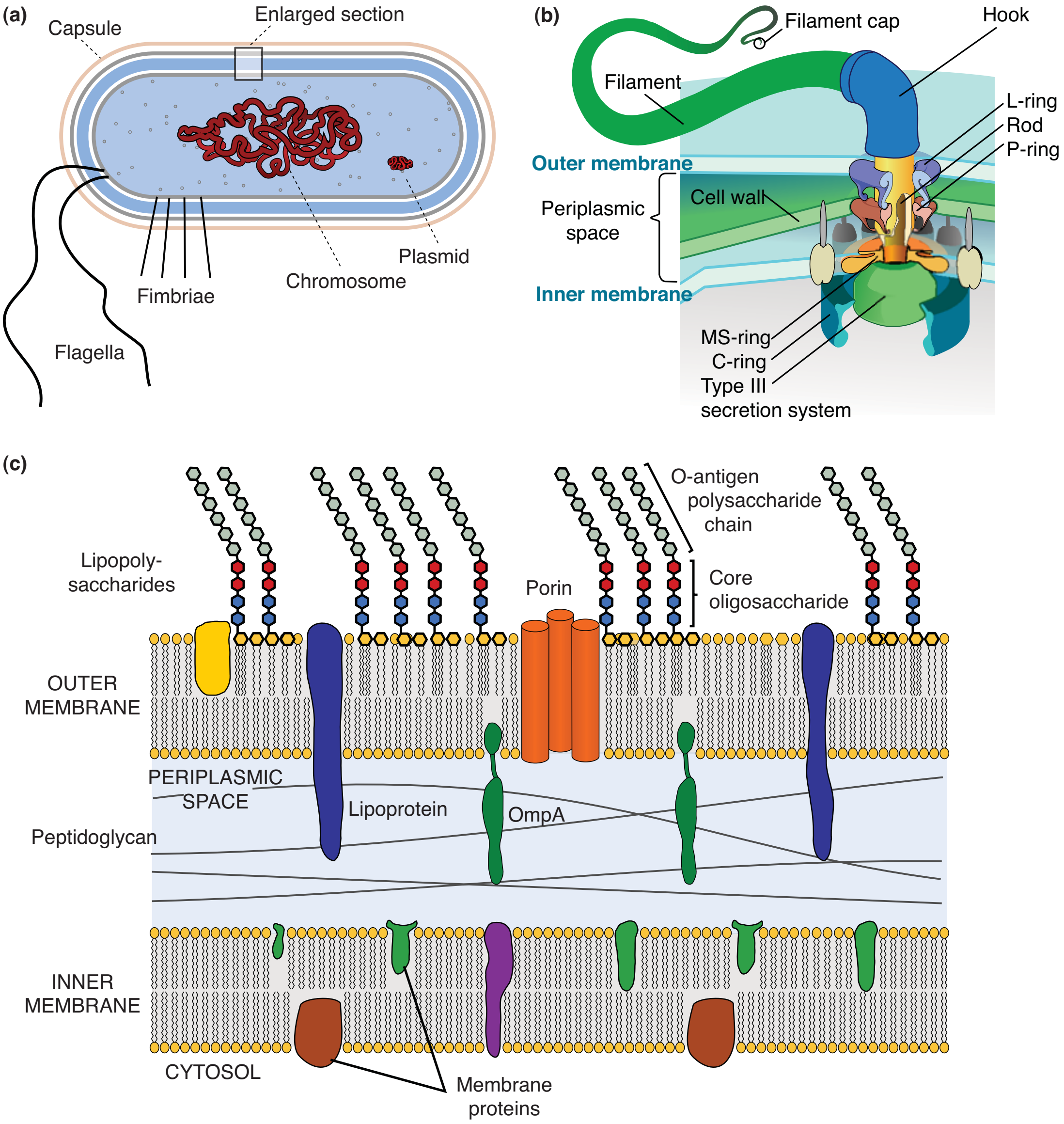


Figure 1.2: Structure of a *Salmonella* cell, flagellum and cell wall. - (a) Structure of a cell, section is enlarged in (c). (b) Structure of a flagellum. (c) Structure of the *Salmonella* cell wall. Reproduced from drawings by Jeff Dahl (a,c) and Mariana Ruiz Villarreal (b).

Flagella are expressed on the surface of *Salmonella* cells. They consist of a basal body embedded in the cell membrane, a central rod attached to a hook which in turn attaches to a helical filament made up of polymerised units of flagellin protein, see Figure 1.2b (70, 71). Rotation of the basal body 'motor' results in movement of the filament which facilitates cell motility. Movement is regulated by sensory networks including chemotaxis proteins embedded in the cell surface, which recognise specific attractant and repellant molecules and signal via the CheA-CheW transmitter complex (70). Over 50 genes are required for flagella assembly (72). Most *Salmonella* have two distinct flagellin genes *fliC* and *fljB*, but express only one at a time, switching between them at a rate of 10^{-3}-10^{-5} (73, 74, 75). This process, known as phase variation, is present in four subspecies of *S. enterica* (including subspecies *enterica* serovars) and is absent from *S. bongori* (76). Phase variation may be a mechanism of avoiding cellular immunity, since FliC has been shown to be a target antigen for *Salmonella*-specific T-cells in a murine model (77). While the ends of the flagellin proteins are conserved, variation within the center of flagellin genes generates distinct flagellar antigens (76, 78). These are used in the Kauffmann-White serotyping scheme for *Salmonella*, which currently lists 70 H (flagellar) antigens (19). Sequence analysis of flagellin genes suggests that recombination between strains and between *fliC* and *fljB* within strains contributes to flagellin variation and the generation of new serovars (76, 79, 80).

A handful of *S. enterica* serovars, including Typhi but not Paratyphi A, express a Vi polysaccharide capsule (81, 82, 83). Vi is also expressed by some strains of *Citrobacter freundii* (84, 85) but has not been detected in any other species. Vi expression is regulated by two loci *viaA* and *viaB* which are separated on the chromosome (84, 86, 87, 88, 89, 90). *ViaA* is present in non-Vi strains, but *viaB* is specific to strains capable of expressing Vi (86, 91). The *viaB* locus includes genes for biosynthesis (*tviA*-*tviE*) and export (*vexA*-*vexE*) of the Vi antigen (90), and is part of the genomic island SPI7 as outlined below (46, 92, 93). The two-component regulatory system *ompR*-*envZ* is also involved in regulation of Vi (94). Vi expression is important for virulence in humans (95). Vi-expressing Typhi strains are more resistant to innate immune defenses (complement-mediated killing and phagocytosis) (96, 97) and Vi can inhibit inflammatory responses in human intestinal epithelial cell lines upon infection with Typhi (98).

Fimbriae are thread-like surface structures expressed in up to 500 copies per cell, which are involved in adhesion to non-phagocytic host cells. They are encoded in fimbrial operons of 3-10 genes (99) and are important for virulence in *Salmonella* (100, 101, 102). Each *S. enterica* serovar studied to date has a different set of fimbrial operons (103). The operons themselves are not unique to one serovar, rather each serovar encodes a distinct combination of fimbriae which contribute to its pathogenicity (101). For example, Typhi contains 13 fimbriae, eight of which are present in Typhimurium, although all are present in at least one other serovar (104). The majority of fimbriae in *Salmonella* are of the chaperone/usher family (99) (including 12 of the 13 fimbrial operons in Typhi). Their biosynthesis requires a periplasmic chaperone, which binds fimbrial subunits as they enter the periplasm (the space between inner and outer membranes) (105, 106) and an outer membrane usher which translocates the chaperone-bound subunits across the outer membrane (107, 108, 109, 110, 111, 112). The other types of fimbriae are type IV pili (113) (including one encoded in the Typhi genome (104, 114)) and nucleator-dependent curli fimbriae (115). Fimbriae contain adhesins that bind to receptors or sugars on host cells. Variation in fimbrial genes, including those encoding adhesins, determines the specificity and affinity of bacterial binding to host cells (116). This binding specificity has roles in bacterial colonisation of specific tissues and cell types and therefore host specificity, pathogenicity and niche adaptation (117, 118, 119, 120).

1.1.2.2 *Salmonella* Pathogenicity Islands

SPIs are clusters of virulence-associated genes that have been horizontally acquired by the *Salmonella* genome and can generally be identified by a base composition that differs from that of the rest of the chromosome (e.g. 42% GC content in SPI1 compared to 52% in the rest of the chromosome) (121). While they generally encode genes associated with virulence traits, the functions of many SPIs are not well understood. SPI1 and SPI2 were identified in 1995 and 1996 respectively and are present in all *S. enterica* (54, 55). They each encode a distinct type III secretion system (TTSS), a needle-like structure that enables bacterial proteins ("secreted effector proteins") to be secreted into the cytosol of host cells (122, 123, 124). In addition to the TTSS aparatus, the SPIs encode regulators and secreted effector proteins (125, 126), although effectors encoded elsewhere in the genome (including in prophage sequences) are also secreted via the SPI-encoded TTSSs (127, 128, 129). For a recent review of effectors and their functions, see (124). The TTSS encoded in SPI1 and SPI2 are genetically distinct, are expressed at different times and perform different functions (121).

SPI1 is 40 kbp in size and contains more than 25 genes including TTSS apparatus, regulators and effectors (54, 125). Distributed throughout *Salmonella* (130, 131) (see Figure 1.1), SPI1 is involved in colonisation of the gastrointestinal tract and can be induced *in vitro* by a shift in pH from acidic to mildly alkaline conditions, consistent with *in vivo* induction upon arrival in the mildly alkaline small intestine after passing through the acidic environment of the stomach (132). The expression of the TTSS is regulated via a complex circuit involving *hilA*, *hilC*, *hilD* and other genes to integrate environmental signals (133, 134). SPI1 is thought to be required for invasion of non-phagocytic cells of the intestinal epithelium (135), via a process that involves ruffling of the host cell membrane and rearrangements of the host cell actin cytoskeleton (136, 137, 138, 139). However a recent study identified *S. enterica* serovar Senftenberg isolates associated with human gastroenteritis that lacked SPI1, demonstrating that it is not essential for intestinal invasion in humans (140).

SPI2 contains two segments that were likely acquired consecutively - the first is 14.5 kbp, present in *S. bongori* as well as *S. enterica* and is not associated with systemic infection (126). The second is 25.3 kbp in size, is restricted to *S. enterica* (130, 131) (see Figure 1.1) and encodes a second TTSS apparatus, regulators, chaperones and secreted effectors (126). This part of SPI2 is required to maintain bacterial growth and replication inside host cells (141). It is associated with survival in macrophages, which facilitates systemic spread and colonisation of host organs (141). Expression of the SPI2 TTSS and effectors is regulated by a network of genes including global regulators *phoP-phoQ* and *ompR-envZ* as well as the SPI1-encoded *hilD*, allowing its induction in response to a variety of different environmental signals (142, 143, 144).

SPI3 contains 10 protein-coding sequences (CDS), and is involved in intramacrophage survival and virulence of Typhimurium in mice (145). It has a mosaic structure and different segments have distinct distributions among *S. bongori* and subspecies of *S. enterica* (145, 146). SPI4 also has a mosaic structure and encodes a type I secretion system (a protein channel (147)) that secretes an adhesin encoded by *siiE* (148, 149), which has been associated with invasion of the intestinal epithelium (149, 150, 151). SPI5 is associated with enteritis but not systemic infections (152). SPI4 and SPI5 are conserved within *S. enterica* subspecies *enterica* (146, 151, 152).

An additional 12 SPIs have been characterised in *S. enterica* (46, 153, 154, 155). SPIs 6-10 were first identified by analysis of the Typhi genome sequence (46). SPI6 and SPI10 encode fimbrial operons, while SPI9 encodes a type I secretion system (46). SPI8 encodes two bacteriocins (proteins toxic to other bacteria). These SPIs are much less conserved among *S. enterica* serovars than SPIs 1-5 (153, 154, 156). SPI7 is a 134 kbp region in the Typhi chromosome encoding genes for biosynthesis of Vi, the virulence-associated *sopE*-prophage and a type IV pilus operon (46, 92, 114). Part of SPI7, including the Vi biosynthesis genes, is also present in *S. enterica* serovar Paratyphi C, *Citrobacter freundii* and some *S. enterica* serovar Dublin strains, but to date has not been reported in any other *Salmonella* (82, 83, 85, 92, 93). SPIs 11-12 were first identified in the genome sequence of *S. enterica* serovar Choleraesuis, but are present in many other serovars (153). SPIs 13-14 were identified in *S. enterica* serovar Gallinarum in a screen for genes involved in infection of chickens and are present in many other serovars (154). SPIs 15-17 were identified in the Typhi chromosome via analysis of variation in base composition across the genome (155). SPI15 has so far only been reported in Typhi, but SPIs 16-17 are present in other serovars (155).

1.1.2.3 Horizontal gene transfer:

Horizontal gene transfer plays an important role in the evolution and adaptation of bacteria, including *Salmonella* (58, 157, 158). DNA can be transferred between bacterial cells via three mechanisms: conjugation, transduction and transformation, see Figure 1.3 (157). Although these mechanisms were once thought of as laboratory peculiarities (159), they have now been shown to occur in nature at high enough frequency to be a major force in bacterial evolution (157, 160, 161, 162, 163, 164).

Conjugation depends on the construction of a conjugative pilus, which is encoded by genes in conjugative plasmids or conjugative transposons (166, 167, 168). Plasmids

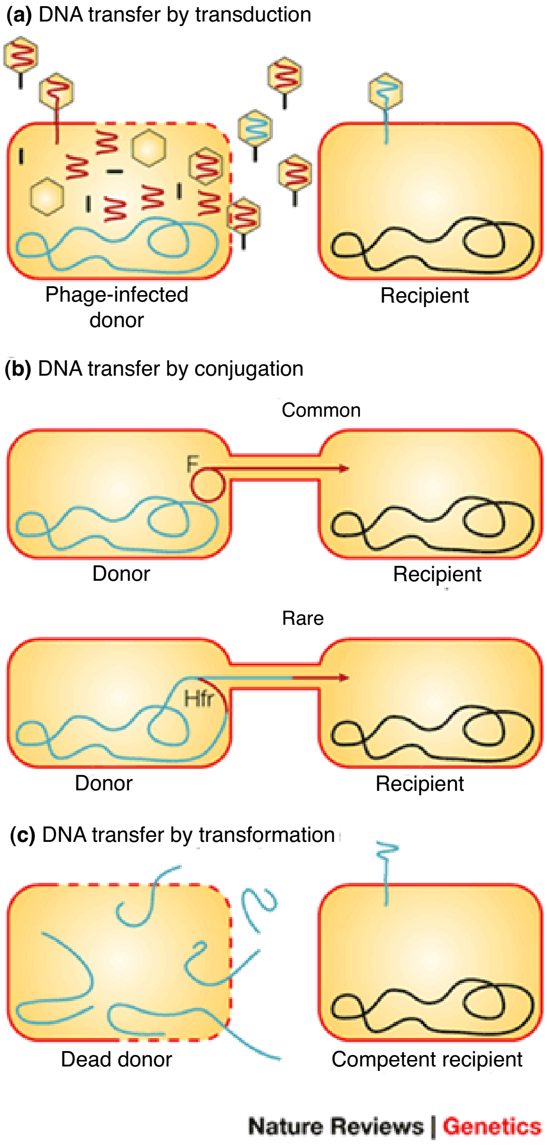

(a) DNA transfer by transduction
Phage-infected donor
Recipient
(b) DNA transfer by conjugation
Common
F
Donor
Recipient
Rare
Hfr
Donor
Recipient
(c) DNA transfer by transformation
Dead donor
Competent recipient
Nature Reviews | Genetics

Figure 1.3: Methods of DNA transfer - Reproduced from (165). Original legend: "(a) Transduction is the phage-mediated transfer of host genetic information. In a phage-infected bacterial cell, fragments of the host DNA are occasionally packaged into phage particles and can then be transferred to a recipient cell. (b) Conjugation is the transfer of DNA from a donor cell to a recipient that requires cell-to-cell contact. Genes on conjugative plasmids, such as the F plasmid, encode products that are necessary for this contact, and replication and transfer of the plasmid to the recipient. When, on rare occasions, the F plasmid becomes integrated into the host chromosome (Hfr), conjugation results in a partial transfer of the donor chromosome. (c) Cells that are competent can take up free DNA from their environment. For all three methods of DNA transfer, the donor chromosomal DNA will only be permanently maintained and expressed in the recipient cell if it is integrated into the recipient genome by physical recombination."

and transposons encoding their own conjugative machinery are referred to as "self-transmissible", while those that are simply transferred via the conjugative machinery of others are referred to as "mobilisable". Once inside the recipient cell, transposons are integrated into the host chromosome or resident plasmids; plasmids themselves can be integrated into the chromosome or remain as independent DNA molecules. Plasmids that use the same mode of replication and maintenance are said to be "incompatible": they are unable to transfer to or reside in the same host and are said to be of the same incompatibility (inc) group (169, 170, 171).

Very few conjugative transposons have been reported in *Salmonella*, although SPI7 (see above) encodes a type IV pilus and may be a conjugative transposon (92, 172). Conjugal transfer of SPI7 has not been demonstrated, however SPI7-mediated conjugal transfer of a small plasmid has been shown, and was dependent on the activity of SPI7-encoded transfer (*tra*) genes but not pilus genes (173). The *Salmonella* Genomic Island 1 (SGI1), first identified in multidrug resistant strains of serovar Typhimurium DT104 (174), includes genes involved in conjugal transfer (175). SGI1 also encodes resistance genes (174, 175), is associated with virulence in some animal models (176) and is mobilisable by conjugation (177). However current evidence suggests it is not self-transmissible, but requires the presence of a conjugative plasmid for transfer (177).

A variety of self-transmissible and mobilisable plasmids, ranging in size from 2-200 kbp and of different incompatibility groups, have been identified in *Salmonella* (178, 179, 180). The most well known are large plasmids encoding virulence or resistance genes, although small plasmids with different or unknown functions are also found. Many *S. enterica* serovars, including some of the most frequently isolated human and farm animal pathogens such as Enteritidis, Typhimurium, Dublin, Choleraesuis and Gallinarum, contain virulence plasmids of 50-100 kbp (179, 181, 182, 183). The plasmids are serovar-specific (179), and some are self-transmissible (184) while some rely on other plasmids for transfer (183, 185). The virulence plasmids encode the *spv* operon which is involved in intramacrophage survival and is required for full virulence in host organisms (185, 186, 187, 188). Plasmid-free isolates of these serovars are rarely found (189).

Resistance plasmids are also well known in *Salmonella* and other bacteria (190, 191, 192). These are usually large, self-transmissible plasmids carrying transposons and other mobile genetic elements that encode resistance to antibiotics (192, 193). Resistance to detergents and heavy metals are also found on plasmids (46, 194, 195, 196). Resistance plasmids in *Salmonlla* can be of different incompatibility groups (197, 198, 199), and are believed to have evolved from plasmids that were circulating in *Salmonella* prior to the use of antibiotics (180). *In vivo* transfer of MDR plasmids into Typhi has been documented (160). Recently, the acquisition of resistance genes by *S. enterica* virulence plasmids has been noted (198, 200, 201, 202, 203). Other small plasmids are found in ~10% of *S. enterica* isolates (178), although their functions are generally unknown (204, 205). Some are capable of phage conversion (altering an isolate's phage susceptibility profile or "phage type") and are used for strain typing in serovar Enteritidis (206, 207, 208, 209, 210, 211).

Transduction depends on bacterial viruses known as bacteriophage. Bacteriophage are transported between hosts in the form of virions or phage particles, made up of a protein coat (capsid) carrying phage DNA (161, 212). The phage DNA includes genes required for synthesis and assembly of the phage particles, but depends on the metabolic machinery of the bacterial host for reproduction. Bacteriophage have two lifestyle modes: productive, whereby new virus particles are produced and released from the cell, usually via cell lysis or bursting (lytic phage); and reductive, whereby the phage genome is not expressed, but becomes integrated into the host chromosome (temperate phage) (213). Generalised transduction occurs during the productive cycle, when bacterial DNA is packaged into the phage capsid by mistake (161). When the transducing phage (carrying bacterial DNA) infect a new bacterial cell, the bacterial DNA is released into the new host cell and may be integrated into the host chromosome or resident plasmids via homologous recombination. Temperate phage, also referred to as prophage, can be activated into the lytic cycle by environmental stresses (213). As they are excised from the chromosome, non-homologous recombination can occur between phage DNA and neighbouring bacterial DNA, leading to the packaging of some host genes ("cargo" genes) along with phage genes into the capsid in a process referred to as specialised transduction (161, 214). Non-phage genes may also be integrated into the phage sequence by transposition (transposase-mediated integration) (215).

Phage cargo genes can contribute to virulence of bacterial pathogens (216), the best known examples being genes encoding toxins including diphtheria toxin (217), Shiga toxin (215, 218) and cholera toxin (219). Phage transduction can also contribute to the spread of antibiotic resistance (220). In *Salmonella*, some effectors secreted by the SPI-encoded type III secretion systems (see above) are phage cargo genes (128). A well-characterised example is *sopE*, which is carried by a phage that can infect serovar Typhimurium isolates *in vitro* (221, 222) and is present in the genomes of Typhi, Paratyphi A, systemic pathovars of Paratyphi B and epidemic strains of Typhimurium and other serovars (46, 47, 49, 50, 221, 222, 223, 224, 225). Two other prophage identified in the Typhimurium genome have been associated with the ability of the serovar to cause systemic infection in mice (226).

Prophage content varies extensively between and within serovars (46, 47, 49, 50, 93, 153, 224, 227). One reason for this is the specificity of phage, which bind specific molecules on the bacterial cell surface (228, 229, 230) and integrate into specific sites in the chromosome, often tRNA sequences (231, 232, 233, 234). For example phage that bind to Vi are only able to infect Typhi and other serovars expressing Vi (235). There are also more complex systems of phage immunity, as the expression of resident prophages is repressed by specific repressor proteins, which also repress expression of incoming phages of a similar type (236, 237). Because of this variation, phage can be used to discriminate among isolates of a given serovar, either by PCR targetting known integration sites (238, 239) or phage typing which involves infecting isolates with a panel of bacteriophage to determine their profile of phage susceptibility (6, 240, 241).

Horizontal DNA transfer can also occur via transformation, involving the active uptake of double-stranded DNA by transformation-competent bacterial cells followed by integration into the genome by homologous recombination. Homologous recombination between *S. enterica* serovars has been documented (23, 56, 242, 243, 244, 245) and probably occurs via a combination of transduction, conjugation and transformation.

1.1.3 Serovar Typhi

S. enterica subspecies *enterica* serovar Typhi (referred to as Typhi hereafter) is the causative agent of typhoid fever in humans. Typhi was first cultured in 1884 and before the advent of modern *Salmonella* nomenclature (18) has been known as *Bacillus typhosus*, *Erbethella typhosa*, *Salmonella typhosa* and *Salmonella typhi* (246). Typhi is the only *S. enterica* serovar known to characteristically express high levels of Vi antigen (expression is low in serovar Paratyphi C (83) and observed in a single clonal lineage of serovar Dublin (82)). As outlined above (1.1.2.2), Vi expression is encoded in the *viaB* locus of SPI7, which is unique to these serovars. Typhi is generally monophasic, harbouring the *fliC* gene but not *fljB*, and expresses the H:d antigen. Thus identification in the laboratory is confirmed by serotyping as O9,12:Hd and Vi antigen (19) (although occasional Typhi isolates may be Vi-negative (247)). Typhi isolates from Indonesia sometimes express unique flagella types H:j and H:z66 (248, 249, 250, 251, 252, 253), which led to the formulation of a hypothesis that Typhi evolved in Indonesia as a biphasic organism before a monomorphic variant arose and became globally disseminated (252). However the discovery that the H:z66 antigen is encoded by a *fljB* gene (254) located on a unique 27 kbp linear plasmid restricted to a specific (and non-ancestral) clone (255, 256) quashed the idea of a biphasic Indonesian ancestor of Typhi. The H:j antigen results from a 261 bp deletion within the chromosomally-encoded *fliC* gene, mediated by homologous recombination between 11 bp repeats within the central part of the gene (253). However the deletion appears only to occur in strains carrying the z66-encoding linear plasmid (252, 253, 256).

The Typhi CT18 and Typhimurium LT2 genomes were the first *Salmonella* genomes to be sequenced (46, 50). The genomes were published in 2001, followed in 2003 by the Typhi Ty2 genome sequence (47). The Typhi and Typhimurium chromosomes differed at <15% of gene loci and showed <2% divergence at the nucleotide level (46, 50). Most of the differences in gene content were due to prophage sequences (seven in Typhi, see Figure 1.4) and the presence of SPI7 (including the *sopE* phage) in Typhi, although several smaller insertions and deletions were identified (46, 50). The Typhi CT18 and Ty2 sequences were less than 0.01% divergent at the nucleotide level and shared >99% of their genes (47). The differences were in prophage (see Figure 1.4b) (224), variants

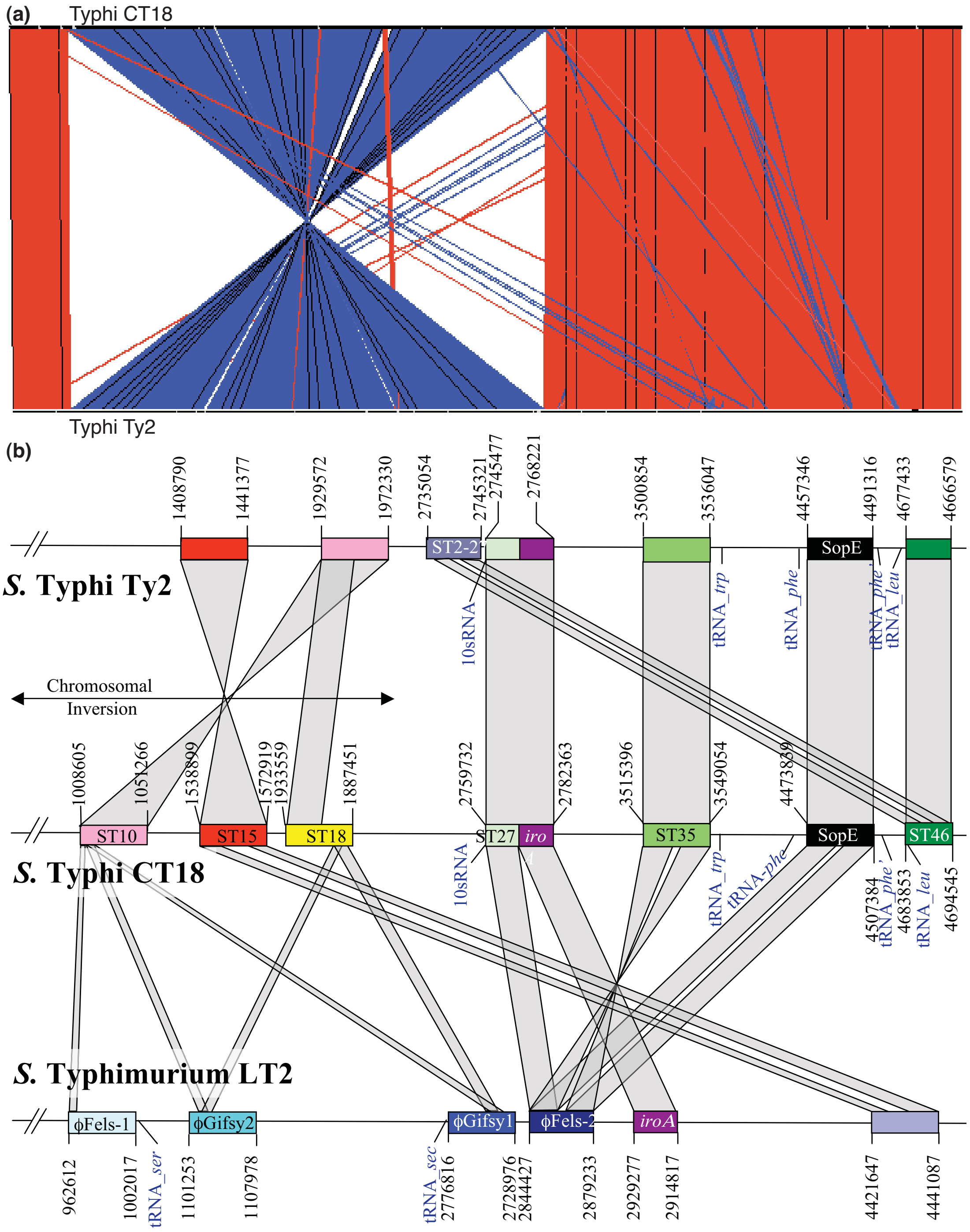


Figure 1.4: Genome rearrangements and phage differences between Typhi CT18 and Ty2 - (a) Linear comparison of Typhi CT18 and Ty2. (b) Prophage in Typhi and Typhimurium genome sequences. Reproduced from (224), original legend: "Illustration of the relative alignments of the prophage regions within the chromosomes of Typhi Ty2, Typhi CT18 and Typhimurium LT2 genomes. Regions displaying significant sequence homology are linked by the grey shading. The co-ordinates of the prophage regions are indicated and similar phage are coloured accordingly. The positions of relevant stable RNA genes are shown."

of SPI15 (155), a deletion in SPI7 in Ty2 and the insertion of *IS*1 elements in CT18 (47). The Typhi CT18 and Ty2 genomes were generally collinear, with the exception of a large inversion between two rRNA operons, see Figure 1.4a (47). The *S. enterica* genomes contain seven near-identical rRNA operons, and rearrangements between them have been found to occur frequently in isolates of Typhi (257, 258, 259), but not other serovars (259, 260, 261, 262). The Typhi genome contains over 200 pseudogenes (46, 47), protein-coding sequences that have been inactivated by nonsense mutations, deletions or frameshifts, preventing the proper expression of the encoded protein. Pseudogenes appear to be more frequent in host-restricted pathogenic bacteria compared to their host-generalist relatives (46, 49, 263, 264, 265, 266). The Typhimurium genome contains only 39 pseudogenes (50), similar to *E. coli* K-12 (74) (267) and other host-generalist bacteria.

Plasmids are occasionally found in Typhi isolates. The Typhi CT18 genome sequence includes two plasmids, a 218 kbp IncHI1 multidrug resistance (MDR) plasmid pHCM1 and a 107 kbp cryptic plasmid pHCM2 (46). MDR plasmids appeared in Typhi in 1972 (268) and have persisted in many regions ever since (16). They are most often of the IncHI1 type (268, 269, 270, 271, 272, 273, 274, 275, 276), although other types have been identified (277). The pHCM2 plasmid shows similarity to the *Yersinia pestis* virulence plasmid pMT1 (46) and is rare in Typhi (278). Other small plasmids not associated with drug resistance are found in Typhi isolates with varying frequency, but with the exception of the z66-encoding linear plasmid, their functions are unknown and they have not been sequenced. A survey of plasmids from Typhi isolated during the pre-antibiotic era found diversity in incompatibility types and sizes (180).

1.1.4 Serovar Paratyphi A

S. enterica subspecies *enterica* serovar Paratyphi A (referred to as Paratyphi A hereafter) is the causative agent of paratyphoid fever in humans, which is generally indistinguishable from typhoid fever (279, 280). Paratyphi A was first identified in 1902 and was briefly known as *Bacillus paratyphi* typus A (281). Unlike Typhi, Paratyphi A does not express Vi and does not contain SPI7 or the *viaB* locus (49). Paratyphi A is monophasic for phase 1 flagella (encoded by *fliC*), due to a frameshift in the *hin* gene which is required for phase switching, although the *fljB* gene is intact (49). Identification in the laboratory is confirmed by serotyping as O1,2,12;Ha (19).

The first Paratyphi A genome was sequenced and published in 2004 (49). The genome was 4.5 Mbp, smaller than Typhi mainly due to the lack of SPI7 and presence of only three prophage. These include the *sopE*-phage, which shared 95-100% amino acid identity with that encoded in the Typhi genome (49). The Typhi and Paratyphi A sequences shared 172 genes that were not present in the sequenced Typhimurium or *E. coli* K-12 genomes, far more than the number of genes shared uniquely by any other pair of these genomes (see Table 1.3) (49). A detailed analysis of nucleotide divergence between Paratyphi A, Typhi, Typhimurium and other serovars was published in 2007 (56). The study showed that pairwise divergence between most serovars was ~1%, whereas one quarter of the Typhi and Paratyphi A genomes were less than 0.2% divergent. The authors concluded that Typhi and Paratyphi A have exchanged large amounts of genomic DNA relatively recently, including many of the 'rare' genes referred to in Table 1.3 (56). Gene order was generally conserved between Paratyphi A and Typhimurium, except for an inversion (between ribosomal operons) of half the chromosome (49). This inversion had been noted previously and was conserved among 12 strains tested (260). The Paratyphi A genome contained 173 annotated pseudogenes, similar to the number in Typhi but largely involving independent mutations and affecting different genes (49).

	Typhi	**Typhimurium**	***E. coli* K12**
Paratyphi A	172	53	0
Typhi		60	15
Typhimurium			48

Table 1.3: Genes unique to pairs of *Salmonella* and *E. coli* genomes - Reproduced from (49). Original legend: "Number of genes shared by a pair of genomes but not the other two genomes, comparing Paratyphi A ATCC9150, Typhi CT18 , Typhimurium LT2 and *E. coli* K-12. Shared genes: >95% identity in a 100-bp window, except for *E. coli* comparison (>75% in a 100-bp window)."

The published Paratyphi A genome was plasmid-free (49). However although Paratyphi A has received much less research attention than Typhi, occasional studies have reported the presence of plasmids in Paratyphi A isolates. MDR Paratyphi A was first reported from India in 1977 (282). A recent study of MDR Paratyphi A isolated from Pakistan between 2002-2004 demonstrated that MDR was associated with IncHI1 plasmids of approximately 220 kbp (283). Prior to this, a large transferable plasmid of 140 MDa (∼230 kbp) was found in 73% of MDR Paratyphi A strains in Bangladesh from 1992-1993 (284) and a plasmid of similar size was reported in China in 2004 (285). In India, a 55 kbp transferable plasmid was associated with MDR Paratyphi A from 1991-2001 (286). A small cryptic plasmid, pGY1, was sequenced from a paratyphoid patient in China in 2005 (287). The plasmid is 3,592 bp in size and contains three CDS, none of which have any similarity to known resistance genes. A putative replication origin was identified by its similarity to those of previously characterised plasmids (287). Small plasmids of 2.2, 5 and 20 kbp were reported among Paratyphi A strains from Kuwait in 1995-1999 (288), while plasmids of 2.2, 3.6, 9.5 and 20 kbp have been reported in Paratyphi A isolates from China (287).

1.2 The disease: enteric fever

The diseases caused by Typhi and Paratyphi A are generally called "typhoid fever" and "paratyphoid fever" respectively, with the term "paratyphoid" also used to describe sytemic infection with Paratyphi B or C. The collective term for systemic disease caused by Typhi or Paratyphi A, B or C is "enteric fever", but sometimes "typhoid fever" is used collectively in this manner as well. Transmission is by the fecal-oral route, where infected individuals excrete bacteria in their feces and urine, which in unsanitary environments can contaminate food or water ingested by other individuals. The infectious dose of Typhi is in the range of 10^3 - 10^9 ingested organisms (95), and carriers can excrete up to 10^9 in a single gram of feces (246).

1.2.1 Pathology and clinical features

Following ingestion of an infectious dose of Typhi or Paratyphi A, the bacterial cells pass through the acidic environment of the stomach to reach the lower small intestine (the ileum), where they begin to invade the intestinal epithelium (289), see Figure 1.5. The initial targets of invasion are likely the M cells (specialised epithelial cells), which transport the bacteria to the underlying lymphoid tissue. Here they invade intestinal lymphoid follicles and the draining mesenteric lymph nodes, allowing some bacteria to spread to the liver and spleen, where they can survive and multiply within mononuclear phagocytic cells (290). This incubation period lasts 7-14 days, after which bacteria are released into the bloodstream (bacteraemia). It is usually at this point that patients experience the onset of fever and other symptoms, but tend not to seek medical treatment for several days (246). If untreated, the bacteria can become widely disseminated during the bacteraemic phase, spreading to the liver, spleen, bone marrow and gall bladder (289, 290). In acute typhoid fever patients, the median concentration of bacteria is 1 colony-forming unit per mL of blood (two-thirds of which are inside phagocytes) (291) and roughly ten times this in bone marrow (292). Bacteria are often excreted in the urine or feces of enteric fever patients (bacterial "shedding"), either via intestinal lesions or following colonisation of the gall bladder (293, 294, 295).

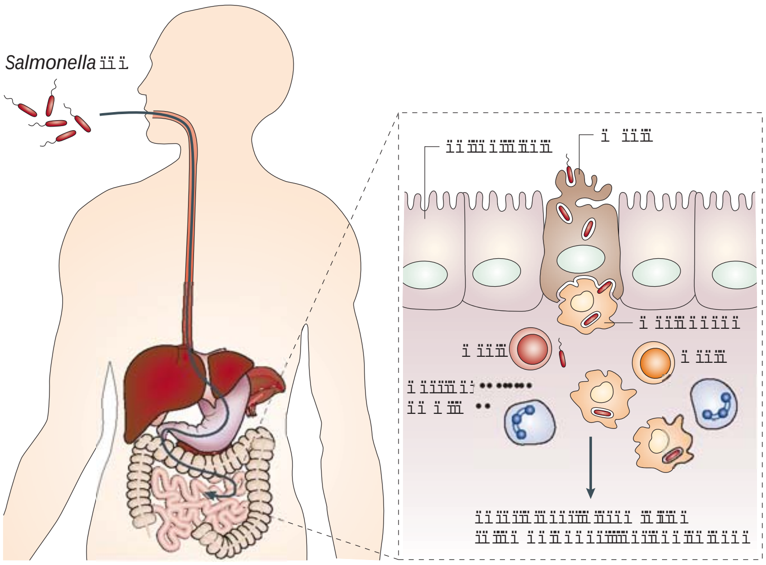


Figure 1.5: Biology of *Salmonella* infection - Reproduced from (296). Original legend: "Orally ingested salmonellae survive at the low pH of the stomach and evade the multiple defences of the small intestine in order to gain access to the epithelium. Salmonellae preferentially enter M cells, which transport them to the lymphoid cells (T and B) in the underlying Peyer's patches. Once across the epithelium, *Salmonella* serotypes that are associated with systemic illness enter intestinal macrophages and disseminate throughout the reticuloendothelial system. By contrast, non-typhoidal *Salmonella* strains induce an early local inflammatory response, which results in the infiltration of PMNs (polymorphonuclear leukocytes) into the intestinal lumen and diarrhoea."

During the bacteraemic phase of infection, patients normally present with persistent fever of up to 40°C, although other symptoms vary widely among patients (3, 297). Malaise, flu-like symptoms and a dull frontal headache are most frequent, although rapid weight loss, poorly localised abdominal discomfort, dry cough and myalgia (muscle pain) are also common. Hepatomegaly or splenomegaly (enlargement of the liver or spleen, respectively) are sometimes found. Other physical signs include coated tongue, tender abdomen and rose spots, which occur in 5-30% of cases (3). A number of complications have been described in patients with typhoid fever. The most common are gastrointestinal bleeding (up to 10% of patients) (3) and intestinal perforation (3% of patients) (298), although extraintestinal complications can also occur. These include encephalopathy (affecting the brain), heart disease, pneumonia (lung infection), osteomyelitis (bone infection) and abcesses of the liver, spleen, kidneys and other organs (299).

Host genetic factors play a role in enteric fever susceptibility and possibly in disease severity. Mutations in toll-like receptor 4 (TLR4), tumour necrosis factor alpha (TNFa) and other MHC class II and III genes have been associated with susceptibility to typhoid fever in Vietnam (300, 301, 302). In contrast, TLR5 and NRAMP1 (natural resistance associated macrophage protein 1) were not associated with typhoid fever in similar Vietnamese populations (303, 304). In Indonesian populations, TNFa was not associated with typhoid or paratyphoid susceptibility but may be associated with disease severity (305). However PARK2 (E3 ubiquitin ligase parkin 2) was associated with susceptibility to typhoid and paratyphoid fever in Indonesian populations (306).

1.2.2 Asymptomatic carriage

Colonisation of the gall bladder by Typhi or Paratyphi A can result in long-term fecal shedding of bacteria. As no other reservoir has been discovered for Typhi or Paratyphi A, this is considered to be the central mechanism by which the disease is transmitted. In untreated patients, up to 10% will shed bacteria for up to 3 months (temporary carriage) (307). Up to 4% of typhoid or paratyphoid fever patients remain chronic carriers for more than a year after the resolution of symptoms, and carriage can persist for much longer (3, 307, 308, 309, 310). This is more likely to occur in patients with underlying pathology of the gall bladder (309), more often affects women than men (309, 311) and is associated with an increased risk of cancer of the gall bladder, pancreas and large bowel (312, 313, 314). Gall bladder carriage of Paratyphi A has been documented (311, 313), but has not been as well studied as Typhi carriage. Asymptomatic gall bladder carriage of Typhi or Paratyphi A can occur in the absence of enteric fever symptoms, with up to 25% of carriers having no history of enteric fever (246, 311). Urinary shedding also occurs, most commonly in patients with urinary tract pathology, and is thought to be associated with urinary schistosomiasis (parasite) infection (293). Gall bladder carriage is most frequently discovered through gall bladder surgery, although detection of Typhi or Vi in blood or stool can be used to identify Typhi carriers (315, 316, 317).

1.2.3 Diagnostics

Given the non-specific nature of the signs and symptoms of uncomplicated enteric fever (318), accurate diagnosis requires culturing of the organism, most commonly from blood (60-80% sensitive) (294). Culturing from bone marrow is more sensitive (up to 95%) (294, 295, 319), but is invasive and rarely performed in resource-poor settings where enteric fever is endemic. Upon culturing, the organism can be identified by biochemical tests and serotyping (19, 297). An alternative diagnostic method is by a serological test - the demonstration of O and H antibodies in patient serum - known as the Widal test (320). However interpretation of the test is not straightforward as the presence of antibodies may be the result of prior infection with Typhi/Paratyphi A or other serotypes (321), vaccination against Typhi or even cross-reactivity with other *Enterobacteriaceae* (297, 320). Furthermore, up to one third of patients do not mount a detectable antibody response or show no detectable rise in antibody titre (246). Rapid PCR-based diagnostic tests have been proposed by a number of researchers. The technique involves amplification of Typhi-specific or Paratyphi A-specific sequences directly from blood, feces or urine. This has the advantage of rapidity, bypassing the need to culture and serotype the organism directly which can take several days, as well as increased sensitivity over blood culture (322, 323, 324). Target sequences proposed include serotype-specific alleles of *wba* cluster genes and/or flagellin (324, 325, 326), and serotype-specific genes such as the *viaB* locus (322, 324). However these and other recently proposed tests vary in sensitivity and are not practical in resource-poor endemic settings, and so a rapid and inexpensive diagnostic test for enteric fever remains an elusive target (318).

1.2.4 Epidemiology

Worldwide, the annual rate of enteric fever cases is approximately 20 million and results in over 200,000 deaths (10). The majority of the burden is in endemic areas in developing countries of Asia, Africa and central and South America where sanitation is poor (10). The highest rates are observed in Southern Asia, in particular India, Pakistan, Vietnam and Indonesia (20-450/100,000 people annually) (10, 327, 327, 328). The 20th century saw a decline in enteric fever in developed countries, for example in the US the annual incidence declined from 7.5/100,000 people in 1940 to 0.2/100,000 people in 1990 (329); in the UK annual case numbers declined from 2,500 in 1936 to less than 500 in 1990-2008 (9, 330). Current incidence rates in developed countries are in the range of 0.1-1/100,000 (12, 331, 332, 333) and the majority of both typhoid and paratyphoid fever cases (>80%) in developed countries are associated with travel to endemic areas (11, 12, 329, 332), in particular India and neighbouring countries (11, 12, 332, 334).

Historically, the vast majority of enteric fever cases have been caused by Typhi (10), however the relative importance of Paratyphi A has been rising over the last 20 years. For example, the proportion of enteric fever cases in the UK caused by Paratyphi A increased from less than 30% in 1990 to 50% in 2001, and has stayed at that level (up to current data from 2008, see Figure 1.6) (330). This may be associated with travellers' use of vaccines against Typhi, which provide little cross-protection against infection with Paratyphi A (see below 1.2.6). Among endemic areas, the situation is most dramatic in China, where Paratyphi A is more prevalent than Typhi (64% in 2001-2002) (335). In Nepal, too, one study found that Paratyphi A increased from 15% of enteric fever cases in 1993 to 35% in 2002 (336), while other studies have reported rates as high as 50% Paratyphi A among both tourists (337) and local residents (338) with enteric fever. However up until 2002 Paratyphi A was still relatively infrequent in some endemic countries, including Pakistan (15%), India (24%) and Indonesia (14%) (335), although rising incidence is beginning to be reported in India (339, 340, 341).

The age distribution of enteric fever patients differs markedly in different populations. In endemic areas with high incidence of typhoid fever the mean age of patients is low, affecting mainly school children, whereas in areas with lower incidence the mean age

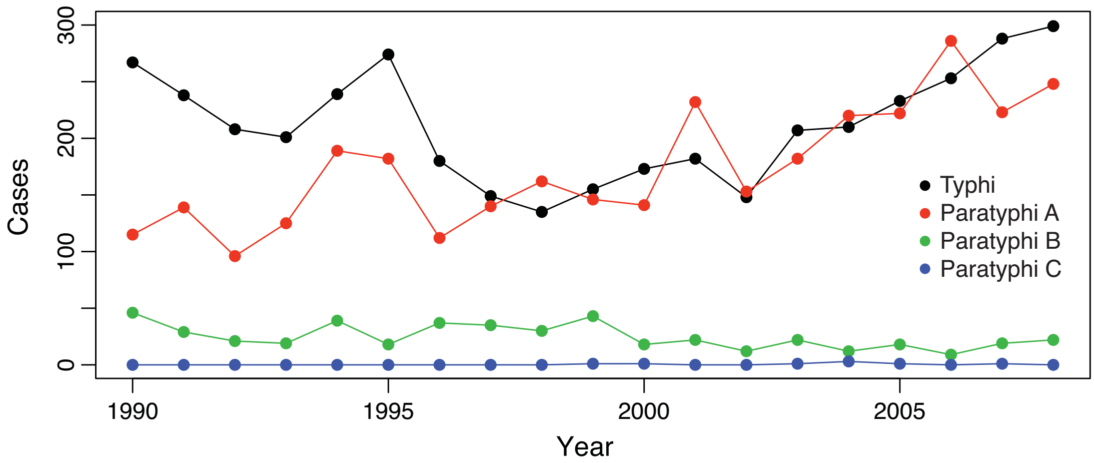


Figure 1.6: Trends in enteric fever incidence in the UK, 1990-2008 - Annual enteric fever cases per year for England, Wales and Northern Ireland, split by serotype. Sourced from publicly available data published online by the Health Protection Agency, London, UK (330).

of patients is higher, affecting mainly young adults (10, 327). In areas of very low incidence, typhoid is more evenly distributed among children and adults under the age of 40 (10). This inverse relationship between incidence rate and median age of patients is considered to reflect acquired immunity among residents of high incidence endemic areas (329). Among travellers to endemic areas, age does not appear to be a factor in acquiring enteric fever (12), consistent with the notion that any immunity in this group is likely to be due to vaccination and not dependent on age (329). In endemic areas, the age distribution of paratyphoid patients is generally higher than that of typhoid patients (279, 342, 343), which may also be related to the lower incidence of Paratyphi A.

Transmission of enteric fever is through fecal- or urine-contaminated food or water. General risk factors among residents of endemic countries include low levels of income and education (343, 344, 345) and large, crowded households (343, 345, 346). Risk factors relating to water sources and hygiene include lack of clean, piped drinking water (342, 344, 347, 348) (in particular drinking unboiled water sourced from rivers or streams (349, 350)), keeping water in open-mouthed containers (344), lack of toilet facilities in the home (342, 344, 349) and lack of regular handwashing (347), particularly without soap (342). Food-related risk factors among residents of endemic countries include consumption of food from street vendors (342, 347, 351), in particular ice-related products (342, 347, 351, 352), and consumption of unwashed fruit or vegetables (344, 346, 350). Contact with typhoid patients is also a risk factor in endemic areas (342, 345, 353). Climatic factors are also recognised, with enteric fever incidence associated with warm weather, increased rainfall and flooding (338, 342, 344, 349, 354). Among travellers to endemic areas, lack of vaccination, poor local sanitation and not following food and water precautions are associated with enteric fever (329). Travellers who stay longer (355), or visit friends and relatives (356) are more likely to become ill. In developed countries, enteric fever cases not associated with travel usually occur in localised outbreaks that can be traced to a single water source (357), food source (358) or carrier (359). Enteric fever rates decline with increasing quality of public water supplies (360), which can be dramatically improved via filtration and to some extent by chlorination (361).

1.2.5 Antibiotic treatment and resistance

A timeline of antibiotic use and resistance in enteric fever is shown in Figure 1.7. The antibiotic chloramphenicol was introduced for the treatment of enteric fever and other bacterial diseases in 1948 (362). Chloramphenicol-resistant typhoid fever was reported two years later (363), but was not common until the early 1970s when a number of chloramphenicol-resistant typhoid outbreaks swept through central and South America and Asia (364, 365, 366, 367). Drug resistance was encoded by the chloramphenicol acetyltransferase gene, *cat*, which catalyses the acetylation of chloramphenicol, leaving the drug unable to bind to and block the the activity of bacterial ribosomes (368, 369). The gene was carried on a plasmid of the IncHI1 type (see above 1.1.2.3) which also carried genes encoding resistance to sulfonamides, tetracyclines and streptomycin antibiotics, but not to ampicillin or co-trimoxazole (364, 365). Ampicillin was first used to treat chloramphenicol-resistant typhoid in 1962 (370, 371) and co-trimoxazole (a combination of the drugs trimethoprim and sulfamethoxazole) was first used in 1968 (372). During a 1972 chloramphenicol-resistant outbreak in Mexico, Typhi isolates resistant to chloramphenicol, sulfonamides, tetracyclines, streptomycin and ampicillin were identified (364). In 1977, a few strains of Typhi and Paratyphi A were found to be resistant to these drugs as well as co-trimoxazole (282, 373). This was the first example of multidrug resistant (MDR) enteric fever, defined as resistance to chloramphenicol, ampicillin and co-trimoxazole. In 1981 plasmid-mediated MDR was confirmed in a typhoid patient, acquired during the course of treatment with chloramphenicol (160). By 1987 MDR had spread to China and Pakistan (374, 375, 376). Outbreaks of MDR typhoid were reported in India in 1990 (377, 378), shortly followed by Malaysia (379), Vietnam (380), Bangladesh (381) and elsewhere (190, 271, 382, 383). Although chloramphenicol-resistant Paratyphi A was first reported in 1977 (282, 373), paratyphoid has predominantly been susceptible to antibiotics (384, 385). However, in recent years the incidence of MDR Paratyphi A isolates has increased, particularly in Pakistan and India where MDR rates as high as 45% of Paratyphi A isolates have been reported (286, 386, 387, 388).

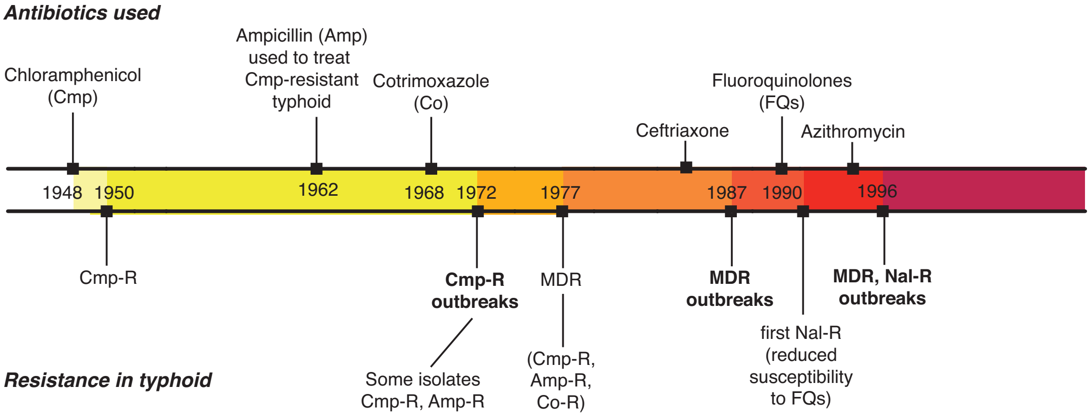


Figure 1.7: Timeline of the use of, and development of resistance to, antibiotics in enteric fever - R = resistant; MDR = multiple drug resistance; Nal = nalidixic acid, the prototype quinolone antibiotic. Details and citations for all events are given in the text.

Fortunately by 1990 a new class of drugs, the fluoroquinolones, were available for the treatment of typhoid fever (389, 390). Fluoroquinolones such as ciprofloxacin and ofloxacin were effective in the treatment of MDR typhoid fever, paratyphoid fever and in asymptomatic carriers (390), were affordable and were effective over short courses of 5-7 days (3). They were widely adopted in areas where MDR enteric fever is common and remain the recommended treatment for uncomplicated enteric fever, including MDR cases (297). However reduced susceptibility to the fluoroquinolones emerged almost immediately in both Typhi and Paratyphi A (16, 391, 392), resulting in increased fever clearance times and sometimes treatment failure in affected patients (393, 394). Fluoroquinolones work by inhibiting the action of topoisomerases such as DNA gyrase, which is necessary for the unwinding of DNA during bacterial replication (395). Decreased susceptibility to the fluoroquinolones is conferred by point mutations in the topoisomerase genes of the bacteria, in particular in codons 83 and 87 of the *gyrA* gene in *Salmonella* (391, 393). Mutations in the topoisomerase genes *gyrB*, *parC* and *parE* can also reduce susceptibility to fluoroquinolones (396, 397, 398). Reduced susceptibility is most often determined by detection of resistance to nalidixic acid (Nal), the prototype quinolone (399). The rate of Nal resistance in Typhi and Paratyphi A increased rapidly in the late 1990s, reaching 97% among southern Vietnamese Typhi isolates, 50% among Typhi isolates on the Indian subcontinent (16, 327, 400) and 80% among Paratyphi A isolates in Nepal (279, 338) by 2004.

The majority of Nal resistant isolates are still susceptible to fluoroquinolones, although their susceptibility is reduced and MICs (minimum inhibitory concentrations) are increased up to 10-fold (393). Full resistance to fluoroquinolones such as ciprofloxacin is reported sporadically for both Typhi and Paratyphi A (401). For the treatment of enteric fever with reduced susceptibility to fluoroquinolones, ceftriaxone (an extended spectrum cephalosporin, first used in 1985 (402)) or azithromycin (a macrolide antibiotic, first used in 1994 (403)) are recommended (297, 404). Resistance to ceftriaxone is occasionally reported in Typhi and Paratyphi A (405, 406). Plasmid-associated fluoroquinolone resistance has been reported in other *S. enterica* serovars but so far not in Typhi or Paratyphi A (407, 408, 409, 410, 411, 412, 413, 414, 415). A decline in MDR typhoid fever has been observed in many regions in the last 15 years (16, 327, 400, 416, 417, 418), presumably associated with the switch to fluoroquinolones and resulting reduction in selective pressure for resistance to the older antibiotics.

1.2.6 Prevention

As the risk factors outlined above highlight, the key to prevention of enteric fever is clean water and good hygiene practices. However, this is generally considered to be a long term goal in developing countries for political and economic reasons, so in the short to medium term vaccination is likely to be the most effective method of prevention. Vaccines against Typhi have been in use since 1896, when a killed whole-cell typhoid vaccine was developed in Britain for use in soldiers fighting in the Boer war in Africa (13). While the vaccine was in widespread use among the British and American military for much of the 20th century (13), its efficacy (73% after three years) was not established until controlled trials in the 1960s, which also demonstrated a high rate of side effects (419). Because of this, whole-killed typhoid vaccines are no longer used (14). Two typhoid vaccines are currently licensed for commercial use: Ty21a (a live attenuated Typhi strain given orally) and Vi (purified Vi antigen given as an intramuscular injection). Ty21a is licensed for use in adults and children over six years of age, has no significant side effects and provides approximately 50% protection over three years (14, 420). The Vi vaccine is licensed for use in adults and children over two years of age, has no significant side effects and provides greater than 60% protection over two years but requires a booster to maintain protection beyond this period (14, 420). A new Vi conjugate vaccine has been trialled in South East Asia, demonstrating protection of 80-90% over 2-4 years in children aged 2-5 years (14, 421, 422). Fever was more frequent among vaccinees than those given placebo (1.3% of vaccinees) (14, 422), however this was not a serious complication and the longer lasting protection and efficacy in young children makes this vaccine a promising prospect for the control of typhoid fever in high incidence endemic areas.

Vaccination against typhoid is currently recommended for travellers to areas where typhoid is endemic (15, 423), as well as for household contacts of typhoid carriers and laboratory workers who handle Typhi (420), although efficacy in these groups has not been demonstrated. Conversely, vaccines are not routinely used in countries where typhoid is endemic and efficacy has been demonstrated. In 1987, mass immunisation of school children in Thailand was highly effective in reducing the incidence of typhoid fever (424), but such programmes have not been adopted in neighbouring countries, which maintain the highest incidence of typhoid in the world (10, 15). There is widespread support for such programmes in the international medical and research community (425), although the cost effectiveness of a typhoid vaccine is a contentious issue and must be weighed against other health concerns in each country (426, 427, 428, 429, 430). There is currently no vaccine available for Paratyphi A, B or

C. The Ty21a vaccine reportedly provides some cross-protection against infection with Paratyphi A and B, which share the O12 antigen with Typhi (280, 431). However the mass immunisation of Thai school children with killed whole-cell typhoid vaccine did not demonstrate any protection against Paratyphi A (424). Not having the Vi antigen, Paratyphi A and Paratyphi B are unaffected by the Vi vaccine (432, 433), although it may provide cross-protection against Paratyphi C which expresses Vi.

1.3 The approach: comparative and population genomics

1.3.1 Population genetics of bacterial pathogens

Bacteria exist in communities or populations of organisms. The genetic structure and dynamics of bacterial populations is shaped by the range of selective pressures acting on individuals in the population, and can differ markedly between bacteria (recently reviewed in (434, 435)). In the case of bacterial and other pathogen populations, selective pressures on the population include interactions with the host, e.g. host immunity (natural or vaccine-associated), natural variation in host genetics, treatment with drugs and disease screening, as well as environmental and ecological factors. By examining the genetic structure of a bacterial pathogen population, insights may be gained into the evolutionary history of the organism, including evidence of selective pressures that can reveal important clues as to its lifestyle and interactions with the host. Furthermore the dynamics of bacterial populations can reveal insights into the evolution of clinically important phenotypes such as virulence, drug resistance and antigenic variation which may be used to decide upon the most appropriate medical or public health interventions. Finally, understanding the population structure of a pathogen is important in order to design molecular epidemiological studies of infectious disease, including determining the most appropriate sampling and molecular methods.

1.3.1.1 Evolution and variation in pathogen populations

Pathogenic lifestyles range broadly from “obligate pathogens” that have no environmental reservoir outside the host and depend on infection and disease for survival and spread, to “opportunistic pathogens” that can spread without causing disease but may spread more quickly by causing host pathology, and “accidental pathogens” for whom causing disease does not promote spread at all (436). Different lifestyles will be subject to (and result from) different selective pressures, which favour the spread of some members (genetic variants) of the population over others. Other factors like growth or contraction of the population, or physical isolation of subpopulations, also contribute to population structure (436). The level and nature of diversity within the population will be influenced by the lifestyle of the pathogen, and will influence the population structure. For example, the simplest model of bacterial evolution is a clonal one, whereby novel mutations (substitution, deletion or insertion of one or a few bases) are passed on to daughter cells during cell division, and new lineages emerge by accumulation of these mutations over generations. In this case of purely asexual reproduction, mutations are in strong linkage disequilibrium, and positive selection for one beneficial mutation arising in a given lineage can result in fixation of all the mutations in the lineage. In the presence of free recombination, mutations are constantly reassorted, resulting in linkage equilibrium and an entirely nonclonal population structure (437). Most bacteria will lie somewhere in between the two extremes of entirely asexual reproduction and free recombination, see for example (438, 439). By examining the diversity of a pathogen population one can infer the degree to which mutation and recombination have contributed to its evolution (437, 440, 441, 442, 443, 444). This is important to guide the design and interpretation of epidemiological studies, as it directly affects the assumptions that can sensibly be made based on the analysis of genetic variation. For example in *S. enterica* subspecies *enterica* most serovars correspond to an essentially clonal group of organisms (445, 446, 447) despite evidence of some recombination within the subspecies (245, 448), making serotyping a useful tool for comparing the causative agents of salmonellosis in human populations around the world (24, 25, 34). In contrast a recent study *Klebsiella pneumoniae* population structure showed that the *cps* operon, which determines the polysaccharide capsule type, is frequently subject to horizontal transfer between sublineages associated with distinct clinical outcomes (449). Thus capsular typing would not be a good choice for comparing the incidence of disease-causing lineages of *K. pneumoniae* over long time periods or on a global scale.

The analysis of extant genetic diversity can be used to reconstruct the evolutionary history of the organism by inferring phylogenetic trees or networks that describe the evolution of the current population from a single common ancestor some time in the past (450). For example, recent studies of pathogenic *E. coli* O157:H7 strains traced the emergence of distinct lineages of O157:H7 associated with different virulence characteristics (451, 452, 453, 454). Similarly, variation in the selective pressures upon different sites in a bacterial genome can result in variation in the level and nature of genetic diversity at those sites (455). Thus one can work backwards from current patterns of diversity within different sites in the genome to infer a history of selection at specific sites (456). For example, by comparing coding sequences from a uropathogenic *E. coli* (UPEC) genome to those of six non-uropathogenic *E. coli* genomes, Chen *et al.* (457) identified 29 genes under positive selection in the UPEC genome including genes known to be important for urinary tract infection.

1.3.1.2 Methods for studying bacterial pathogen populations

A variety of different methods have been developed for analysing the structure of bacterial populations. Each aims to subdivide the population based on discriminatory markers (typing), and to uncover the evolutionary relationships between those subdivided groups. An important goal of pathogen typing is often to compare populations over time and between geographical locations, therefore the value of a typing scheme depends not only on the discriminatory power and phylogenetic informativeness of the genetic markers used, but the ability to standardise and compare results over time and between laboratories. Based on these considerations, the current gold standard for bacterial typing is multi-locus sequence typing (MLST), which involves sequencing and comparison of a defined set of housekeeping gene fragments (458). Based on the combination of alleles at each of the gene fragments, each bacterial isolate is assigned a sequence type (ST) which can be directly compared with those of other isolates, for example using eBURST (459, 460). The assignment of STs to isolates based on sequence data is standardised via the use of international databases (e.g. (461, 462, 463, 464, 465, 466)), which facilitates direct comparison of population genetic data between laboratories and over time (467, 468, 469). MLST is based on direct determination of nucleotide sequence data, and comparative analysis of the sequences themselves is phylogenetically informative and can even be used to analyse recombination, provided there is variation within the sequenced gene fragments (470).

Unfortunately, Typhi and Paratyphi A exhibit so little nucleotide variation as to be considered “monomorphic pathogens”, for which MLST provides virtually no discriminatory power (1). The same applies to many other important human pathogens, including *Bacillus anthracis* (the causative agent of anthrax), *Yersinia pestis* (plague), *Mycobacterium tuberculosis* (tuberculosis), *Mycobacterium leprae* (leprosy) and *Shigella sonnei* (shigellosis) (471). Studies of population structure in Typhi, Paratyphi A and other monomorphic pathogens have relied on indirect typing of sequence variation including pulsed-field gel electrophoresis (PFGE), phage typing, insertion sequence (IS) typing and ribotyping. The traditional method of discriminating among Typhi isolates is phage typing (472). This involves testing the susceptibility of isolates to lysis by each of a panel of bacteriophages and comparing the patterns of bacteriophage susceptibility, or “phage types”, between isolates. This is usually done by reference laboratories. More recently, molecular typing techniques have been introduced, the most popular being PFGE (473, 474). This technique involves digesting genomic DNA with restriction enzymes and separating the resulting fragments on a pulsed-field gel. The number and sizes of the fragments depends on the distribution of restriction sites around the genome - thus mutations resulting in formation or destruction of restriction sites will affect the number and size of fragments. Fragment sizes will also be affected by gain or loss of DNA, including prophages, and by genomic rearrangements (359). Ribotyping and *IS*200 typing rely on digestion of DNA into fragments and detection of ribosomal RNA or *IS*200 probe sequences within gel-separated fragments by Southern blotting (475, 476).

These techniques are difficult to standardise and are not easily amenable to phylogenetic inference. In Typhi, PFGE profiles, ribotypes and phage types are not strongly correlated with SNP types (2). *IS*200 typing is not discriminatory within the Typhi population (6, 476) and *IS*100 typing gave an inaccurate phylogenetic picture for *Y. pestis* (43). PFGE and ribotyping are the most discriminatory within Typhi and are closely correlated, phage typing is less discriminatory and is not generally correlated with ribotyping or PFGE (6, 474). Relatively little work has been done on typing in Paratyphi A, although phage typing, *IS*200 typing, ribotyping and PFGE have been reported (8, 477, 478, 479, 480). Typing of VNTR (variable number tandem repeat) sequences have been proposed for *Salmonella* (481) including Typhi (482, 483). However a standardised set of VNTR typing loci has yet to be established for analysis of Typhi or Paratyphi A populations and the phylogenetic informativeness of the approach in these populations has not been demonstrated (43, 483). There is therefore a need to develop phylogenetically informative, standardised and reproducible methods for typing within the Typhi and Paratyphi A populations. The optimal approach would be sequence-based, and since MLST does not provide enough resolution (1) the next step is to consider analysis of much larger sequences and ultimately the whole genome.

1.3.2 Genome sequencing of bacterial pathogens

In 1977, Sanger and colleagues sequenced the first complete microbial genome - bacteriophage phi X174 (484). This was followed by other phage and viral genomes (485) until the first bacterial genome was completed nearly 20 years later. The *E. coli* genome sequencing project began in the mid-1980s and the 4.6 Mbp genome was completed in 1997 (486); however in the meantime a shotgun sequencing approach was used to sequence the complete 1.8 Mbp genome of the human bacterial pathogen *Haemophilus influenzae* in 1995 (487). The number of bacterial genome sequences available has risen steadily (see Figure 1.8), with the Genomes OnLine Database reporting over 2,500 bacterial genomes at the end of 2008, including over 750 complete genome sequences (488). Nearly 60% of these bacterial genomes are from pathogens (489), and genome-level analysis has led to the discovery of novel virulence genes and pathogenicity islands, as well as novel insights into the evolution of bacterial pathogens (490, 491). Whole-genome sequence data has also led to novel techniques for analysing bacterial pathogens at the population level, including DNA arrays to analyse variations in gene content and expression (492), and the identification of SNPs (43), small insertions/deletions (indels) (493) and VNTRs (481) with which to analyse bacterial populations.

Until recently DNA sequencing has essentially relied on Sanger's original chemistry, implemented in more automated and increasingly optimised ways thanks to numerous technological developments (recently reviewed in (494, 495)). The throughput of capillary-based Sanger sequencing technology has reached 1.6 million bp per machine per day (495), and by the end of 2008 nearly 100 billion bp of sequence had been deposited in the GenBank sequence database (496). Because of the quantity of data generated by DNA sequencing, data analysis is of central importance to the process of

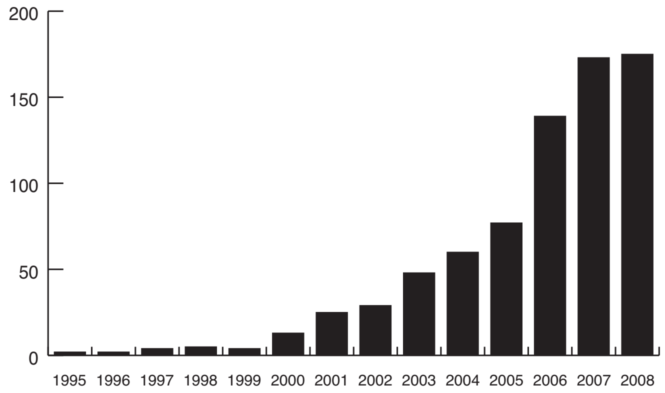


Figure 1.8: Complete bacterial genome sequences deposited in public databases - Data sourced from Genomes OnLine Database (488), July 2009. Note the data shown is the number of genome sequences deposited during each year, not the cumulative number of genome sequences in the databases.

generating genome sequences (reviewed in (497)). Capillary-based Sanger sequencing generates reads >500 bp in length (495), which must be assembled into contiguous sequences ("contigs") based on alignment of overlapping sequences (see Figure 1.9) (497). DNA fragments can be sequenced from both ends, generating "paired-end" reads which can be helpful in determining how reads fit together in the assembly. Each base in the genome is usually covered by at least 6-8 overlapping reads, giving a "read depth" or "coverage depth" of 6-8x. Thus each base in the contig sequence is supported by multiple data points (that is, bases from multiple reads), and is essentially a majority-rule "consensus" of those data points. Each base in each read can be assigned a quality score, which indicates the likelihood of it being an error. The most widely accepted quality score is the "phred" score, where a score of 10 corresponds to an error probability of 0.1, 20 corresponds to 0.01, 30 corresponds to 0.001, etc (498). These quality scores can be used in the determination of consensus sequences, so that low-quality bases are not given equal weight in a simple majority-rule consensus, and a phred-like quality score can be calculated for the consensus base itself.

Once reads have been assembled into contigs, the contigs can be arranged in correct order and orientation with the help of additional information (e.g. paired-end reads, comparison to similar genomes) to generate a scaffold (497). Gaps in the scaffold can be closed ("gap closure") using additional PCR and sequencing experiments. To-

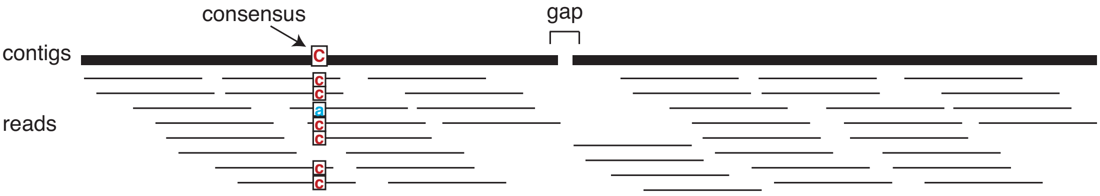


Figure 1.9: Sequence assembly - Reads are assembled into contigs (contiguous sequences) based on overlapping sequences. Gaps between contigs can be closed using additional PCR and sequencing experiments.

gether with additional experiments to resolve difficult areas like repeats or low quality sequences, this process is known as “finishing”. A “complete” genome is usually considered to be one that it entirely finished, with all gaps closed and high quality consensus bases at each position. However, many “complete” sequences have been published that contain ambiguous base calls, for example the Typhi Ty2 genome sequence (47). Once the genome is finished, or at least assembled into contigs, gene prediction and annotation can be performed (reviewed in (499, 500, 501)).

1.3.2.1 Comparative genomics

The availability of genome sequences for more and more bacteria has allowed comparisons between closely related pathogenic and non-pathogenic bacteria at the whole genome level. Comparisons of genomes separated by different phylogenetic distances can offer different kinds of insights into evolution (502). For example in bacteria, comparisons between serovars of *S. enterica* subspecies *enterica* offer different lessons from comparisons between subspecies, species of the same genera, or across genera.

Most genome sequence comparisons to date have compared isolates from different species or subspecies. They have highlighted the dynamic nature of bacterial evolution, providing evidence for the importance of horizontal DNA transfer and the acquistion of novel functions, as well as gene loss or inactivation, gene duplication and genome rearrangements (reviewed in (491)). For example, comparative analysis of *S. enterica* genomes led to the identification of many SPIs, as described in 1.1.2.2 above. Comparative genome analyses have also identified pathogenicity islands in other genera including *Staphylococcus* (503) and *Yersinia* (266), as well as horizontal acquisition of virulence genes via prophage, for example most recently in *Streptococcus* (504). Whole-genome comparisons have detected associations between specific genes and pathogenic phenotypes, for example the enterohemorrhagic *E. coli* O157:H7 genomes contained over 1,000 genes that were not present in the *E. coli* K-12 genome, including over 100 predicted to have virulence functions (505, 506). Comparative genome analysis has provided evidence of reductive evolution (i.e. loss or degradation of coding sequences) in host-adapted *S. enterica* serovars including Typhi (46), Paratyphi A (49) and Gallinarum (227). A similar trend has been observed in human-adapted species of *Bordetella* (265), *Mycobacterium* (264, 507), *Yersinia* (266, 508) and other genera. A recent study used genome sequences to compare gene content among pathogenic bacteria from a range of genera (509), an approach which may be useful for identifying targets for anti-bacterial therapeutics in the future.

Genomic comparisons between isolates of the same subspecies or even serovar have yielded further insights into the evolution of bacterial pathogens. Until recently there were few examples of whole genome sequences from multiple isolates of a single bacterial subspecies or serotype, however the availability of a single genome sequence allows the construction of DNA microarrays which can be used to interrogate gene content within a collection of isolates (492). In *Salmonella*, DNA arrays have been used to demonstrate differences in gene content between species and subspecies (510), between *S. enterica* serovars (50) and between isolates of a single serovar (49, 511). These studies have provided evidence for horizontal DNA transfer between serovars (103), including many that have not yet been sequenced, as well as highlighting specific chromosomal regions that vary within the Typhi population (prophage, SPIs and deletions) (511) or the Paratyphi A population (mostly prophage) (49). The comparison of two Typhi genome sequences (CT18 and Ty2) in 2003 revealed low levels of nucleotide variation between the isolates but some large-scale differences in prophage sequences (47). Genomic comparisons of *M. tuberculosis* isolates using genome sequences and array data revealed over 150 deletions within the population, affecting 224 genes (5.5% of coding sequences) (512). These deletions have been used as markers for epidemiological studies, which suggested certain sublineages of *M. tuberculosis* characterised by specific deletion profiles were associated with severe disease in infected patients (493). Comparative sequence analysis can also be used to identify genes under selection or associated with virulence within a particular population, including the examples given above regarding selection in uropathogenic *E. coli* (457) and the distribution of virulence genes in *E. coli* O157:H7 (453). Further examples include *H. pylori*, where severe disease is associated with the presence of a toxin and secretion system (513), and *N. meningitidis*, where DNA arrays were used to demonstrate a strong association between hypervirulence and the presence of a bacteriophage (514).

Comparative analysis of whole genome sequences has revealed vast differences in gene content between members of a single bacterial species, leading to the definition of the "pan-genome" (515). The pan-genome is the total number of genes associated with an organism and includes the core genome (genes that are conserved among all strains) as well as rarer genes that are present in some but not all strains. The pan-genome can only be characterised by sequencing multiple isolates, but the number of isolates required to adequately represent the pan-genome varies widely between bacteria. For example, in *Streptococcus agalactiae*, where the pan-genome was first described, it was estimated that each new genome would contribute over 30 novel genes to the pan-genome (515). In *B. anthracis*, it was predicted that the entire pan-genome would be sampled with just four genome sequences (515). A recent analysis of 20 complete genome sequences from *E. coli* found a core genome of ~2,000 genes and a pan-genome of nearly 18,000 genes (516). *E. coli* is incredibly phenotypically diverse, and it is likely that the *S. enterica* pan-genome is far less variable. Analysis of array data from *S. enterica* suggests that approximately three quarters of any given genome (~3,000) is core (103), although the extent of rare genes remains unknown. Within the Typhi and Paratyphi A populations, the situation is likely more akin to that of *B. anthracis*, with array data identifying only 254 CT18 genes as missing from other Typhi isolates, of which 90% were prophage genes or SPI7 genes, as SPI7 can be deleted from Typhi strains (511, 517).

1.3.2.2 SNP analysis

Single nucleotide polymorphisms (SNPs) are increasingly being used for phylogenetic analysis of bacteria, particularly monomorphic bacteria (471). SNPs are the result of substitution mutations, most often caused by uncorrected errors during DNA replication, which have become fixed within a subpopulation. The most common replication error is demethylation of cytosine to uracil (C->T) (518), thus the most common bimorphic SNP allele combinations observed are C/T or G/A (the same mutation inspected on the opposite strand) (519). If a novel SNP increases the fitness of a bacterium it may be positively selected, resulting in the novel variant becoming fixed in the local population. If the SNP decreases the organism's fitness it may be negatively selected, resulting in the novel variant being purged from the population. If the fitness difference is negligible (the SNP is neutral) or the population is small, a novel SNP may become fixed or purged by chance (genetic drift). In the absence of allele reassortment by recombination (1.1.2.3 and 1.3.1.1), SNPs are entirely vertically inherited and accumulate randomly in the bacterial genome over time. Thus SNPs provide a very strong phylogenetic signal with which to reconstruct the evolutionary history of an extant group of isolates (43, 451, 454, 520, 521). Recombination can disrupt the simple vertical inheritance of SNPs. However depending on the frequency of recombination within a bacterial population, the phylogenetic signal can often still be discerned from SNP variation (443, 454, 516). Homoplasy, i.e. identity by convergent evolution as opposed to identity by descent, is much rarer among SNPs than other kinds of genetic variants. Thus SNPs are more reliable for phylogenetic inference, which assumes identity by descent. SNPs can also be used to detect selection, most commonly by comparing the rate of nonsynonymous SNPs (dN) to the rate of synonymous SNPs (dS) within a given locus (see e.g. (457, 522)). In the absence of selection against nonsynonymous SNPs, the ratio $\frac{dN}{dS}$ should be approximately 1; positive or diversifying selection, which favours novel variation at the protein level (i.e. nonsynonymous SNPs), will result in $\frac{dN}{dS} > 1$; negative or purifying selection, which favours maintainance of the original protein sequence, will purge nonsynonymous SNPs from the population resulting in $\frac{dN}{dS} < 1$. However the interpretation of $\frac{dN}{dS}$ data needs to consider the context of the kind of population under study (level of phylogenetic distance, population size and structure) and the type of sequences examined (e.g. whole genomes, genes, protein domains, individual codons) (523, 524, 525, 526). One way to achieve this is the use of phylogenetic methods that incorporate models of nucleotide substitution, population growth and other factors into phylogenetic inference (527, 528, 529).

SNPs can be detected in a variety of ways, including denaturing high-performance liquid chromatography (dHPLC), oligonucleotide arrays and sequence analysis (including MLST) (530). The most comprehensive way to detect SNPs across the whole genome is by comparison of whole genome sequences, which is of greatest importance for monomorphic bacteria where genetic variation is too low to enable detection of SNPs from analysis of tiny fractions of the genome such as that provided by MLST. However, whichever method is used to detect SNPs, the selection of samples is crucial in order to provide an unbiased picture of variation and phylogenetic structure within a population. This is because only those SNPs lying on the evolutionary pathway separating the sampled isolates can be discovered (520, 531). Therefore in order to minimise this "discovery bias", it is important that the sampled isolates are as distantly genetically related as possible (see Figure 1.10) (471). Depending on the geographical distribution of the organism, sampling from multiple geographical locations may help in this regard. For example, SNPs discovered by comparing the first two complete *M. tuberculosis* genome sequences, both from American sources, with a strain of *M. bovis* were used in studies for typing *M. tuberculosis* isolates (532). However mutation discovery in a global collection of isolates showed significant geographical clustering of isolates and placed the two initially sequenced isolates very close together on the phylogenetic tree (533). Sample size is also a consideration, with larger samples providing greater opportunities to discover more variation, but only if phylogenetically diverse isolates are chosen (471). In the case of opportunistic bacterial pathogens, or those that colonise a variety of hosts or environmental niches, it is important to include samples from carriers, other animal hosts and/or environmental isolates in addition to human disease isolates (534, 535).

The history of SNP analysis in Typhi illustrates these points well. In 2002, the first MLST study of Typhi was published (1). Twenty-six isolates were examined at seven ~500 bp gene fragments, providing 3,336 kbp of sequence (0.07 % of the genome) in which only two SNPs were detected (1). Later, dHPLC was performed on 200 gene fragments of ~500 bp in 105 Typhi isolates (2). This is essentially the same strategy as MLST, but expanded to include 200 rather than seven gene fragments, covering nearly 2% of the genome. In order to minimise discovery bias, the 105 isolates were chosen from a diverse range of geographic locations (in Asia, South America and Africa) and

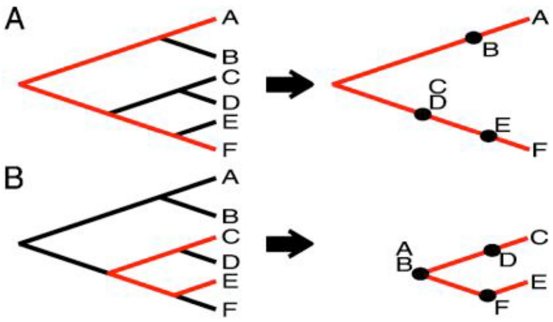


Figure 1.10: Phylogenetic discovery bias - Reproduced from (520). Original legend: "Evolutionary model showing the consequences of biased character discovery for nonhomoplastic molecular markers. (Left) The 'true' path structure of OTUs AF is shown. (A) When OTUs A and F are used for comparative character (i.e., SNPs) discovery, only mutations on the connecting evolutionary path (red) will be discovered, resulting in the disappearance of all secondary branches but showing accurate node positions of all other OTUs. (B) Similarly, if C and E are used for character discovery, only mutations on the connecting path will be discovered, causing A and B to collapse at a single point. Again, accurate node positions are retained." (OTU = operational taxonomic unit)

a diverse range of phage types, ribotypes and genome arrangements (2). Only 66 of the genes were polymorphic, with a total of 82 SNPs detected in the study, which was published in 2006 (2). The SNPs described a single, parsimonious phylogenetic tree which included multiple lineages and diversification events (see Figure 2.1). In 2007, another research group attempted to use SNPs detected between the two published Typhi genome sequences (CT18 and Ty2) to type 73 Typhi isolates (536). Thirty-six genes containing SNPs were amplified by PCR and each amplicon was digested using a restriction enzyme whose target site included the SNP locus (536). Phylogenetic analysis of the resulting data revealed 574 equally parsimonious trees, the consensus of which essentially described a line between CT18 and Ty2, with a large cluster in the middle, illustrating the problem of discovery bias. Unsurprisingly, this central cluster contained 80% of the isolates tested including a diverse range of haplotypes described in the 2006 paper (2).

Once SNPs have been detected among a discovery set of isolates, SNPs can be used to analyse population structure among a larger collection of isolates by a variety of SNP typing methods (reviewed in (530, 537)). The throughput of SNP typing methods ranges from a few SNPs to hundreds of thousands of SNPs, and from one sample up to hundreds of samples at a time. Several ultra-high throughput SNP typing techniques (targeting >500,000 SNPs) have been developed for human genotyping (reviewed in (538)) and provide far greater resolution than is required for most applications in bacteria. SNP typing studies in bacteria have generally focused on small numbers of SNPs, most recent examples include *L. monocytogenes* (8 SNPs) (539), *M. leprae* (3 SNPs) (540), *S. aureus* (9 SNPs) (541) and *Brucella* species (7 SNPs) (542). These studies use low-throughput SNP typing techniques such as allele-specific primer extension (539, 543), real-time PCR methods (542) or restriction digestion of PCR products (536, 540). These approaches are feasible up to ∼40 SNPs (536, 544) but are difficult to scale up further. Medium-throughput typing methods have been used for bacteria, including multiplex Sequenom assays (Sequenom (545)) (84 SNPs, Typhi) (256) and other mass-spectrometry based methods (546), hairpin primer assays (96-212 SNPs, *M. tuberculosis, E. coli* O157:H7) (547, 548) and SNaPshot primer extension (Applied Biosystems) (148 SNPs, *M. tuberculosis*) (532). Pyrosequencing has been used for low throughput SNP typing in *B. anthracis* (4 SNPs) (549) but could be feasibly scaled up to hundreds of SNPs (550). A high throughput method using molecular inversion probes (MIPs) (551) was recently used to type >1,500 SNP loci in *Franciscella tularensis* (544); this technology is currently scalable up to >20,000 SNPs.

1.3.2.3 New high throughput sequencing technology

Recent technological developments have led to the availability of multiple "next-generation" sequencing platforms, which offer multiple-log increases in data throughput compared to capillary-based Sanger sequencing (552, 553). Two of these platforms, 454 and Solexa, were adopted at the Sanger Institute in 2006; these and other technologies are described in detail in (552). The next-generation sequencers generate shorter reads than capillary-based sequencing (35-250 bp) and are therefore much harder to assemble (554). Most genome sequences generated by these technologies are never finished, rather they are analysed by mapping reads to a finished reference sequence, or assembled into contigs but rarely followed through to gap closure.

The 454 platform (454 Life Sciences, later Roche) was the first to become commercially available (555). In brief, DNA is fragmented and bound to microbeads on which the fragment is amplifed (one fragment per bead) (552). Pyrosequencing is used to determine the sequence of the clonal fragments clustered on each bead (one template fragment = one read per bead) (555). Amplification and sequencing is highly parallelised, generating up to 1.6 million reads per run (555, 556). The initial platform, known as the GS20, generated reads of approximately 100 bp in length. During the course of the current project, in late 2007, the FLX platform was introduced to replace the GS20 (557). The FLX can generate reads of 200-250 bp in length. Since the completion of sequencing for the current project, modifications have been introduced allowing the generation of paired-end reads using the 454 FLX (557). The most common error in 454 data is insertions/deletions (indels) within homopolymeric tracts (runs of a single nucleotide, e.g. AAAAA). This is because during pyrosequencing, the number of bases incorporated at each step must be inferred from the magnitude of the fluorescent signal generated, which loses accuracy with increasing number of bases (555, 557). Base calling for 454 data is performed using proprietary software (555). The first reports of 454 (GS20) sequencing in 2005 involved the sequencing of bacterial genomes, namely *Mycoplasma genitalium* (1 isolate) (555) and *M. tuberculosis* (4 isolates) (558). In 2006 the genome sequence of a laboratory strain of *Campylobacter jejuni* was reported using GS20 and compared to two reference sequences (559). In 2007 an isolate of *Acinetobacter baumannii*, for which no reference genome sequence was available, was sequenced with 454 GS20 followed by gap closure involving >10,000 PCR reactions and 2,200 capillary sequencing reactions (560).

The Solexa platform has been in use at the Sanger Institute since late 2006. The platform is now produced by Illumina and marketed as the Illumina Genome Analyzer (GA), but is still widely known as Solexa sequencing and will be referred to as ‘Solexa’ hereafter. Briefly, libraries are constructed by any method that generates a mixture of adaptor-flanked DNA fragments up to several hundred bp in length (561). Fragments are amplified using PCR primers tethered to a solid substrate, so that all amplicons arising from a single DNA template are physically clustered on an array (552). Sequencing proceeds by cycles of single-base extension using a mixture of four fluorescently labelled reversible terminator nucleotides, followed by four-channel imaging to determine which base has been incorporated into each cluster during the cycle, and cleaving of the fluorescent and terminator modifiers to prepare for extension during the next cycle (562).

Several million distinguishable clusters can be generated on a single array, generating one read per cluster. Each flow cell (in which the sequencing reactions take place) is divided into eight distinct lanes to which different samples may be added, allowing for example eight different bacterial genomes to be sequenced simultaneously. During the course of the current project, read lengths were limited to 36 bp and paired-end reads were not available. Recent developments have allowed read lengths to increase (that is, more cycles per run) without compromising base call quality, and paired-end sequencing is now possible (561). Throughput has also increased, from 0.5 Gbp per run to 1-2 Gbp per 36 bp paired-end run (562). At the Sanger Institute, raw data is processed via an informatics pipeline that calls bases and calibrates phred-like quality scores (562).

454, Solexa and other next-generation sequencing technologies allow vast amounts of sequence data to be generated in much shorter timespans and for lower cost than capillary-based Sanger sequencing (see Table 1.4). However the data analysis is more challenging due to the shorter read lengths, lower per-base accuracy and high number of reads (see Table 1.4) (554). There are two general approaches to analysing short read data: mapping (aligning) the reads to a reference sequence, or attempting to assemble them *de novo*. Read mapping must be able to accommodate mismatches and gaps (caused by SNPs, indels or sequencing errors), which in short reads comprises a much larger proportion of the alignment than in longer reads. Mapping algorithms may also take into account individual base qualities, which can improve the accuracy of read mapping (563). Mapped reads can be used to identify SNPs and small deletions compared to the reference sequence (564), and paired-end reads allow large insertions and deletions to be identified with confidence (565). *De novo* assembly of next-generation sequence data is also complicated by short read lengths, low accuracies and quantity of data (552). 454 data can be assembled using the proprietary software *Newbler* (454 Life Sciences/Roche). Assembly from single-end 36 bp Solexa reads is possible and can be improved by increasing read depth (566, 567). The use of paired-end reads can also improve the assembly (567), although many repetitive sequences can not be resolved without dedicated experiments for gap closure (554).

Platform	Throughput	Read length	Accuracy per base
Sanger/capillary (ABI 3730xl)	0.08 Mb/run, 1 Mb/d	$500 - 1,000$ bp	>99%
454 GS20	20 – 50 Mb/run (8 h)	100 bp	>96%
454 FLX	400600 Mb/run (10 h)	250 bp	99.5%
Solexa	215 Gb/run (2 – 8 days)	35 – 75 bp	98 - 99%

Table 1.4: Throughput and accuracy of next-generation sequencing technologies - Modified from (553), except information on 454 GS20 which was sourced from (555, 559, 560).

1.4 Thesis outline

In this thesis, the evolution and population structure of Typhi and Paratyphi A are investigated using whole genome sequence analysis. In Chapter 2, genome-wide sequence data is generated for 17 novel Typhi isolates using a combination of 454 and Solexa sequencing. Comparative analysis of these and the published genomes focuses on the detection of SNPs, deletions and variation in prophage content. Plasmid and SPI content is also examined. In Chapter 3, variation within the Paratyphi A population is investigated via comparative analysis of a novel finished genome sequence and five novel genomes sequenced with Solexa. The analysis focuses on the detection of SNPs and other variants, and provides the first glimpse of the phylogenetic structure of Paratyphi A. A novel approach is taken to identify and quantify genome-wide SNPs within a global collection of 160 Paratyphi A isolates, by sequencing pooled DNA samples using Solexa. Analysis involves validation of the approach, followed by investigation of the global population structure of Paratyphi A revealed by the pooled sequence data. Chapter 4 uses genome-wide comparisons of multiple Typhi and Paratyphi A genome sequences to investigate the convergent evolution of these pathogens. The analysis focuses on the accumulation of pseudogenes in each population and assesses the extent to which recombination between the two serovars has contributed to their convergence. Chapter 5 shifts the focus to the population of IncHI1 plasmids responsible for the spread of drug resistance within Typhi and Paratyphi A populations. A novel MDR IncHI1 plasmid sequence from Paratyphi A is compared to available plasmid sequences from Typhi and Typhimurium, with a focus on the accumulation of resistance-encoding mobile elements within the plasmids and the evidence for transfer of these elements between plasmids. Novel and published data from other IncHI1 plasmids isolated from enteric pathogens are also compared, revealing important aspects of the evolution of the plasmids and their movement between pathogen populations. Finally, Chapter 6 presents a novel high throughput SNP typing method for Typhi, drawing on chromosomal SNPs identified in Chapter 2 and IncHI1 plasmid SNPs and resistance genes identified in Chapter 5. Following validation of the method, SNP typing is applied to a global collection of Typhi isolates as well as localised collections of Typhi isolates from four endemic areas, focusing on the distribution of distinct SNP types in time and space, and the spread of IncHI1 plasmids within the Typhi population.

Chapter 2

Genomic sequence variation in Typhi

2.1 Introduction

Current knowledge of genetic variation in Typhi is limited, although it is evident that there is very little variation at the SNP level. The two fully sequenced Typhi genomes CT18 and Ty2 differed by 450 SNPs, 14 insertions/deletions including three prophage sequences and two IS, and a rearrangement between copies of the ribosomal RNA (*rrn*) operon (47). MLST analysis of 26 Typhi isolates detected only two SNPs within the 3,336 bp analysed (1) (note while this publication states there were three SNPs, this was later found to be incorrect). The study of 199 gene fragments from a global collection of 105 Typhi isolates detected 82 SNPs (2), approximately one SNP per 1,080 bp. SNPs associated with conferring resistance to fluoroquinolones have been reported in the *gyrA* gene (16, 397), however most variation in antibiotic resistance is attributed to the gain and loss of self-transmissable IncHI1 plasmids (46, 275). Thirteen insertions/deletions have been detected by microarray analysis of gene content among nine isolates (511), including four phage sequences and two insertion sequences (IS). A large deletion (labelled X in (511)) was later identified as loss of SPI7 (92), which can occur spontaneously in the laboratory and may not reflect natural variation in Typhi (517, 568). Rearrangements between the seven *rrn* operons occur frequently in the Typhi population (569), and this has been suggested to be associated with restoring balance between the replication origin and terminus following large genomic insertions/deletions (259), although it remains unclear whether the rearrangements result in significant functional changes, for example in gene expression.

The study by Roumagnac *et al.* (2) of genetic variation in 199 fragments from 105 isolates represented a leap forward in our understanding of Typhi evolution. The 82 SNPs they discovered resolved into a rooted, fully parsimonious phylogenetic tree defining 85 genetically distinct Typhi haplotypes (H1-H85, see Figure 2.1). Each node in the tree was represented by extant Typhi isolates, demonstrating that clonal replacement does not occur in Typhi, rather distinct lineages continue to persist in the population. The data provided a basis for estimating the age (10,000-43,000 years) and effective population size ($N_e = 230,000 - 1,000,000$) of Typhi. Furthermore, the distinct haplotypes defined by the 82 SNPs provide a basis for differentiating between isolates, providing an easily-interpretable alternative to the dominant techniques of PFGE, phage typing and ribotyping. The authors themselves typed 450 Typhi isolates at these SNP loci and showed that *gyrA* mutations associated with fluroquinolone resistance were over-represented within a particular haplotype (H58), which was also the most common haplotype to be found among isolates of the previous 10 years. This suggests that H58 may have undergone a clonal expansion in recent years, associated with a high rate of resistance to fluoroquinolones, currently the drug of choice for the treatment of typhoid fever. Typing on the basis of these 82 SNPs has been applied to the study of 150 Typhi isolates from Jakarta, Indonesia (256). This study demonstrated for the first time the co-circulation of multiple distinct lineages of Typhi within a defined urban area. It was also able to demonstrate that the presence of the z66-encoding linear plasmid was restricted to a single haplotype of Typhi (H59), suggesting that this rare variant, unique to Indonesia, is the result of a single plasmid acquisition event followed by clonal expansion of the recipient strain.

While it provided many novel insights, the study by Roumagnac *et al.* (2) was limited to 88,739 bp, just 1.85% of the Typhi genome, and despite including 105 isolates yielded only 82 SNPs. It was clear therefore that to discover the true extent of SNPs and other variations in Typhi, and discover novel variations providing increased resolution in typing, would require comparative analysis at the whole-genome level. Recent advances in sequencing technology make this feasible using 454 and Solexa sequencing (see 1.3.2.3), although development of appropriate analysis methods is ongoing. At the time the present study began, the 454 GS20 platform generated reads of ~100 bp using pyrosequencing (555), while the Solexa platform generated shorter reads of 35 bp (562). Both provided ~10-fold coverage of the 4.8 Mbp Typhi genome in a single experiment, and it was decided to use a combination of these platforms to sequence 17 novel Typhi genomes. The choice of isolates for sequencing has in the past been driven by clinical phenotype or simply availability. However isolate selection is critically important for comparative analysis, which can only uncover mutations that differ between the sampled isolates, a phenomenon known as discovery bias (see 1.3.2.2). To limit discovery bias as much as possible, isolate choice was guided by the phylogenetic tree defined by Roumagnac *et al.* (2) (Figure 2.1). Ten Typhi isolates from central haplotype clusters and radial haplotype groups (Figure 2.1) were chosen for sequencing using 454, which allows *de novo* assembly and therefore analysis of insertions (including prophage and IS) as well as deletions and SNPs across a broad range of Typhi lineages. Note that the publicly available sequences CT18 and Ty2 provide sequence coverage of two additional radial haplotype groups; these were also resequenced using Solexa, in order to check the published sequences and assess error rates for Solexa sequencing. To gain additional insight into SNP variation among recently expanding haplotypes, Solexa sequencing was used to generate short reads from an additional six isolates from the H58 group, which has undergone recent clonal expansion in South East Asia (2, 570) and a second isolate from the H59 group, the z66-associated lineage that is common in Indonesia (256). Three isolates, including one H58 and one H59 isolate, were sequenced using both platforms.

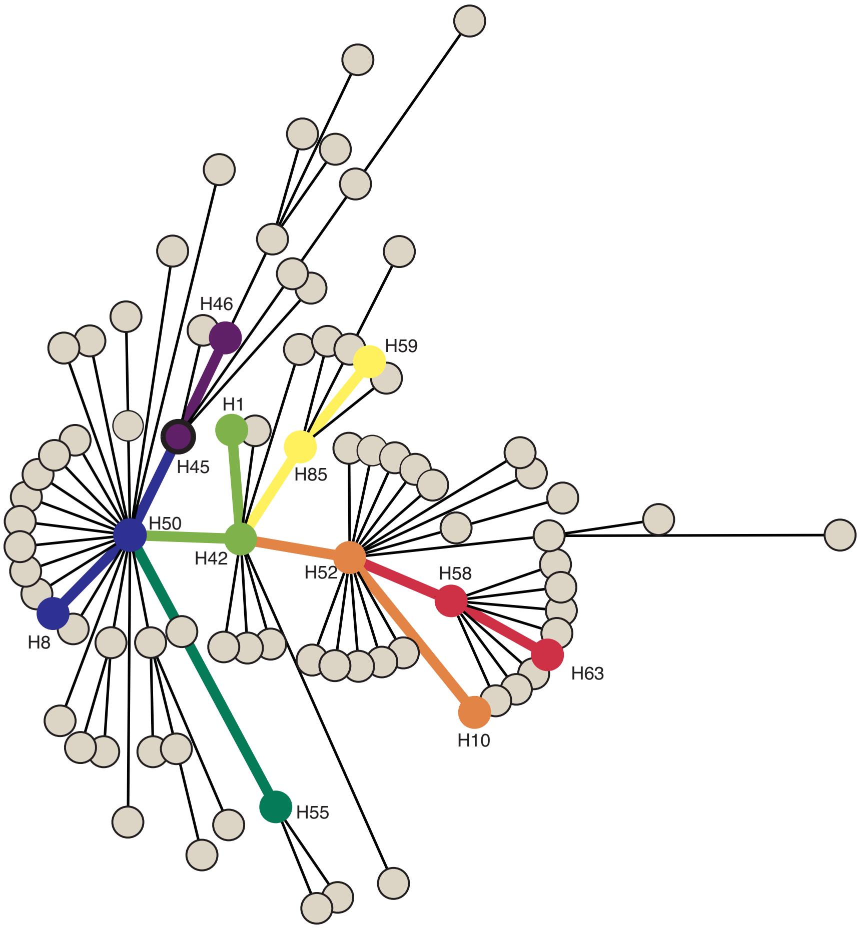


Figure 2.1: Phylogenetic tree guiding selection of Typhi isolates for sequencing - Phylogenetic tree defined previously from analysis of 97 SNPs in 481 Typhi isolates (2); root is H45. Haplotypes from which isolates were chosen for sequencing, and the branches joining them, are coloured according to the same colour scheme as Figure 2.6.

2.1.1 Aims

The work presented in this chapter is a comprehensive genome-wide survey of genetic variation among multiple Typhi lineages. As this involved a novel approach using two new high-throughput sequencing platforms, the first aim was to determine appropriate methods to detect single nucleotide variation using this novel data. Having established suitable methods, the aims of the analysis were to:

- determine the quality and quantity of genetic differences between distinct Typhi lineages, and how they may be differentiated; and
- gain insights into the evolution of Typhi, including the nature and frequency of genetic changes and any evidence of selective pressures.

2.2 Methods

2.2.1 Bacterial strains and DNA

Details of Typhi isolates used in this study are provided in Table 2.1. Bacterial cells were pelleted by centrifugation and DNA was prepared using the Wizard Genomic DNA Kit (Promega) as per manufacturers instructions. DNA preparation was performed by Dr Stephen Baker and Dr Satheesh Nair at the Sanger Institute.

Isolate	Country	Year	Haplotype	454	Solexa	Plasmid
E00-7866[1]	Morocco	2000	H46	10.5x	-	nd
E01-6750[1]	Senegal	2001	H52	8.16x	-	nd
E02-1180[1]	India	2002	H45	13.1x	-	nd
E98-0664[1]	Kenya	1998	H55	10.8x	-	nd
E98-2068[1]	Bangladesh	1998	H42	10.9x	-	nd
J185SM[2]	Indonesia	1985	H85	13.5x	-	nd
M223[3]	unkown	1939	H8	11.1x	-	nd
404ty[4]	Indonesia	1983	H59	8.49x	24.6x	pBSSB1
AG3[2]	Vietnam	2004	H58	10.1x	13.1x	nd
E98-3139[1]	Mexico	1998	H50	11.1x	5.40x	nd
150(98)S[1]	Vietnam	1998	H63	-	8.60x	nd
8(04)N[1]	Vietnam	2004	H58	-	13.1x	nd
CT18[2]	Vietnam	1993	H1	-	9.80x	IncHI1 (pHCM1), pHCM2
E02-2759[1]	India	2002	H58	-	65.5x	pHCM2
E03-4983[1]	Indonesia	2003	H59	-	7.42x	pBSSB1
E03-9804[1]	Nepal	2003	H58	-	8.19x	IncHI1
ISP-03-07467[1]	Morocco	2003	H58	-	7.87x	IncHI1
ISP-04-06979[1]	Central Africa	2004	H58	-	72.9x	IncHI1
Ty2[4]	Russia	1916	H10	-	8.60x	nd

Table 2.1: Typhi isolates sequenced in this study - Isolates were provided by: [1]Francois-Xavier Weill, Institut Pasteur, Paris, France; [2]Christiane Dolecek, Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam; [3]Barry Holmes, National Collection of Type Cultures, Colindale, UK; [4]Gordon Dougan, Wellcome Trust Sanger Institute, Cambridge, UK. Columns give country and year of isolation; haplotypes as defined in (2) and Figure 2.1; read depth from 454 and/or Solexa sequencing; plasmid content (nd=none detected, pBSSB1=z66-encoding linear plasmid).

2.2.2 DNA sequencing

Eight Typhi isolates were sequenced using a 454 Life Sciences GS20 sequencer, and an additional two isolates (M223, E02-1180) were sequenced using the 454 Life Sciences FLX sequencer. Twelve isolates were sequenced using Solexa. All steps in library preparation and sequence were performed by Ian Goodhead and Richard Rance and the Sanger Institute, according to the manufacturer's specifications. Single end reads were sequenced in all cases. Two isolates, E02-2759 and ISP-04-06979, were each sequenced over seven Solexa lanes during protocol optimisation and thus have much higher coverage than other isolates, which were sequenced in one Solexa lane each.

Sanger sequencing of PCR products was used to confirm insertion and deletion sites. Primers used for PCR and sequencing are provided in Table 2.2. PCR was performed in a 25μL volume using PCR Supermix Taq Polymerase (Invitrogen) and cycled on an MJ Research thermal cycler. Products were checked on a 0.8% agarose gel, and purified using QIAquick PCR Purification Kit (QIAGEN). PCR and purification was performed by myself, sequencing of PCR products was performed by the sequencing team at the Sanger Institute.

2.2.3 Plasmid identification

In order to verify the presence and size of plasmids within Typhi isolates, plasmid DNA was prepared from 19 Typhi isolates using an alkaline lysis method originally described by Kado and Liu (571). The resulting plasmid DNA was separated by electrophoresis in 0.7% agarose gels made with 1x E buffer. Gels were run at 90 V for 3 h, stained with ethidium bromide and photographed. High purity plasmid DNA was isolated for transformation using alkaline lysis and either AgarACE purification (Promega) or ultra-centrifugation based upon a method described by Taghavi *et al.* (572). All plasmid isolation experiments were performed by Stephen Baker at the Sanger Institute.

All plasmids detected in this way were represented in the sequence data for their host isolates and were identified by mapping to known plasmid sequences (using BLASTN for 454 contigs and Maq for Solexa reads).

Ins/Del	Strain	Forward/reverse primers	Additional sequencing primers
del A	E03-4983	TCGGCTGGAGCTAGAGAGTC, TTCACGTCCACATTCACGTT	
del B	E00-7866	AAAGTACAGGCCGGTCTCCT, CCGATAGCCCCTCTATGGAT	
del C	AG3	AAGACAACGCCAGCAGAGTTG, AATGCTGGCCAACTTCACTC	
del E	AG3	AATAGGCCTCATCACGTTCG, CAAACCGTTGAATCGGAAGT	
del F	E98-3139	CGCAATGAGCATACCTATCG, AAGCACACGACGAACAAATG	
del G	E00-7866	TCTCCCTGAGGAATCTGGTG, AAAACACCGGACAAGTCTGC	GTGAAGAAAAGCGGCTTCG, GTTGCAAGGGCGGCTTAG, GGTATTGTCGCCATTGTGC
del H	ISP-03-07467	ATTAAACCCAACGCCAACAG, GGCGAGTCTGAGCGATAAAG	CCATCGCAGACAGGACAATA, ATGAACTGGGTAGGCAAGCA
del I	E01-6750	ACGAAACGACGGGATAAGTG, GGCAAAAAGCTGGTTAAACG	
del J	E03-4983	TTCACTGCATAGCCACCATC, TACACCCCGAAAGAAACTGC	
del K	E00-7866	TGATAGAGCAGCGCATTGAC, CCGATTTCGACTGGCTGGAC	GCTACTGACGGGGTGGTG, AAGCTGCACGTAATCAGCAA
del L	E00-7866	ACGGCGTCATAACTCTCCAG, TGTCGGACGTACAGAAGAGC	
del M	E00-7866	ACAGACGGCGCAATTTATTC, GGTCATCGCGTATGAAGTCC	
del N	E02-2759	GGCCATACACTCAACCAACC, CGCCTTATCCAGCCTACATT	
del P	E98-3139	GAAGCCATTGATGAAGCACA, CACCAGCAACGACGACTCTA	
del Q	E00-7866	TGCGCTACTCAAAGACATGG, TTGATGTGGGTCAGCAAGTC	
del R	E00-7866	CAGGGAGCTCTTGGCAATAC, ACCCATTCTGGCTGAAACTG	
del S	Ty2	GACAGCATGGTGGCAAAGTT, ACCCATTCTGGCTGAAACTG	
ST16	E98-2068	GGTTCAGCAAGTGGGTTTTC, ACACCTTCGCCAGTCATTTC	
ST20a(1)	M223	CGAAAACCAACGTCACCTTT, CAAAGCAACGGAAGAATTCAA	
ST20a(2)	M223	TGCCAAGGTTCTTGATTGTG, GGAAGACTCGCTGATTTTGC	GGGATCATCGCAGCATTAGT, CGCAGAAACTGCAACACAAT
ST2-27	E01-6750	CGCGTGATATCGCCTTTATT, TACTGTCCTGTGCGATTTGC	

Ins/Del	Strain	Forward/reverse primers	Additional sequencing primers
ST36	E01-6750, 404ty	ATATCCACCAGCGAGTCCAC, TTACAGTGCGACTCCACCAG	
SPI-15b	404ty	CGGGCAAAGTTGCTTATCTC, CTGTGGGACGCTAAGTCCTC	ACCGACCGGAAAACGTTAAG, AACCACGAGCAAGCATCTG
SPI-15b	404ty	GCTTGGAAGACTCCAGAACG, TCAGCCTGTGTGTTCTTTGG	AGCGTCTTTTGTCATGGTCA, CAGGGTCTTAATCGCCAGAG, CCATCTCAGGCTTACCGAAG, CCCCTGCGCATTTAGATAGA
SPI-15b	404ty	GTGCGTTAAGCTCCTCAACC, GGCTAGGCATCTCGACACTC	GCCCAGCTACAGGTCAAAGA

Table 2.2: Primers used for PCR and sequencing of deletions and insertion sites in the Typhi gneome - Ins=insertion, Del=deletion. Deletion boundaries are given in Table 2.11. Forward and reverse primers were used for PCR; forward, reverse and additional primers were used for sequencing.

2.2.4 Phylogenetic analysis

SNPs lying within recombined regions (see 2.2.7 below) or within repeat regions (see 2.3.1.5) were excluded from analysis, leaving 1,964 SNP calls. Alleles were checked by Camila Mazzoni (Environmental Research Institute, Cork, Ireland) against an independent whole-genome multiple alignment of all 454 and published Typhi sequences generated using Kodon (Applied Maths). Alleles could be confirmed in all 19 Typhi isolates for 1,787 (90%) SNPs. These support a single maximum parsimony tree, determined using the `mix` algorithm in the `phylip` package (573) (Figure 2.6), consistent with the reference phylogenetic tree (Figure 2.1).

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

$\frac{dN}{dS}$ was calculated according to the formula $\frac{N/n}{S/s}$, where N=sum of nonsynonymous SNPs, n = nonsynonymous sites in non-repetitive protein-coding sequences (n_i above, where i = nonsynonymous), S = sum of synonymous SNPs, s = synonymous sites in non-repetitive protein-coding sequences (n_i above, where i = synonymous). The mean $\frac{dN}{dS}$ since the last common ancestor was calculated by weighting $\frac{dN}{dS}$ for H59 isolates by $\frac{1}{2}$, H58 isolates by $\frac{1}{7}$ and all other isolates by 1, so that each haplotype contributes equally. The error reported (0.053) is one standard deviation of this weighted mean.

2.2.6 Transition bias

The number of possible mutations of each type (synonymous, nonsynonymous or intergenic; transition or transversion) within each of 64 possible ancestral codons was counted. This was used to determine the total number $n_{i,j}$ of possible mutations of each type in the Typhi CT18 genome as follows:

$$n_{i,j} = \sum_{m=1}^{64} n_{i,j,m} * N_m,$$

where:
i = synonymous, nonsynonymous or intergenic,
j = transition or transversion,
m = codon,
$n_{i,j,m}$ = number of mutations of type i, j possible starting from codon m;
N_m = number of times codon m appears in Typhi CT18 (excluding repeat regions).

Transition bias in each class was calculated as follows:

$$\text{ts bias}_i = \frac{s_{i,ts}/s_{i,tv}}{n_{i,ts}/n_{i,tv}}$$

$$95\%\ \text{C.I.} = [\ \frac{s_{i,ts} - se_i/s_{i,tv} + se_i}{n_{i,ts}/n_{i,tv}}\ ,\ \frac{s_{i,ts} + se_i/s_{i,tv} - se_i}{n_{i,ts}/n_{i,tv}}\ ],$$

where:
$s_{i,j}$ = number of SNPs observed of type i, j,
n_i = number of possible mutations type i, that is $\sum_{j=1}^{2} n_{i,j}$,
$p_i = \frac{s_{i,ts}}{n_i}$,
$se_i = \sqrt{\frac{p_i(1-p_i)}{n_i}}$.

2.2.7 Detection of recombination events

Didelot *et al.*(56) found the distribution of gene-wise divergence between Typhi and its closest relatives (other serovars of *S. enterica*) predominantly followed a normal distribution with mean 1%, and the authors used divergence below 0.3% as a cut-off for identifying potential recombination events with other serovars. In order to identify potential recombination events in the present study, SNP calls from each Typhi isolate were checked for SNP clusters, defined as >3 SNPs within 1,000bp (SNPs relative to CT18, i.e. >0.3% divergence from CT18). For this analysis, SNP calls from 454 data were filtered only on consensus base quality and neighbourhood base quality. Alignments of potentially recombined sequence with other bacterial sequences were constructed using ClustalX (574) and nucleotide divergence levels calculated using the `dnadist` algorithm in the `phylip` package (573). In order to identify potential sources for the recombined DNA, the variant Typhi sequences were aligned with their homologs in Typhi CT18, *E. coli* K-12, *S. flexneri* and *S. enterica* serovars Typhimurium, Paratyphi A, Choleraesuis and Enteritidis, identified by BLASTN search of the EMBL database.

2.2.8 Evidence of expression from published microarray data

Microarray data from a study of gene expression in Typhi was available in the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/). The study (GEO accession GDS231) included 24 microarray experiments on RNA extracted at five time points after treatment with 1 mM peroxide (575). No Typhi expression data was available in the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/), as at January 17, 2008. A gene was considered to be expressed (but not necessarily differentially expressed at different time points) if its measured expression level ranked in the top 80% of genes in one or more experiment. Most of the genes examined ranked in the top 10% in at least one experiment; the maximum percentile rank of each putative pseudogene is shown in Appendix A.

2.2.9 Accession codes

Raw sequence data generated in this study is available in the EBI Whole Genome Shotgun (WGS) database (454 de novo assembled contigs, accessions CAAQ - CAAZ) and European Short Read Archive (Solexa reads, accession ERA000001). In addition, mapped assemblies of all 454 and Solexa datasets, including plasmid sequences, are available online at http://www.sanger.ac.uk/Projects/S_typhi. Accession IDs of published genome sequences used for comparative analysis, including determining ancestral alleles, are: Typhi strain CT18 (AL51338), Typhi strain Ty2 (AE014613), Typhimurium strain LT2 (AE006468), Paratyphi A strain ATCC9150 (CP000026) and Choleraesuis strain SC-B67 (AE017220); *E. coli* K12 (NC_000913) and *Shigella flexneri*

5 strain 8401 (CP000266). Enteritidis strain PT4 sequence was downloaded from http://www.sanger.ac.uk/Projects/Salmonella. Accession IDs of plasmid sequences used for comparative analysis are: pHCM1 (AL513383), pHCM2 (AL513384), pBSSB1 (AM419040), pAKU_1 (AM412236).

2.3 Results

In order to capture as much information as possible about the distribution of genomic variation in the Typhi population, DNA prepared from CT18, Ty2 and seventeen other isolates was subjected to a combination of 454 and Solexa sequencing (Table 2.1, see Methods). Since the resulting sequence data was among the first to be generated with these new technologies, and data analysis methods were still being developed, it was important to determine the best methods for detecting single nucleotide polymorphisms (SNPs) from the data.

2.3.1 Assessment of SNP detection methods

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

At 100 bp on average, 454 reads were long enough to be assembled *de novo* into contigs (i.e. without reference to any other sequence). Thus two approaches were available for the detection of SNPs from 454 data: (i) align reads directly to a reference sequence, or (ii) assemble reads into contigs and align contigs to a reference sequence. To determine the best method for analysing 454 data, SNP detection error rates were calculated using real and simulated reads, and real and simulated contigs. Reads were aligned to a reference using ssahaSNP (http://www.sanger.ac.uk/Software/analysis/ssahaSNP/). Contigs were aligned to a reference using MUMmer (v3.19, `nucmer` algorithm) (576). These free, opensource software packages combine alignment with SNP detection, and provide flexible parseable output which facilitates high-throughput analysis. Unlike ssahaSNP, most other software specifically developed for detecting SNPs from sequencing reads utilise raw fluorescence data generated by capillary sequencing (e.g. SNPdetector, NovoSNP, PolyBayes), which is not appropriate for analysing data generated by 454 pyrosequencing. 454 provide their own SNP detection algorithm within their mapped assembly software, however this proprietary software is something of a ‘black box', and appears to miss a large number of true SNPs (e.g. it detected on average 40 SNPs per Typhi genome compared to CT18, whereas the results of all other analyses reported here suggest 5-10 times this level of variation). MUMmer is also unique in that it facilitates fast and free whole-genome comparison combined with SNP detection; other options include diffseq (EMBOSS package, (577)) which can't handle genome rearrangements, Mauve (578) which does not include SNP detection, and the proprietary software package Kodon (Bionumerics).

Typhi reads were simulated by introducing 200-1000 SNPs into the CT18 finished sequence and randomly sampling 100 bp reads at 8x, 10x, 20x and 40x read depth. Substitution errors and insertion/deletion errors were also introduced into the simulated reads, at rates determined from ssahaSNP analysis of real 454 data from *Streptococcus suis* (exponential distributions with means 2% and 1% respectively, see Figure 2.2). Simulated Typhi reads were aligned back to the CT18 reference sequence using ssahaSNP and false positive and false negative rates calculated. A minimum depth of 3 or 5 reads was required to call a SNP, and the minimum proportion of reads calling a SNP (out of those reads mapped to the SNP locus) was allowed to vary between 0.5-0.9. SNPs introduced within repetitive sequences (defined in 2.3.1.5) were excluded from analysis.

Contigs were simulated from reads, by determining the number of times each base in the reference sequence was sampled and breaking contigs at any point where a base was covered by ≤ 1 read. Note that it was not possible to assemble simulated 454 reads directly, as the 454 software Newbler performs assembly using raw pyrosequencing data (i.e. in signal space) and not base-called reads (i.e. in nucleotide space). Simulated contigs displayed a realistic relationship between read depth and genomic coverage (% of the reference genome covered by contig sequences) (Figure 2.3), however contigs became unrealistically large at high read depth (Figure 2.4) as real assemblies are limited by the presence of unresolvable repeat sequences.

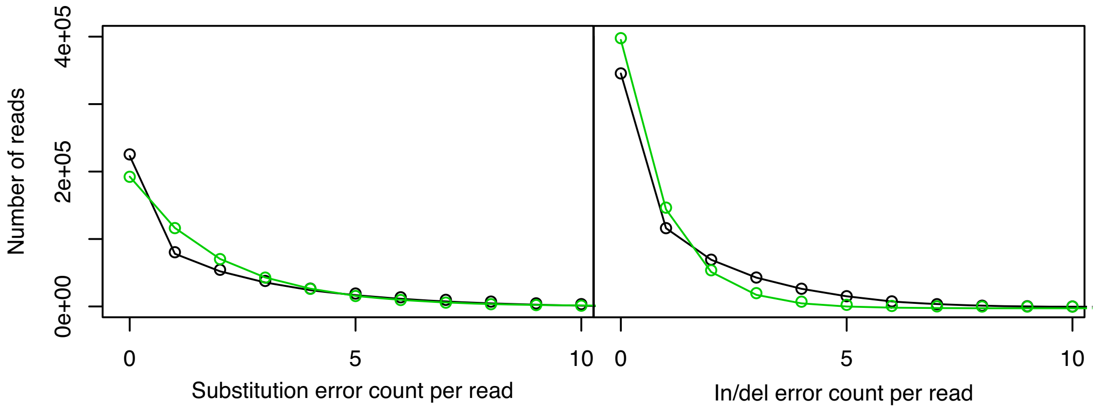


Figure 2.2: Error models for 454 reads - Real frequencies of errors in 454 reads from *Streptococcus suis* P_17 are shown in black; these were determined by mapping reads to the finished sequence. Note *S. suis* data was used because both 454 data and finished sequence was available for the same isolate; no such data exists for Typhi. Frequencies expected under an exponential distribution are shown in green.

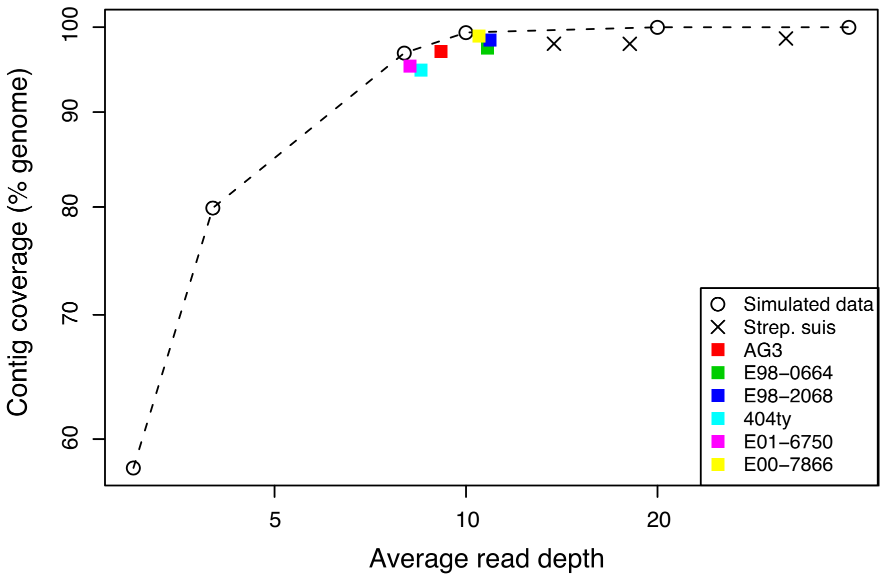


Figure 2.3: Read depth vs genome coverage for real and simulated 454 data - Coloured squares represent real data from Typhi strains.

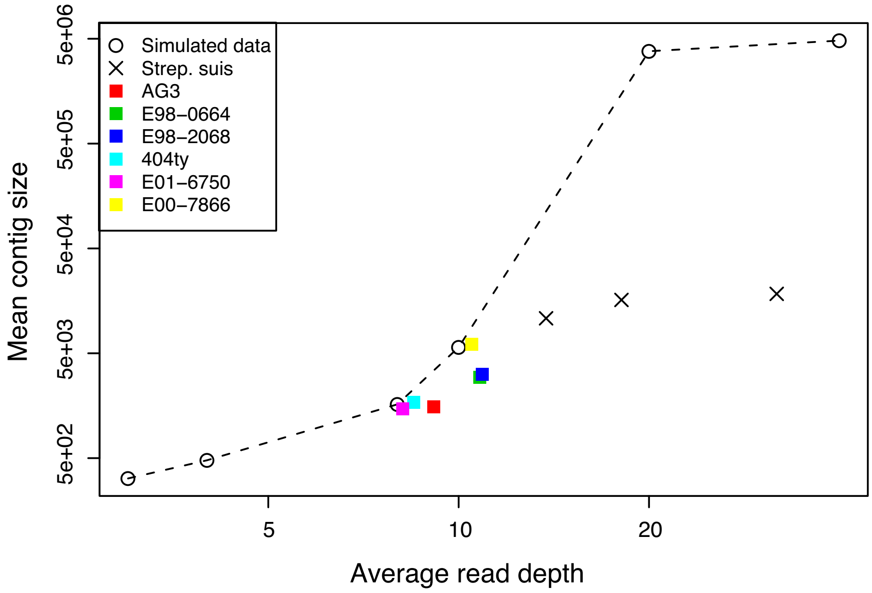


Figure 2.4: Read depth vs mean contig size for real and simulated 454 data - Coloured squares represent real data from Typhi strains.

The error rates estimated from simulated data are shown in Table 2.3. Using contigs, false positive rates were low (< 3SNPs per strain) and false negative rates were acceptable (≤11% of SNPs introduced during simulation but not detected) for read depths ≥8. Using reads, false positive rates were minimal with high coverage (≥40x) and could be controlled at lower depths (≥8x) using more stringent parameters. However false negative rates were much higher at these more stringent parameter settings, especially at lower read depths (e.g. 30-50% of SNPs undetected at 10x read depth). Thus analysis of assembled contigs appears to be a more accurate method for SNP detection in 454 GS20 data.

False positive rates were further investigated by comparing real data generated by 454 sequencing of *S. suis* strain P_17 to the finished sequence of the same strain, sequenced previously at the Sanger Institute (http://www.sanger.ac.uk/Projects/S_suis/). The data included read sequences and *de novo*-assembled contig sequences from two 454 runs, providing 20x and 27x coverage of the 2.0 Mbp genome. Using ssahaSNP to align reads directly to the finished sequence (minimum depth=5, p=0.5) resulted in 104 SNP calls, all of which lay in repetitive sequences in the finished genome. Contigs

Read	**Contigs**	**Reads (ssahaSNP), depth≥3**			**Reads (ssahaSNP), depth≥5**		
depth	**(MUMmer)**	**p≥0.5**	**p≥0.7**	**p≥0.9**	**p≥0.5**	**p≥0.7**	**p≥0.9**
8x	3-11%	21-36%	22-38%	35-48%	50-66%	50-66%	58-73%
	0-3	1972-2512	545-725	20-50	618-967	53-111	0
10x	0-5%	10-21%	11-24%	24-38%	28-45%	30-47%	42-57%
	0-1	1251-1637	256-389	3-27	487-743	29-73	0
20x	0-1%	0-2%	0-3%	8-24%	0.5-6%	0.5-7%	9-24%
	0	48-84	1-11	0	33-63	0	0
40x	0-1%	0-0.5%	0-0.5%	8-18%	0-0.5%	0-0.5%	8-18%
	0	0	0	0	0	0	0

Table 2.3: Error rates in SNP detection using simulated sequence data - False negative rate (% of simulated SNPs that were not detected) and number of false positive SNP calls using analysis of simulated contigs and reads. For read analysis, various quality cut-offs were trialled including minimum read depth at SNP locus of 3 or 5, and minimum proportion (p=0.5-0.9) of mapped reads that must include the SNP allele.

were compared to the finished sequence using MUMmer, which identifed no SNPs (note MUMmer does not report SNPs in repetitive sequence). The data are consistent with the simulated data which suggested near-zero false positive rates for SNP detection in 454 of samples of ≥ 20x read depth, using either contigs or reads.

The total number of SNPs determined by analysis of reads or contigs was compared for ten Typhi isolates sequenced on the 454 platform to 8-13x read depth, see Table 2.4. Note these numbers exclude SNPs called in repetitive sequence (see 2.3.1.5 below) and SNPs with a consensus base quality of <30 in assembled contigs. Analysis of contigs detected 303-652 SNPs per isolate, while analysis of reads detected 206-755 SNPs per isolate (using a cut-off of depth ≥5 to call a SNP). Of the 2,243 SNPs detected from contigs, 61.4% were detected by reads analysis with depth ≥5, and 79.5% with depth ≥3. Conversely, only 40-50% of the SNPs detected from reads analysis were detected by contig analysis. This is consistent with low error rates for contig analysis and high error rates for read analysis, as suggested by the simulations at read depths in this range (8-10x). It is possible that reads analysis has identified hundreds of genuine SNPs that were missed by contig analysis, but given the high false positive rate estimated from reads simulated at 8-10x depth (Table 2.3), the former explanation appears more likely. Thus for the remainder of this study, SNP detection from 454 data was performed by analysing assembled contig sequences rather than reads.

			Reads, depth≥3		**Reads, depth≥5**	
Isolate	**Read depth**	**Contigs**	**p≥0.5**	**p≥0.9**	**p≥0.5**	**p≥0.9**
E00-7866	10.5	652	672	605	597	530
E01-6750	8.16	347	243	230	118	105
E02-1180	13.1	653	1018	857	755	753
E98-0664	10.8	380	317	313	176	172
E98-2068	10.9	303	266	261	130	125
J185SM	13.5	373	394	313	239	189
M223	11.1	489	894	463	712	331
404ty	8.5	435	342	333	140	131
AG3	10.1	365	425	338	327	258
E98-3139	5.4	426	271	237	235	206

Table 2.4: SNPs detected in Typhi 454 data by analysis of contigs and reads - Number of SNPs detected by comparison of Typhi 454 contigs or reads to the CT18 reference. Contigs were analysed using MUMmer, reads with ssahaSNP. For read analysis, various quality cut-offs were trialled including minimum read depth at SNP locus of 3 or 5, and minimum proportion (p=0.5 or 0.9) of mapped reads that must include the SNP allele.

2.3.1.2 Solexa data

Solexa reads were too short to be assembled effectively using available software, thus were mapped directly to the CT18 reference sequence using Maq v0.6.0 (564), which was also used to generate primary SNP calls. Maq was chosen over ssahaSNP because it uses a more sophisticated method that calculates a "mapping quality" for each read aligned to the reference sequence, which it takes into account during SNP calling and quality estimation. However, Maq is unable to handle reads of >80 bp and so was not considered for analysis of 454 data.

2.3.1.3 Determining quality filters for SNP detection

To avoid SNP calls due to errors in assembly or base calling in 454 contigs, SNPs with low base call quality or low neighbourhood quality were filtered out before further analysis. SNP calls close to contig ends were also filtered out, as base call errors are more common in the low read-depth regions at the ends of contigs. Similarly, SNPs called from analysis of Solexa reads were filtered out when read depth or consensus base quality was low. Appropriate thresholds for all filters were determined by comparison of results from three isolates sequenced using both 454 and Solexa (AG3, E98-3139, 404ty). The set of unfiltered 454 SNP calls (using MUMmer and *de novo* assembled contigs) that overlapped with unfiltered Solexa SNP calls (using Maq and Solexa reads) was used to approximate the 'true' set of SNPs for each of the three isolates (excluding any calls in repetitive sequences). The distribution of quality parameters among 'true' SNP calls and those detected by only one platform (Figure 2.5) were used to define the threshold values for each platform, given in Table 2.5.

454	**Solexa**
consensus base call quality ≥ 30	read depth ≥ 5
$\leq 2/10$ surrounding bases with quality <30	Maq consensus base quality ≥ 30
>15 bp from end of a contig	no heterozygous base calls

Table 2.5: Thresholds for filters used during SNP calling - For 454 sequences, SNPs were detected in assembled contigs using MUMmer; consensus base call quality is calculated during contig assembly. For Solexa sequences, SNPs were detected in reads using Maq; consensus base call and quality is calculated by Maq for each position to which reads are aligned; a consensus base call can include heterozygous base calls (given by IUB codes e.g. Y = C and T), but all such loci were filtered out of SNP analysis.

2.3.1.4 Estimating error rates for SNP detection

The comparison of results from three isolates sequenced using both 454 and Solexa was also used to estimate error rate. SNP calls filtered using the cut-offs determined in 2.3.1.3 were compared to the 'true' SNP set (overlap between unfiltered SNP calls from 454 and Solexa) to estimate the false positive rate: platform-specific calls/'true set', mean 2.7%; and the recovery rate of 'true' SNPs: filtered calls/'true' set, mean 81%

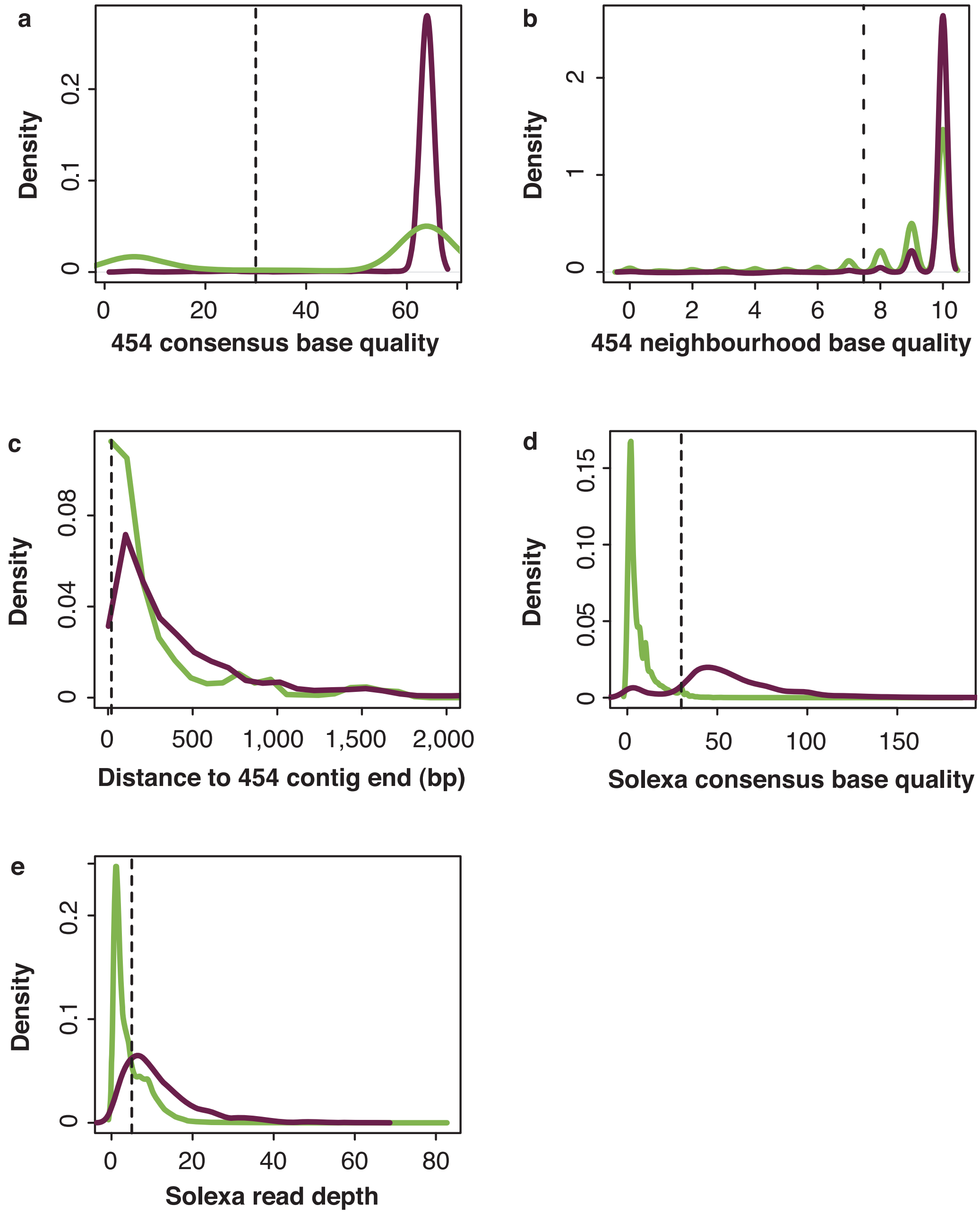


Figure 2.5: Distribution of quality parameters for SNP detection - Purple lines show distribution of each parameter among SNPs called in both 454 and Solexa data, pooled for three isolates (AG3, E98-3139, 404ty). Green lines show distribution of each parameter among SNPs called in only 454 or Solexa data but not both. Dashed lines indicate the cut-off value of each parameter used in subsequent SNP detection analysis. (a) Consensus base quality at SNP locus in 454 *de novo* assembly. (b) Number of bases 5 bp either side of SNP that have consensus base quality >30 in 454 *de novo* assembly. (c) Distance from SNP locus to nearest 454 contig end. (d) Consensus base quality at SNP locus in Solexa mapping. (e) Solexa read depth at SNP locus.

recovery rate (see Table 2.6). The recovery rate was estimated again after checking each isolate for alleles at SNP loci identified in any of the three isolates (62-97% recovery) or in any other isolate (82-99% recovery).

Strain	Depth	Overlap	False pos.	Recovery	Post-check (3)	Post-check (all)
454:						
AG3	10.1	356	4.3%	87.4%	89.0%	97.6%
404ty	8.5	307	1.1%	89.6%	91.9%	97.4%
E98-3139	11.1	517	0.0%	77.9%	77.9%	82.0%
mean	-	-	*1.8%*	*85.0%*	*86.3%*	*92.3%*
Solexa:						
AG3	13.1	356	2.3%	85.4%	91.3%	99.7%
404ty	24.6	307	4.5%	97.1%	97.1%	99.3%
E98-3139	5.4	517	1.2%	48.7%	62.1%	82.6%
mean	-	-	*2.7%*	*77.1%*	*83.5%*	*93.9%*

Table 2.6: Estimated measures of SNP detection accuracy - Estimates of false positive (false pos.) rate and percentage of SNPs recovered (recovery), assuming SNPs called independently in both 454 and Solexa data sets represent the 'true' set of SNPs in each isolate. All sensitivity estimates were made after filtering SNP calls by the stated quality criteria 2.3.1.4. Post-check = sensitivity after checking each isolate for SNP alleles at all loci for which a high-quality SNP was detected in another isolate (using data from these 3 isolates, or all 19 isolates).

Typhi isolates CT18 and Ty2, which have previously been sequenced and finished (46, 47), were resequenced using Solexa and mapped to the published sequences using Maq as described above (2.3.1.2). At the cut-offs used for SNP detection in this study (2.3.1.3), 42 differences were detected between the Solexa and published CT18 sequences. Checking the capillary sequencing traces at these loci showed that 35 of these differences were due to errors in the published sequence (approximately 1 in 140 kbp). The remaining seven differences are likely errors in the Solexa sequence, with quality scores in the range 30-37. Two of these differences occur in repetitive regions excluded from our study, resulting in an estimate of five false SNP calls for this data set. Nineteen differences were detected between the Solexa and published sequences for Ty2. The Ty2 capillary sequencing traces were not available for checking, however three of these loci were given ambiguous base calls in the published sequence and at a further six loci the Ty2 Solexa base call matched those of the remaining eighteen

Typhi isolates. It is therefore presumed that these nine differences were errors in the published sequence, leaving an estimated ten errors in the Solexa data (quality scores 30-54). It is also possible that some of these SNP calls represent genuine mutations arising in the laboratory. The Ty2 sequence was assembled with reference to the CT18 sequence (47), thus it is unsurprising that the Ty2 published sequence had a lower error rate (9 vs 35 bases). However the Ty2 sequence was not fully finished (47), thus the CT18 is considered the more reliable reference sequence for this study. Overall, the comparison of CT18 and Ty2 Solexa data with finished sequence data suggested 5-10 false SNP calls per Solexa genome, which agrees with that estimated by comparison of 454 and Solexa sequence data (6-14 SNPs per genome, Table 2.6).

2.3.1.5 Minimisation of potential errors

SNP analysis focused on the non-repetitive component of the genome and did not attempt to identify single base indels. Repetitive sequences, including VNTRs, exact repeats of ≥20 bp, >95% identical repeats of >50bp, phage and insertion sequences (IS), account for 7.4% of the CT18 genome (Table 2.7). In this study, these classes of repetitive sequences were excluded from SNP analysis as (a) non-identical repeats can appear indistinguishable from SNPs, particularly with short sequencing reads (100-250 bp for 454, 25 bp for Solexa), (b) assembly and mapping of short reads are unreliable in repetitive regions, and (c) repeated regions may be subject to different selective pressures compared to the rest of the genome, e.g. recombination between repeat copies. All prophage sequences were excluded on the grounds that they are subject to horizontal transfer and may therefore confuse phylogenetic signals, in addition to concerns regarding sequence similarity between prophage of different origins. 11% of initial SNP calls lay within repetitive or phage sequences (Table 2.7) and were excluded fro analysis.

SNPs called within 10bp of another SNP or gap within a single genome ("mismatch clusters") were examined further to identify if they were due to errors in or near homopolymeric tracts, which can be problematic for 454 pyrosequencing (555), or due to misassembly or misalignment of sequences in non-identical repeats. To eliminate these errors, mismatch cluster SNPs were removed if they were (a) within or adjacent to a tract of three or more identical bp (44%), or (b) BLASTN search of the surrounding

(a) Genomic bp	Excluded	Included	Total	% Included
Intergenic	260657 bp	510034 bp	770691 bp	66.2
rRNA	32797 bp	0 bp	32797 bp	0.0
tRNA	5159 bp	1024 bp	6183 bp	16.6
Protein coding	56905 bp	3942461 bp	3999366 bp	98.6
All bases	355518 bp	4453519 bp	4809037 bp	92.6
(b) Genes				
Total	390	4210	4600	91.5
IS elements	36	0		
Phage-like	10	0		
Phage	324	0		
Other	20	0		

Table 2.7: Repetitive Typhi CT18 sequences excluded from SNP detection anlaysis - Details of (a) genomic nucleotides and (b) genes in the CT18 genome that were included or excluded from SNP detection analysis.

region (50bp each side) returned multiple hits of >80% identity in the CT18 reference genome (25%). The remaining mismatch cluster SNPs were manually inspected for potential alignment errors (contig alignments with CT18), assembly errors (reads alignments with CT18) or recombination with a source outside Typhi (contig alignments with other bacteria, by BLASTN search of EMBL prokaryote database). Of these, 38% were consistent with recombination (see below and Table 2.10), 52% appeared to be assembly errors and 10% (22) were deemed to be real SNPs, with properly aligned reads consistently containing the SNP allele.

Filtered SNP calls were combined into a single list of SNP loci, and the allele at each locus determined in each of the 19 Typhi sequences and additional *S. enterica* serovars (using fasta3 search for 454 contigs or finished sequences, and Maq consensus base calls for Solexa data). This allowed recovery of some SNPs that were initially rejected in one isolate due to low confidence, but detected with high confidence in a second isolate. For example, in the three strains sequenced using both 454 and Solexa, it was estimated that this form of allele checking could improve recovery rates by up to 33% (Table 2.6). Detection of alleles in other *S. enterica* serovars provided an outgroup for phylogenetic analysis.

Nonsense SNPs were verified by manually inspecting multiple alignments of all 454 and Solexa reads mapping to each nonsense SNP locus. SNP calls were not verified by capillary sequencing, as this would be extremely labour intensive and contribute very little increase in depth of coverage to what is already available in the 454 and Solexa data sets. However it is expected that SNP detection errors will be randomly distributed within the Typhi genome and should not introduce significant bias into the analysis which would invalidate the conclusions drawn.

2.3.2 SNP analysis

In summary, *de novo* assembled 454 contigs and Solexa reads were aligned to the finished CT18 sequence using MUMmer and Maq, respectively, and filtered as described (2.3.1.3, 2.3.1.5). SNP calls from 19 strains were merged, resulting in 1,964 high quality SNPs, approximately 1 in every 2,300 bp of non-repetitive genomic sequence. Details of these SNPs are availabe as Supplementary Material in (579). Complete allele data from 19 Typhi genomes were determined for 1,787 SNPs (missing data were due to low coverage or deletion of SNP loci in one or more isolates). Alleles were also determined in several other *S. enterica* serovars (2.2.9) to differentiate ancestral from derived alleles and provide an outgroup for phylogenetic analysis.

A rooted maximum parsimony tree was fit to this set of 1,787 SNPs. The tree was consistent with the previously defined minimum spanning tree based on 82 SNPs (2) (Figure 2.1), while providing better estimates of branch lengths and greatly increasing resolution, particularly within the H58 and H59 groups (Figure 2.6). Only ten SNPs (0.56%) did not fit the previously determined phylogenetic tree, two of which are confirmed examples of convergent evolution at sites under adaptive selection in *gyrA* (see 2.3.2.2 below). Thus there is little reason to suspect high error rates among allele assignments, or to doubt the phylogenetic tree structure shown in Figure 2.6. Using this phylogenetic tree, mutations were grouped into relative age groups including: (a) recent mutations, furthest from the root and lying on intra-haplotype branches, (b) intermediate mutations, lying on haplotype-specific branches, and (c) older mutations, lying on branches closest to the root and shared by multiple haplotypes. The distribution of SNPs and other variants in each group is shown in Table 2.8.

SNPs were more common in non-protein-coding sequences (mean 0.051% divergence), with 86.7% of SNPs in protein-coding sequences (mean 0.043% divergence) which make up 88.5% of the non-repetitive CT18 genome (χ^2 test, p=0.01). Transition mutations (purine to purine G<->A, or pyrimidine to pyrimidine C<->T) were much more frequent than transversion mutations (purine <-> pyrimidine): 24-fold (95% confidence interval 17-40) higher among synonymous SNPs, 16-fold (14-20) higher among nonsynonymous SNPs and 13-fold (9-21) higher among non-coding SNPs. Note these rates are normalised to the number of available sites for transitions and transversions in each class of SNPs (see 2.2.6 above), so are not explained by the fact that e.g. transitions are more likely to be synonymous than transversions. The mutation bias towards transitions has been determined experimentally to be 2-fold in Typhimurium (580), thus the much higher bias indicated here may reflect selection bias in addition to mutation bias in favour of transitions. By far the most common mutations (75%) were the transitions $G->A$ and $C->T$, consistent with the observation that deamination of cytosine to uracil frequently escapes DNA repair (518).

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

The mean $\frac{dN}{dS}$ of each isolate compared with the last common ancestor was 0.66 ± 0.053 (s.d.) (see 2.2.5), suggesting either a weak trend in the direction of stabilising selection since the last common ancestor of Typhi, or a combination of stabilising selection in some genes and diversifying selection in others. Since there is little evidence of diversifying selection in any Typhi genes (see below, Table 2.9), weak stabilising selection is most likely. The weakness of the signal for stabilising selection observed here may be due to too little time for selection to act, and/or genetic drift due to low effective population size. Rocha *et al.* (526) showed that in closely related bacteria the reciprocal of $\frac{dN}{dS}$, $1/\frac{dN}{dS}$, is related to time. Their simulations indicated that when population size was large this relationship was linear, but when effective population size was small genetic drift became more important and $1/\frac{dN}{dS}$ reached a plateau. The relationship of $1/\frac{dN}{dS}$ to the number of intergenic SNPs for pairwise comparisons of sequenced Typhi isolates was non-linear (Figure 2.7a). Intergenic SNPs serve as an approximation of time, as they are less likely to be under purifying selection than SNPs in coding regions.

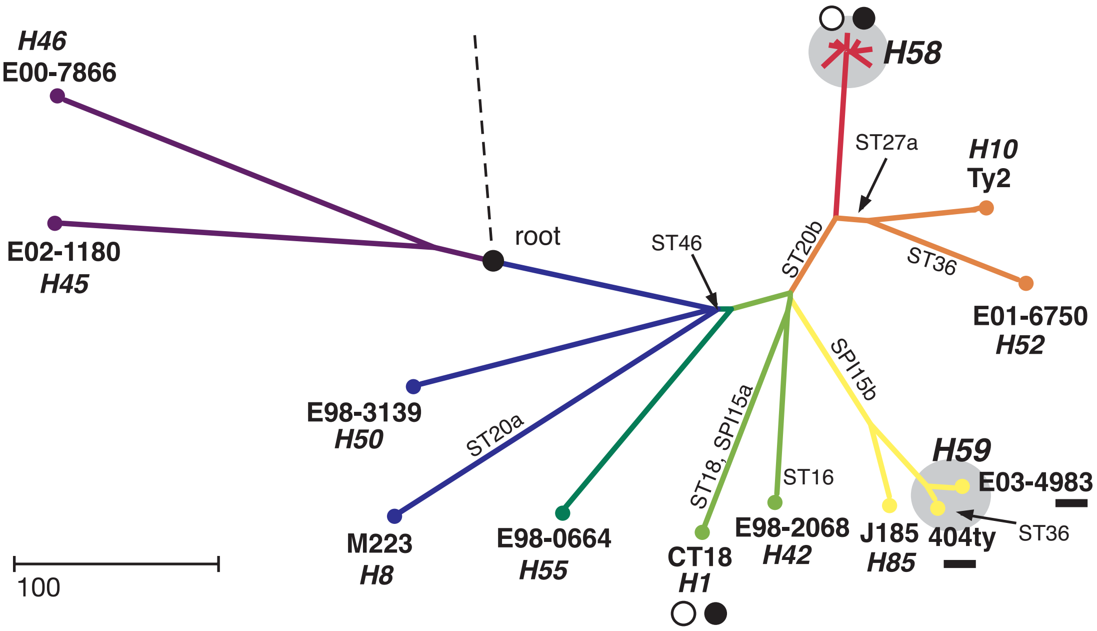


Figure 2.6: Phylogenetic tree of Typhi based on SNP data - Branch colours and lengths are consistent with Figure 2.11; branch lengths are measured in number of SNPs, scale as indicated. Black circle indicates the ancestral root, dashed line represents the link to other *Salmonella*; phage (ST) and SPI15 insertion events are labelled on the branches on which they occurred; plasmids detected in each isolate are indicated by filled circles (IncHI1 multidrug resistance plasmids), open circles (cryptic plasmid pHCM2) and filled lines (linear plasmid pBSSB1 carrying z66 flagella variant); shaded ovals group together multiple isolates of the same haplotype.

Variation type	**(i) Intra-haplotype**		**(ii) Inter-haplotype**		**(iii) Conserved**		**Total**
Deletions	5		8		7		20
Phage insertions	n/a		5		4		9
Plasmids	3		2		0		5
SNPs (complete)	93		1356		338		1787
- Intergenic	6	(6.5%)	177	(13.1%)	44	(13.0%)	227
- Synonymous	21	(22.6%)	477	(35.2%)	106	(31.4%)	604
- Nonsynonymous	61	(65.6%)	663	(48.9%)	176	(52.1%)	900
- Nonsense	5	(5.4%)	39	(2.9%)	12	(3.6%)	56
- dN/dS	0.98		0.46		0.52		0.49
SNPs (incomplete)	19		122		35		176
- Intergenic	4	(21.1%)	24	(19.7%)	6	(17.1%)	34
- Synonymous	3	(15.8%)	41	(33.6%)	12	(34.3%)	56
- Nonsynonymous	12	(63.2%)	57	(46.7%)	17	(48.6%)	86
- Nonsense	0	(0.0%)	0	(0.0%)	0	(0.0%)	0
- dN/dS	1.24		0.44		0.44		0.48

Table 2.8: Genetic variation detected in 19 Typhi genomes - Frequency of mutations in three relative age groups; percentages give relative frequency of each SNP class within each age group. SNP data is split into two groups depending on whether alleles could be reliably determined for all isolates (complete allele data) or not (incomplete allele data). Total counts of each variant are given in the last column, which also gives the $\frac{dN}{dS}$ ratio calculated across all three relative-time groups.

However intergenic SNPs may have regulatory or other functions which may be under selection, so as an alternative measure $\frac{dN}{dS}$ was also calculated among SNPs of different relative ages (a-c above), which confirmed a non-linear trajectory (Figure 2.7b). In the light of the previously described model (526), these patterns are consistent with genetic drift in Typhi due to a small effective population size, which appears likely as Typhi has no known reservoir outside of humans. A small effective population size ($N_e = 2.3$ x $10^5 - 1.0$ x 10^6) has been calculated previously using Bayesian skyline plots based on 82 SNPs in 105 Typhi isolates (2).

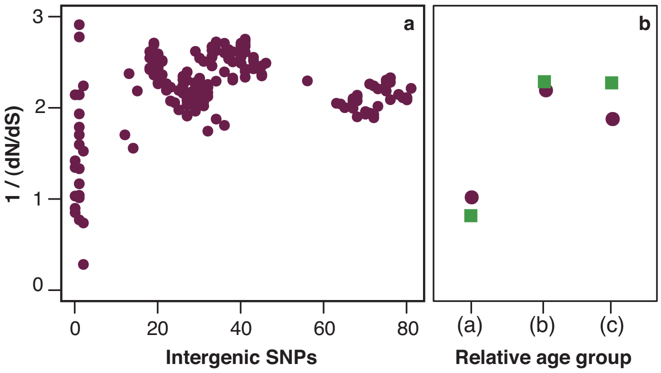


Figure 2.7: Trajectory of $\frac{dN}{dS}$ over time in Typhi - Y-axis is the reciprocal of $\frac{dN}{dS}$, or $1/\frac{dN}{dS}$. (a) Pairwise $1/\frac{dN}{dS}$ between 19 Typhi isolates vs pairwise number of intergenic SNPs. (b) $1/\frac{dN}{dS}$ for SNPs in three relative age groups (a=youngest, c=oldest), calculated from SNPs with complete allele data in 19 isolates (purple circles) and all SNPs including those with incomplete allele data (green squares).

2.3.2.2 Potential signals of selection

There was very little evidence of adaptive selection in Typhi genes, which would be represented by an overabundance of nonsynonymous SNPs or independent changes in the same or nearby amino acid residues. Nearly three quarters (72%) of Typhi genes contained no SNPs and the distribution of SNPs per gene followed a Poisson distribution in the range 0-6 SNPs per gene, shown in Figure 2.8. However, there were a few exceptions, listed in Table 2.9. Three genes (*yehU*, *tviE* and STY2875) contained more than six SNPs, which deviates from the Poisson model. STY2875 is an exceptionally large gene (3,625 bp compared to the genome mean of 910 bp), which may account for the number of SNPs. However *yehU* and *tviE* are small (562-579 bp) and thus the high number of SNPs may be evidence of diversifying selection in these genes, the second of which is encoded in SPI7 and involved with Vi synthesis (92). Ten SNPs did not fit the phylogenetic tree, which may indicate either recombination or convergent evolution, whereby the same mutation arose independently in different lineages (note however that independent SNP typing (Chapter 6) suggested that two of these SNPs, indicated in Table 2.9, were not actually homoplasic). If the latter explanation is true it would suggest the possibility of adaptive selection at these sites, which include nonsynonymous SNPs in two membrane proteins (STY1204 and *yadG*) and two nonsynonymous SNPs in *gyrA* that are known to increase resistance to fluoroquinolones, the class of antibiotics currently recommended for treatment of typhoid fever (2, 3, 397). Fifteen genes contained clusters of nonsynonymous SNPs, whereby two residues within five amino acids were mutated, which may indicate adaptive selection in localised regions of the encoded protein (Table 2.9).

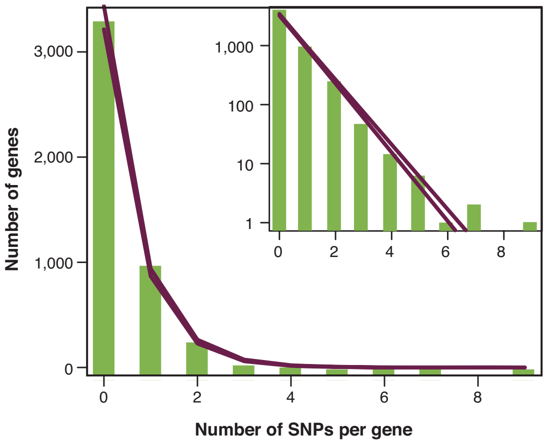


Figure 2.8: Distribution of number of SNPs per Typhi gene - Lines indicate 95% confidence interval of mean predicted values under a Poisson distribution fitted to the data shown in green. Inset shows gene count on a log scale to better show deviation from the Poisson model at high numbers of SNPs per gene.

Of the 26 genes exhibiting potential signals of adaptive selection, half encode proteins that are surface-exposed, exported or secreted, or affect synthesis of such proteins (highlighted in Table 2.9). These weak signals may reflect selective pressures stemming

Gene	SNPs	cluster	homoplasy	Name	Length	Function
STY2389	9*	465,470**	-	*yehU*	562	*response to stimulus*
STY2875	7*	-	-		3625	*membrane protein*
STY4656	7*	263,266	-	*tviE*	579	*Vi synthesis*
STY4318	6*	-	-	*bigA*	1870	*outer membrane protein*
STY2499	3	83,87	non x 2 (83,87)	*gyrA*	879	topoisomerase
STY1204	2	-	non (188)		403	*membrane transporter*
STY0194	1	-	non (37)	*yadG*	309	*membrane transporter*
(STY0347)	1	-	non (563)	*tsaC*	896	*fimbriae*
(STY1689)	3	-	syn (35)	*ydhD*	116	
STY3775	2	-	syn (418)	*priA*	733	DNA replication
STY1674	1	-	syn (79)	*pdxH*	219	metabolism
STY3838	0	-	44 bp upstream	*fdhD*	268	respiration
STY4805	2	-	186 bp upstream		407	metabolism
STY0042	2	10,11	-		498	*secreted protein*
STY0223	3	47,51	-	*hemL*	427	metabolism
STY0565	2	7,8**	-	*gcl*	594	metabolism
STY0970	3	30,31	-		66	
STY1264	2	58,59	-	*sifA*	337	*secreted effector*
STY1515	2	47,48	-		388	
STY2388	4	131,131**	-	*yehT*	240	*response to stimulus*
STY3222	2	9,12	-		212	*membrane protein*
STY3297	2	199,203	-	*ordL*	434	metabolism
STY4161	3	41,44	-	*yhjY*	235	*membrane protein*
STY4314	5	32,35	-	*gph*	84	DNA repair
STY4890	5	12,12**	-	*cstA*	717	*membrane transporter*
STY4659	4	-	-	*tviD*	832	*Vi synthesis*

Table 2.9: Genes with potential signals of adaptive selection - *=deviation from Poisson model of SNPs per gene; **=clustered nsSNPs are in the same isolates; ns=nonsynonymous, syn=synonymous. Functional group in italics indicates the gene is predicted to encode surface exposed or secreted protein, or to be required for synthesis of such proteins. Note that two apparent homoplasies were found by independent SNP typing to be non-homoplasic polymorphisms (see Chapter 6), the genes affected are shown in brackets.

from interactions with the human host (581), including selection for more virulent mutants or those with novel antigenic variants that better escape immunity in the human population. The genes identified here as potentially under selection warrant further investigation, illustrating the value of this approach which could potentially be adapted to genetic association studies in pathogenic bacteria, similar to those performed routinely in eukaryotes (582). However, most genes whose products are released by the bacterial cell or are surface-exposed showed no evidence of adaptive evolution. For example, with the exception of the SPI1 effector protein *sifA* (Table 2.9), no other known secreted effector proteins showed evidence of potential immune selection.

2.3.2.3 Recombination

Other than the few SNPs that do not fit the phylogenetic tree, which are potentially due to convergent evolution, there was no evidence of recombination between Typhi isolates and very little evidence of recombination with other bacteria (see 2.2.7). A 25 kbp import from Typhimurium was identified in Typhi isolate 404ty, however when investigated this was found to have been introduced artificially in the laboratory during the production of an *aroA* knock-out mutant. This region includes all the SNPs initially used to define 404ty as haplotype H2 rather than H59 in (2). Since the present study is concerned with “wild” Typhi variation, the SNPs in the imported region of 404ty were excluded from the phylogenetic analysis and 404ty was reassigned to haplotype H59. Many of the other SNP clusters identified using this approach were within phage sequences, which are likely due to misalignment, recombination between phage genes in the Typhi chromosome or recombination with novel phage. However 14 potential recombination events were detected in small stretches (50-270 bp) within non-phage genes, summarised in Table 2.10. Sequencing reads aligning to the CT18 sequence in these 14 potential recombined regions consistently included the SNPs and consequently do not represent a mixed population. Thus, the apparent variants reflect genuine differences between the sequenced DNA and the Typhi CT18 reference sequence rather than DNA contamination. The majority of these potential sequence imports (nine) were detected in isolate M223, all of which shared close similarity with *E. coli* sequences (Table 2.10). Large-scale recombination has been identified between Typhi and Paratyphi A (56). However this occurred before the evolution of the common ancestor of extant Typhi (see Chapter 4), which now appears to be genetically isolated.

Isolate	**Size**	**Gene**	**CT18**	***E. coli***	***S. enterica***	**Serovar**
404ty	50	STY4499	0.153	0.210	*0.129	T,P,E
E98-0664	150	STY2627a	0.046	N/A	*0.011	T
E98-2068	100	STY1289	0.145	0.204	*0.136	E
E00-7866	100	STY4499	0.109	*0.000	0.097	T,P
E00-7866	100	STY2853	0.040	*0.036	0.040	Typhi,T,P,C
M223	150	STY1428	0.164	*0.000	0.157	T,P,E,C
M223	230	STY1901	0.091	*0.004	0.091	Typhi,T,P,C
M223	200	STY2125	0.111	*0.016	0.105	T
M223	270	STY2546	0.056	*0.008	0.056	Typhi,P,E
M223	160	STY2768	0.123	*0.038	0.123	Typhi,P,C
M223	100	STY2970	0.071	*0.010	0.060	T
M223	60	STY3459	0.213	*0.000	0.118	T,E,C
M223	230	STY3907	0.040	*0.000	0.040	Typhi,T,P,C
M223	250	STY4250	0.166	*0.008	0.165	T

Table 2.10: Recombination events detected in Typhi isolates - Size gives estimated size of recombined region (bp). Divergence was measured between *de novo* assembled contig sequence for the Typhi isolate and homologous sequence from Typhi CT18, other *S. enterica* serovars and *E. coli* (for *E. coli* and *S. enterica* divergence is the minimum between the imported sequence and sequences in EMBL as of December 2007). For each potential imported sequence, the closest species is indicated with a *, and the closest *S. enterica* serovar is indicated in the last column (T=Typhimurium, P=Paratyphi A, E=Enteritidis, C=Choleraesuis).

2.3.3 Gene acquisition

Since 454 reads were long enough to be assembled, DNA insertion events could be identified among 454-sequenced Typhi isolates and confirmed by PCR and capillary sequencing (see 2.2.2). The distribution of insertions and deletions in the Typhi genome and among isolates is shown in Figure 2.9. Three *IS*1 insertions were previously identified in the CT18 genome. Comparative analysis found no evidence of insertions at these sites in the remaining 19 Typhi genomes, however an *IS*1 element was detected at another site in H58 isolates (Figure 2.9). These most likely originated from IncHI1 plasmids, which encode *IS*1 genes and were detected only in CT18 and some H58 strains (see 2.3.3.3 below).

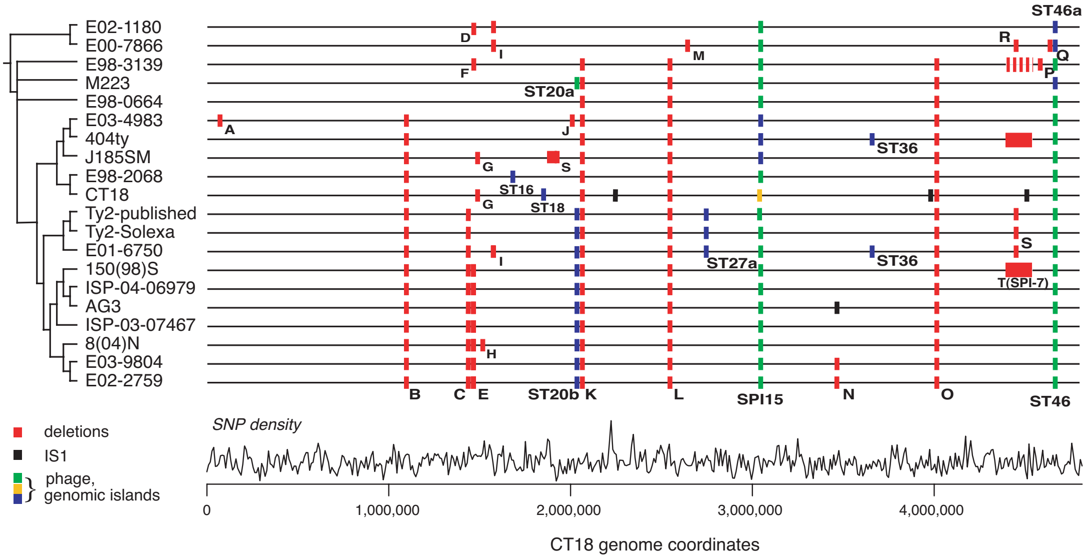


Figure 2.9: Distribution of prophage, IS elements and deletions in the Typhi genome and phylogenetic tree - The phylogenetic tree based on SNPs is shown on the left, the distribution of these SNPs in the genome is shown at the bottom. The genomic positions of deletions, prophage and IS insertions are shown using the colours indicated. Deletions are labelled the same as in Table 2.11.

2.3.3.1 Prophage sequences

CT18 harbours seven well defined prophage-like elements (Figure 1.4) (46, 224) and while some of these were conserved in all sequenced isolates, several novel phage were also identified. Figures 2.6 and 2.9 show the occurrence of phage insertion events in the phylogenetic tree, and the number of insertion events occurring in each relative age group is shown in Table 2.8. The complete ST18 phage of CT18 was not identified in any other genome, although the central region of this phage was present within the ST10 phage in all but CT18. It is therefore hypothesised that the other isolates carry the ancestral version of ST10, which recombined in CT18 with the recently acquired ST18 phage. The CT18 phage ST46 lies within SPI10 in most sequenced isolates, however a different phage ST46a appeared to be integrated at the same site (tRNA-*Leu*, a hot-spot for horizontal gene transfer (232)) in isolates E02-1180, E00-7866 and M223. This is consistent with the acquisition of one phage (ST46a) within SPI10 in a common ancestor of extant Typhi, followed by replacement of this phage at a point in the Typhi lineage shown in Figure 2.6. The ST2-27 phage previously identified in Ty2 (47, 224) was inserted at the same site in the closely related isolate E01-6750. A novel 28 kbp phage ST36, similar to the P2-like phage WPhi (583), was inserted at identical sites in the distantly related isolates 404ty and E01-6750. This highlights that prophage are not reliable markers of genetic relatedness among Typhi isolates. E98-2068 contained a novel 38 kbp phage ST16, similar to a Mu-like phage inserted in the uropathogenic *E. coli* strain UT189 (GenBank:NC_007946). A 20 kbp region identified in Ty2 (47) was conserved in the related E01-6750 and H58/H63 isolates, inserted within tRNA-*Asn* (ST20b). A region with similar sequence was also identified in M223, however the insertion site was a neighbouring copy of tRNA-*Asn* (ST20a).

2.3.3.2 Genomic islands

Variation was also identified within the 6 kbp genomic island SPI15 (155). This region includes an integrase gene adjacent to four hypothetical genes and was inserted within tRNA-*Gly*, generating direct flanking repeats. The region appeared to exist in three forms among the sequenced Typhi: (a) CT18; (b) J185SM, 404ty and E03-4983; (c) all other isolates (see Figure 2.9). In each case the insertion site and direct repeats were identical, but three distinct but related alleles were present for the integrase gene (95% amino acid identity between a and b, 70% between a, b and c). All three forms contained a probable phage regulatory gene with similarity to the Pfam protein domains Phage_pRha and PB091963. However each form contained a unique set of cargo genes, the function of which is unknown. The cargo genes encode proteins with matches to

Pfam B protein domains (PC023776, PB098004, PB017807, PB194640, PB127141) so far found only in other human pathogens including *Shigella flexneri*, *Yersinia enterocolitica*, *Erwinia carotovara subspecies atroseptica*, *Vibrio cholera*, *Leishmania major* and *Trypanosoma brucei*. These genes merit further investigation because of their potential contribution to virulence.

2.3.3.3 Plasmids

Plasmids were detected in seven of the sequenced Typhi isolates and fell into three types (Table 2.1, Figure 2.6). CT18 harbours two plasmids, pHCM1 and pHCM2 (46). The pHCM1 plasmid is of the IncHI1 incompatibility type, which is often associated with multiple drug resistance in Typhi (275, 276). IncHI1 multidrug resistance plasmids were also detected in three of the H58 isolates, however these were more closely related to pAKU_1, which was sequenced from an isolate of Paratyphi A (283) (99.7% sequence identity to pAKU_1, 98.4% to pHCM1, see Chapter 5). The pHCM2 plasmid, predicted to be associated with virulence (46, 278), was also identified in an H58 isolate (>99.9% sequence identity to pHCM2, 100% coverage of pHCM2 with no deletions). The presence of these plasmids in Typhi isolates of distantly related types (Figure 2.6) shows that independent acquisitions of similar plasmids have occurred in different Typhi isolates. In contrast the linear plasmid pBSSB1, which encodes the Typhi z66 flagella variant (255), was found only in the two H59 isolates (>99.99% sequence identity to pBSSB1), consistent with an earlier study (256).

2.3.4 Loss of gene function

2.3.4.1 Genomic deletions

Genomic insertions were rare in the sequenced isolates, but deletions were twice as common and more conserved (see Table 2.8). Note that in many comparative studies insertions and deletions are indistinguishable, but were able to be separated in this study using the rooted phylogenetic tree. All deletions were checked by capillary sequencing of amplicons generated by PCR reactions covering the deletion boundaries (see 2.2.2 above). The deletions range in size from 60-6,560 bp and some correspond to variant regions previously identified using DNA microarrays (511) (see Table 2.11). Most of the deleted regions include protein-coding sequences, resulting in partial or total deletion of 42 Typhi genes. The distribution of deletions is shown in Table 2.11 and illustrated in Figure 2.9.

ID	Type	Size	CT18 Position	Num. genes	Genes
A	dr (8bp)	418 bp	69831-70249	2	STY0068, STY0069
B	-	368 bp	1097535	1	STY1131
C	hp (5bp)	783 bp	1438314-1439052	2	STY1485, STY1486
D	-	1451 bp	1463356-1464807	1	STY1505 (glgX)
E	hp (6bp)	993 bp	1466586-1467578	3	STY1507, STY1508**, STY1509
F	-	180 bp	1468803-1468953	0	n/a
G	-	1260 bp	1490158	1	STY1536
H	-	6560 bp	1518586-1525146	8	STY1568-STY1575
I	-	1840 bp	1576079-1577919	3	STY1648-STY1650
J	-	241 bp	2011802-2012043	1	STY2167 (fliC)
K	dr (8bp)	155 bp	2067704	1	STY2238
L	dr (12bp)	102 bp	2550673	1	STY2717 (internal deletion)
M	-	75 bp	2647832-2647907	1	STY2791
N	-	720 bp	3470820-3471540	2	STY3617, STY3618
O	dr (12bp)	60 bp	4021842	1	STY4162 (internal deletion)
P	dr (8bp)	121 bp	4591367-4592273	3	STY4728**, STY4728a**, STY4729
Q	-	841 bp	4645486-4646324	2	STY4786, STY4787
R	-	5795 bp	4453436-4459232	8	STY4575-STY4582
S	-	1116 bp	4458116-4459232	2	STY4580, STY4582
T	-	133.5 kbp	4409500-4543100	149	STY4521-STY4680

ID	Strain(s) harbouring the deletion	Array ID
A	E03-4983	
B	CT18, E98-2068, J185SM, H59, Ty2, E01-6750, H58/H63	
C	Ty2, E01-6750, H58/H63	III
D	E02-1180	
E	H58/63	IV
F	E98-3139	IV
G	*CT18, J185SM	
H	8(04)N	
I	*E01-6750, E00-7866, E02-1180	
J	E03-4983	
K	E98-3139, M223, E98-0664, CT18, E98-2068, J185SM, H59, Ty2, E01-6750, H58/H63	
L	E98-3139, M223, E98-0664, CT18, E98-2068, J185SM, H59, Ty2, E01-6750, H58/H63	
M	E00-7866	
N	E02-2759, 8(04)N	
O	E98-3139, M223, E98-0664, CT18, E98-2068, J185SM, H59, Ty2, E01-6750, H58/H63	
P	E98-3139	
Q	E00-7866	
R	E00-7866	XI
S	Ty2, E01-6750	XII
T	*404ty, 150(98)S	X

Table 2.11: Genomic deletions detected in this study - Affected genes=genes partially or entirely deleted; dr=direct flanking repeats of 6-8 bp; *=deletion is not consistent with single event on phylogenetic tree; **=gene is a pseudogene in other isolates. Some deletions overlap with deleted regions detected by microarray (511), region labels defined in that study are given by Array ID.

In addition SPI7, which harbours genes required for synthesis of the polysaccharide Vi capsule (92) was missing from 404ty and 150(98)S. The isolate E98-3139 appeared to be a mixed population in regards to SPI7 as its coverage in both 454 and Solexa reads was approximately 25% that of genomic coverage (see Figure 2.10). Note that the low mapping coverage in this region is most likely due to deletion of SPI7 rather than replacement with a similar island, as deletion is known to occur during culture (517, 568) and no alternative island could be assembled from 454 reads. No other SPIs were deleted from the sequenced Typhi, indicating they are relatively stable in the genome (although we observed three variants of the 6 kbp SPI15 as described above).

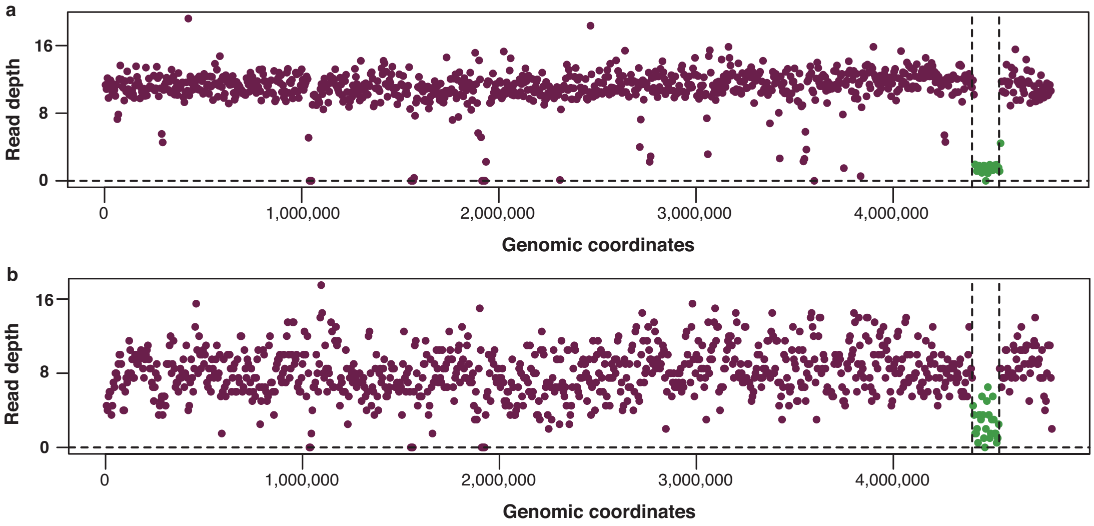


Figure 2.10: Coverage of SPI7 in Typhi isolate E98-3139 - Genome positions correspond to CT18 genomic coordinates, SPI7 is located between the two vertical dashed lines; coverage in this region is highlighted in green; horizontal dashed line shows zero coverage. (a) Read depth in 454 data. (b) Read depth in Solexa data.

2.3.4.2 Accumulation of pseudogenes

In addition to the 42 genes affected by deletion events, 55 nonsense SNPs were detected that had occurred since the last common ancestor of Typhi. These introduce stop codons into protein-coding genes, thereby cutting short translation. Read-through of stop codons has been reported (584), however the described mechanism applies to only two of the nonsense SNPs we detected. There was some evidence of selection against nonsense SNPs, with a lower rate of occurrence than nonsynonymous SNPs. Nevertheless, many nonsense SNPs were fixed, making up 2.9% of SNPs in the intermediate and oldest age groups (Table 2.8).

By mapping the deletions and nonsense SNPs to the phylogenetic tree we found that 92 novel pseudogenes have accumulated among the sequenced Typhi isolates since their last common ancestor (Appendix A), which itself harboured ∼180 pseudogenes (46, 47). Many of these genes fall into gene categories (metabolism, cobalamin utilisation, peptide or sugar transport, fimbriae) previously associated with pseudogenes in host-restricted pathogens (266) (see Appendix A). Figure 2.11 shows the rate of accumulation of inactivating mutations in each branch of the phylogenetic tree. Nearly all of these genes showed evidence of expression in Typhi according to microarray data accessible at the NCBI GEO database (see 2.2.8, Appendix A), thus most of the nonsense and deletion mutations observed in this study probably result in inactivation of previously functional genes. Since the losses have occurred independently in different lineages, Typhi isolates at different points in the phylogenetic tree have slightly varying complements of functional genes, which may affect their pathogenic potential. This may contribute to the differences observed in clinical manifestations of typhoid fever in different regions (3). It is interesting to note that different lineages display variation in the relative rates of accumulation of SNPs and inactivating mutations (line slopes in Figure 2.11). This may be due to variation in mutation rates, or different selective pressures for or against pseudogene formation, in particular lineages.

Since only 3% of possible SNPs in the Typhi genome are nonsense SNPs, we expect only 1-2 false nonsense SNP calls overall (3% of the estimated total of 53 false SNP calls). This constitutes ∼2% of genes inactivated by nonsense or deletion mutation, which would make little difference to conclusions regarding the continuous accumulation of pseudogenes. In addition, frameshift mutations were not analysed in this study since single base insertions or deletions were difficult to detect reliably from 454 and Solexa sequence data (0.6% indel error rate for 454 data (555); see 3.3.1.4 and Figure 3.4 for analysis of feasibility for Solexa data). However most of the genes identified as differentially inactivated between CT18 and Ty2 were due to frameshift mutations (20 frameshifts vs 4 nonsense SNPs and 2 deletions) (46, 47), thus it is likely that many more pseudogenes may have accumulated in the Typhi population than those caused by nonsense SNPs or deletions. Therefore, while the current analysis demonstrates that gene loss is ongoing in Typhi, the extent of this phenomenon is likely underestimated.

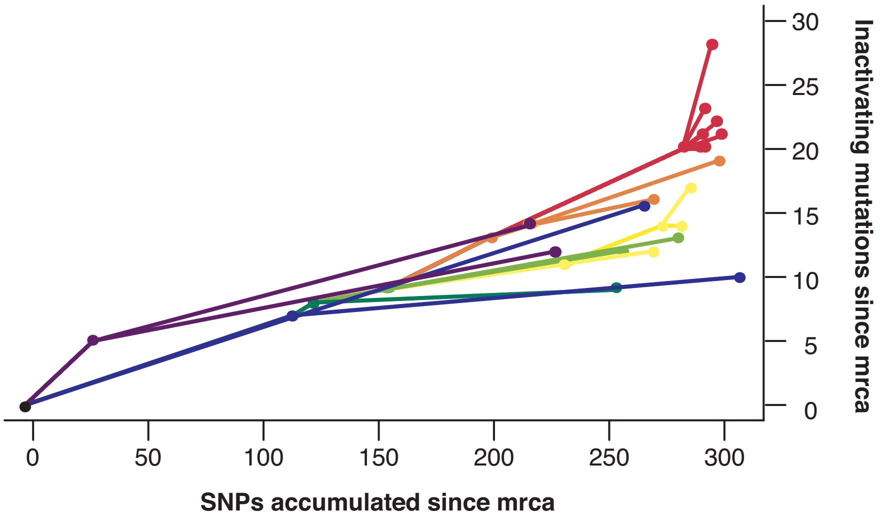


Figure 2.11: Accumulation of gene-inactivating mutations in Typhi lineages - Points correspond to bifurcations in the phylogenetic tree in Figure 2.6, y-axis shows the total number of genes inactivated by deletion or nonsense mutation up to that bifurcation. Each line represents the accumulation of mutations in a particular isolate since the most recent common ancestor (mrca) of all 19 genomes, branches are coloured as in Figure 2.6.

2.4 Discussion

2.4.1 Strengths and limitations of the study

This study employed novel high-throughput sequencing technologies to compare whole genome sequences from 19 Typhi isolates. At the time of writing, few whole-genome intraspecies comparisons of this scale existed for bacteria (515, 585) and none at this level of sub-species resolution. Some experimentation was therefore required to determine the best methods for analysis, particularly for SNP detection as this is most important for phylogenetic inference, but also most dependent on absolute accuracy of sequence alignments.

For this study, the analysis of 454 data was of central concern, as this platform was used to sequence representative isolates from central and radial haplotype nodes thus forming the basis of phylogenetic analysis, as well as providing evidence of insertion and deletion events via assembled sequence. Analysis of simulated and experimental Typhi data suggested that analysis of assembled contigs provided more accurate SNP detection than direct analysis of 454 reads (Tables 2.3, 2.4; 2.3.1.1). This is likely due to the fact that the 454 *de novo* assembler Newbler 1.0.5.3 operates in flow space, meaning it assembles raw pyrosequencing signal data directly into consensus base calls as opposed to converting signals to base calls in individual reads and assembling the resulting read sequences. Thus a base call in the assembled contigs represents the consensus of multiple flow signals, which is expected to be more accurate than the collection of base calls in individual reads, and so contigs assembled in this way are expected to be more reliable for correctly identifying SNPs (Guido Kopal, Roche, personal communication, February 2007). Note however that both 454 data generation and the Newbler assembly software have changed since this study was completed, and so analysis of new data will require a reassessment of SNP detection approaches that take into account the platform and software in use at the time.

The false positive rate for SNP detection in this study was estimated to be around 2-3% of detected SNPs (Table 2.6). This seems high expressed as a percentage, but it reflects more on the paucity of true SNPs within the Typhi population rather than a high error rate in sequencing. Even at an upper estimate of 10 false SNPs per genome (2.3.1.4), this would be only ∼1 error per 500,000 bp of sequence, which is considered an acceptable error rate even for finished genome sequences. The error rate could be determined more accurately by independently testing each SNP locus. To do this by Sanger capillary sequencing would make little sense, as it would be extraordinarly labour intensive: to confirm or disprove a SNP call from ∼10 Solexa or 454 reads would require at least the same depth of sequencing, i.e. generating ∼20,000 targeted sequencing reads to confirm ∼2,000 SNPs, or ∼2,000 experiments to check just 10% of SNP calls. Fortunately, over 75% of SNPs identified in this study have been successfuly typed using a high throughput genotyping system (see Chapter 6), which provided independent confirmation of >98.5% of SNPs tested. Thus we can say with some confidence that the false positive rate of SNP detection was quite low in this study, and should not affect the conclusions regarding phylogenetic structure and overall patterns of nucleotide substitutions.

An independent test of false negative rates is much more difficult. Data simulation suggested this should be below 10% at the read depths used in this study (2.3), although this did not take into account the full effects of quality filtering (2.3.1.3). However estimations made by comparing results from 454 and Solexa data suggest that using quality filtering, in combination with checking alleles in all isolates at all SNP loci detected in any isolate (2.3.1.5), should result in detection of $\geq$90% of SNPs. Failing to detect all SNPs is unlikely to have an affect on phylogenetic inference, particularly given the allele checking procedures (2.3.1.5). It may reduce power to detect genes under selection, although there is no obvious reason to suspect that undetected SNPs should be concentrated in particular regions of the genome which would affect conclusions regarding particular genes. Clusters of SNP calls made in one strain were manually inspected at both the read and contig level to separate true SNP clusters from errors. This approach avoided the inclusion of 198 dubious SNP calls while allowing several real SNP clusters to be considered in the analysis (2.3.1.5). The exclusion of repetitive sequences from SNP analysis (2.3.1.5) will undoubtedly blind us to some genuine variation in the Typhi genome. However this is essential to avoid a large amount of alignment and assembly errors (2.3.1.5), and all estimates of rates and patterns of variation ($\frac{dN}{/}dS$, transition bias, G+C content, etc) were adjusted to reflect the exclusion of repetitive sequences. Only 20 genes were excluded that were not phage or IS elements (2.7) and manual inspection of SNP calls in these gene sequences revealed no evidence of variation that could not be explained by mis-alignment of the repeated sequences. Thus while it must be accepted that a small number of SNPs will have gone undetected in this study, there is no evidence to suggest that this results in systematic bias that would invalidate the conclusions drawn.

Genomic insertions and deletions identified from 454 contigs, and deletions identified from Solexa reads, were all confirmed by PCR and Sanger capillary sequencing (2.2.2, Table 2.11). In each case, capillary sequencing confirmed the exact boundaries of the insertion/deletion events that were detected from 454 or Solexa sequence data. Thus 454 sequencing can be used to characterise insertion and deletion events, and short read Solexa data can be relied upon to detect deletion boundaries. Plasmid detection was also highly successful using 454 or Solexa sequencing of genomic DNA. The presence of plasmids in each strain had been inferred previously by resistance testing (suggestive of IncHI1 plasmids) and z66 screening (suggestive of plasmid pBSSB1), and was confirmed by plasmid isolation experiments performed by Stephen Baker at the Sanger

Institute (2.2.3). All expected plasmids were successfully detected from sequence data and all isolates in which plasmid sequence was detected tested positive for plasmids of the expected size by plasmid isolation. Thus 454 and Solexa sequencing of genomic DNA were both highly successful at detecting plasmid sequences. In this case plasmid isolation confirmed that all plasmids sequenced were present as independent replicons within the Typhi isolate, however it possible for plasmid sequences to be integrated into the genome. This may be difficult to detect from sequence data, although successful identification of phage insertion sites from 454 assemblies suggests this may be possible.

A clear limitation of the present study is the inability to resolve small indel events involving one or a few bases. This is particularly difficult for the 454 platform, which has a tendency to make errors in homopolymeric tracts. This is because during pyrosequencing, the number of nucleotides incorporated in a single flow is estimated from the amount of fluorescence emitted, which becomes less precise with higher numbers of nucleotides. For example, the difference in fluorescence emitted when one vs two nucleotides is incorporated is relatively easy to discern, but the difference in flourescence emitted upon incorporation of five vs six nucleotides is much harder. This is particularly unfortunate because the natural bacterial DNA replication machinery makes precisely the same sorts of errors (586), making it impossible to distinguish sequencing errors of this kind from genuine mutations. Unfortunately, on the 454 GS20 platform used here, the problem was known to extend to subsequent flows, so that base calls made after a homopolymeric tract could sometimes be wrong too. This problem has been greatly reduced in newer versions of the 454 platform and analysis software, but in data from the present study, single base insertions and deletions cannot be trusted. This is illustrated by the fact that MUMmer identified >15,000 indels in non-repetitive regions of each set of 454 contigs, compared to <30 indels detected between CT18 and Ty2 finished sequences. The failure to identify small indels almost certainly results in an underestimate of the rate of accumulation of pseudogenes in the Typhi genome. As stated above (2.3.4.2), comparison between CT18 and Ty2 revealed indels resulting in frameshifts in 20 genes, more than three times the number of genes inactivated by nonsense SNPs or deletions between the two genomes (47). It is also possible that being blind to small indels results in an underestimate of the level of variation within some genes, potentially obscuring signals of antigenic variation. While there is little to be done about this using current data and software, this problem will become increasingly tractable using Solexa sequencing either to identify and correct errors in 454 data, or to detect small indels directly by alignment of reads to reference (587).

2.4.2 Differences between Typhi lineages

The phylogenetic tree shown in Figure 2.6 defines 12 distinct Typhi lineages. These isolates include five chosen from central nodes of a minimum spanning tree resulting from analysis of over 100 isolates (Figure 2.1), so it is likely that the internal branches of the phylogenetic tree capture a significant proportion of the common evolutionary history of the Typhi population. Thus mutations lying on these internal branches, including SNPs, deletions and the insertions of prophages ST46 and ST20b, are likely to be informative markers for discriminating within the broader Typhi population. It is important to note that isolates from internal nodes (haplotypes) of the minimum spanning tree (Figure 2.1) are not in any sense 'ancestral' to those from radial nodes, despite the impression given by the appearance of the minimum spanning tree. For example, while H50 appears to have diversified into multiple haplotypes including H8 and H52 (Figure 2.1), the common ancestor of isolates E98-3139 (H50), E98-0664 (H52) and M223 (H8) is no closer to the H50 isolate E98-3139 (Figure 2.6). It was simply by chance that the 1.85% of the genome analysed by Roumagnac *et al* (2) happened to include regions that differentiate E98-0664 and M223, but not E98-3139, from their common ancestor. Because it is based on genome-wide data, the phylogenetic tree generated in this study more accurately reflects the fact that all lineages of Typhi have continued to diversify at equivalent rates, as the total branch lengths from any isolate back to the root are roughly equal (Figure 2.6).

It is difficult to estimate how much of the underlying variation in the Typhi population has been captured in this sequencing study. However we do now have a much better picture of how much variation can be expected between two distinct Typhi lineages. Figure 2.12 shows the distribution of the number of SNPs detected between pairs of the 12 Typhi lineages shown in Figure 2.6 (using 404ty to represent the 404ty/E03-4983 lineage and AG3 to represent the H58 lineage). There are peaks at ∼300 SNPs and ∼550 SNPs, reflecting the very early divergence of E00-7866 (H46)/E02-1180 (H45) from the other lineages. Note roughly half of SNPs are nonsynonymous, thus lineages differed by an average of 150 nonsynonymous SNPs. Any two lineages differed by an average of 5 deletions (range 0-15) and 1 or 2 prophage (range 0-5), the distributions are shown in Figure 2.13a-b. However prophage insertions tended to be specific to individual lineages, while deletions were more frequently conserved (see Figures 2.6 and 2.9, Table 2.11). Thus the deletions identified in this study would make good genetic markers whereas the prophage insertions would not. There is likely to be some functional variation between lineages due to nonsynonymous SNPs (in particular nonsense SNP)s, deletions and small indels. Lineages differed by an average of 150 nonsynonymous SNPs and four pseudogenes differentially inactivated by nonsense SNPs and deletions (Figure 2.13). Note that the true difference in pseudogene complement is likely to be larger, due to frameshifts which could not be detected from this data; more than 75% of pseudogenes that differed between CT18 and Ty2 were due to frameshifts, thus it is likely that the mean variation between lineages is more in the order of 12 pseudogenes than four. The effect of these mutations is unknown, however it is likely that they do contribute to some phenotypic differences. It is also possible that prophage variation contributes to differences in genetic function between Typhi lineages, although no obvious phage cargo genes were identified in this study. The variation observed in SPI15 may also contribute to phenotypic variation between lineages, although the function of its cargo remains a mystery.

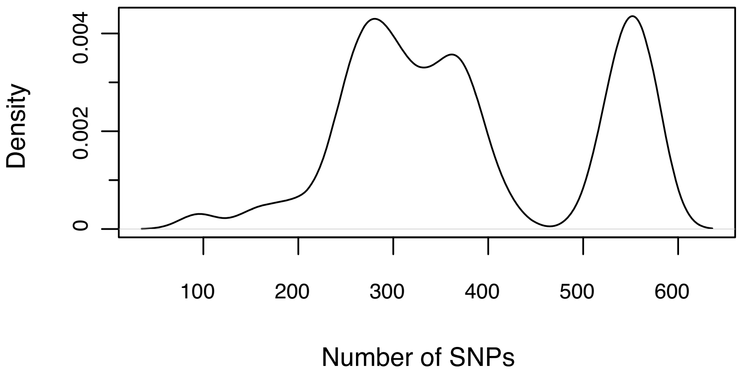


Figure 2.12: Distribution of number of SNPs between pairs of Typhi lineages - The number of SNPs between every possible pair of 12 Typhi lineages was calculated (using AG3 to represent the H58 lineage and 404ty to represent the H59 lineage), the distribution of SNP numbers between pairs is shown.

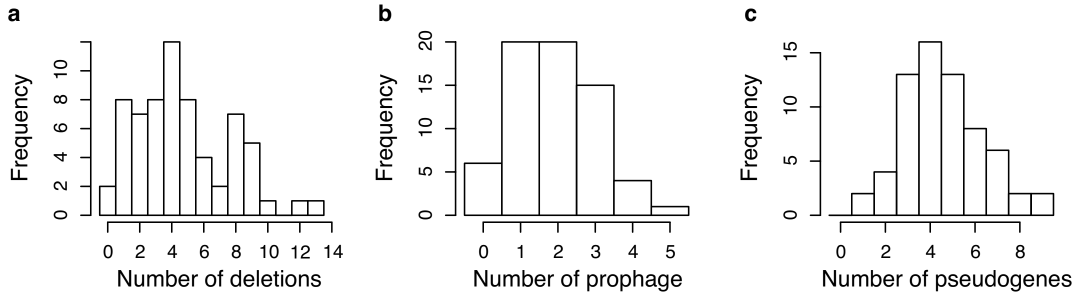


Figure 2.13: Distribution of number of deletions, prophage and pseudogenes between pairs of Typhi lineages - The number of deletions, prophage insertions and pseudogenes that differed between every possible pair of 12 Typhi lineages was calculated (using AG3 to represent the H58 lineage and 404ty to represent the H59 lineage). The distribution of these counts is shown for deletions, prophage and pseudogenes in a-c respectively.

2.4.2.1 Antibiotic resistance and the H58 lineage

Four of the sequenced isolates were multidrug resistant (MDR) and harboured IncHI1 MDR plasmids (2.3.3.3). These include the sequenced MDR strain CT18 (H1) and three H58 isolates (E98-9804, ISP-03-07467, ISP-04-06979). Five of the sequenced isolates were resistant to nalidixic acid (Nal), a marker for resistance to fluoroquinolones which are currently recommended for the treatment of typhoid fever. These all contained SNPs in *gyrA*, see Table 2.12. Roumagnac *et al.* reported a much higher frequency of nalidixic acid resistance among H58-derived strains compared to other haplotypes (2). No resistance plasmids were identified among the other sequenced isolates, thus antibiotic resistance of all kinds appears to be over-represented in the H58 haplotype background. Roumagnac *et al.* also showed this haplotype has experienced a recent proliferation, especially in South East Asia, and the concentration of resistance in this group may provide a mechanism for recent selection via the treatment of human infections (2). Different *gyrA* mutations contribute to Nal resistance in H58 strains although some remain Nal sensitive (Table 2.12, (2)). Thus, while the H58-defining SNPs do not themselves confer Nal resistance, H58 may provide a genetic background whereby *gyrA* mutants can survive more easily in the population. An alternative hypothesis is that the selective advantage of the MDR plasmid (and potentially H58-specific chromosomal mutations) may have resulted in an early proliferation of MDR H58 such that, when fluoroquinolones were introduced in response to rising rates of MDR typhoid, MDR

H58 strains were more frequently exposed to the novel drugs compared to other lineages which were still sensitive to the old drugs. This scenario would result in increased selective pressure for fluoroquinolone resistance in H58 over other lineages, and may account for both the higher frequency of nalidixic acid resistance and the variety of mutations conferring this resistance.

Isolate	Haplotype	GyrA	Nal	IncHI1 plasmid	MDR
E00-7866	H46	wt	S	no	S
E02-1180	H45	Gly87	R	no	S
M223	H8	wt	S	no	S
E98-3139	H50	Phe83	R	no	S
E98-0664	H55	wt	S	no	S
E98-2068	H42	wt	S	no	S
CT18	H1	wt	S	yes	MDR
J185SM	H85	wt	S	no	S
404ty	H59	wt	S	no	S
E03-4983	H59	wt	S	no	S
E01-6750	H52	wt	S	no	S
Ty2	H10	wt	S	no	S
AG3	H58	wt	S	no	S
150(98)S	H63	Phe83	R	no	S
8(04)N	H58	Gly87	R	no	S
E02-2759	H58	wt	S	no	S
E03-9804	H58	Phe83	R	yes	MDR
ISP-03-07467	H58	wt	S	yes	MDR
ISP-04-06979	H58	Phe83	R	yes	MDR

Table 2.12: Drug resistance phenotypes and genetic variants for sequenced Typhi isolates - S=sensititive, R=resistant, MDR=multidrug resistant.

In this study 106 H58-specific SNPs were detected, including 57 nonsynonymous substitutions. Three of these introduce stop codons within genes encoding lipoprotein B precursor *rlpB* (STY0698), DNA-binding protein *stpA* (STY3001) and ATP-dependent RNA helicase *dbpA* (STY1410). The *stpA* gene also contains a SNP introducing a novel stop codon in E98-3139, which also harbours the GyrA-Phe83 mutation. StpA is a homolog of the transcriptional repressor H-NS involved in repressing transcription of the porin gene *ompS1* and possibly many other Typhi genes (588, 589). The occurrence of independent inactivating mutations in *stpA*, on different haplotype backgrounds, suggests that this gene may be subject to negative selection. One hypothesis is that inactivation of *stpA* enables cells to better tolerate the mutations in GyrA, perhaps by the suppression of some response normally induced by StpA. This could be investigated by looking for evidence of an association between *stpA* inactivation and the GyrA-Phe83 mutation on different haplotype backgrounds and by selection on media containing nalidixic acid. In addition, the H58 isolates shared a deletion affecting the aminotransferase gene STY1507 and hypothetical gene STY1509, and they shared a SNP within the RNA gene *csrB* which regulates activity of the carbon storage regulator *csrA* (STY2947) (590). CsrA regulates the expression of SPI1 genes and genes secreted via the SPI1-encoded TTSS in Typhimurium (591), so the mutation in *csrB* may have an effect on virulence.

2.4.3 Insights into the evolution of Typhi

2.4.3.1 Adaptive selection in Typhi genes

The very low level of nucleotide variation detected between Typhi genomes makes it difficult to conclude much about selection on individual Typhi genes. Most genes (72%) contained no SNPs at all, although it is possible that some of these may harbour frameshift or other small indel mutations that could not be detected. The usual approach of detecting selection by calculating $\frac{dN}{dS}$ for a particular gene would be inappropriate in this study, as there is not enough variation to work with and there is some doubt as to whether the statistic is useful for analysing variation within rather than between species (525) (see 2.4.3.2 below). However the variation data generated in this study were carefully examined for evidence of unusual variation within particular genes which may indicate adaptive selection, and very little was found (Table 2.9).

The lack of evidence for adaptive selection in general is in contrast with the known adaptive selection for mutations in *gyrA* associated with fluoroquinolone resistance. The signal of selection in *gyrA* was detected in the present study as clustered, homoplasic nonsynonymous SNPs in neighbouring codons 83 and 87. Two other genes contained homoplasic nonsynonymous SNPs (Table 2.9), one of which (*yadG*) is the membrane component of an efflux protein in *E. coli* (592) and may therefore be associated with antibiotic resistance in Typhi (efflux proteins can act as pumps to remove antibiotics from the bacterial cell (593)). However, no genes besides *gyrA* contained multiple homoplasic SNPs and few contain multiple nonsynonymous SNPs at all, consistent with the hypothesis of genetic drift in the Typhi genome. The adaptive mutations evident in the *gyrA* gene highlight the strong selective pressure on the Typhi genome associated with antibiotic use in the human population. This is not particularly surprising, as the fitness advantage associated with increased antibiotic resistance is likely to be very strong. However the lack of similar evidence for other adaptive mutations suggests that Typhi is under relatively little selective pressure from its host or the environment in general.

The limited evidence of selection in Typhi gene sequences is particularly striking when compared to patterns observed among other human bacterial pathogens, which display a variety of mechanisms for antigenic variation. For example, antigenic variation is achieved by extensive recombination in the *Helicobacter pylori* and *Chlamydia trachomatis* populations (594, 595), while in *Mycobacterium tuberculosis* antigenic variation is associated with duplication and diversification of antigen-associated gene families (596). In contrast, only three Typhi genes contained more than six SNPs and just sixteen genes contained independent nonsynonymous SNPs in the same or neighbouring amino acids (see Table 2.9). While these may represent cases of antigenic variation, the level of variation is low, with most of the SNPs unique to a single haplotype and therefore most haplotypes sharing identical sequences. Similarly, while there was some evidence of import of small fragments of non-Typhi sequences (see Table 2.10), the only indication of possible recombination between Typhi isolates were eight SNPs that do not fit the phylogenetic tree (Table 2.9), which could equally be due to convergent evolution. The sparsity of direct sequence evidence for antigenic variation in Typhi suggests that this pathogen is not under strong selective pressures from the human immune system. Clearly that immune system has some ability to recognise and protect against Typhi infection, as whole cell and Vi vaccines do provide protection, although it is incomplete (estimated at around 50-60% protection, see 1.2.6). It is possible that Typhi has a different strategy for immune evasion, perhaps related to its inhabiting priveleged intracellular niches. However, it cannot be ruled out that Typhi may posses as yet unidentified mechanisms of generating antigenic diversity or that prophage genes, which were excluded from SNP analysis in this study, may play a role.

2.4.3.2 Evolutionary dynamics of the Typhi population

Kryazhimskiy and Plotkin recently suggested that $\frac{dN}{dS}$ is inappropriate for analysis of variation within a population (525), based on models that incorporate extensive recombination and high mutation rates resulting in a level of variability beyond that observed in Typhi. Although the models do not apply particularly well to the Typhi population, it is clear from Kryazhimskiy and Plotkin's study that care must be taken when applying a statistic designed for analysis of interspecies variation to analyse intraspecies variation. The problem with using $\frac{dN}{dS}$ to analyse intraspecies variation lies in the difference between substitutions (which have become fixed in a population and are therefore meaningful for comparing a particular sequence between populations), and mutations (which are not fixed in the population or within subpopulations, are subject to high rates of flux due to recombination and selective sweeps, and so should not be used to assess selection on a particular sequence within a population). However the use of $\frac{dN}{dS}$ in the present study (2.3.2.1) is not an attempt to assess selective pressures acting on particular sequences, but to assess the extent to which nonsynonymous mutations are conserved or removed from the Typhi population. The $\frac{dN}{dS}$ data are interpreted in the context of models of mutations within bacterial populations (526) and take into account the phylogenetic structure of, and lack of evidence for recombination within, the Typhi population.

The patterns of $\frac{dN}{dS}$ shown in Figure 2.7 suggest that, genome-wide, there is some degree of purifying selection in the Typhi population: nonsynonymous mutations appear to arise with the same frequency as synonymous mutations ($\frac{dN}{dS} \sim 1$ among recent intra-haplotype SNPs) but are less frequently conserved ($\frac{dN}{dS} \sim \frac{1}{2}$ among SNPs on internal or haplotype-specific branches). The lack of difference in $\frac{dN}{dS}$ between SNPs on internal or haplotype-specific branches suggests that this purifying effect happens relatively quickly but is not an ongoing process, which is consistent with a small population characterised by genetic isolation and drift rather than recombination and selective sweeps resulting in clonal replacement (526). This picture of the Typhi population is consistent with the lack of evidence for recombination within the population (2.3.2.3), the small estimated population size (2) and the persistence of distinct haplotypes through time and geographical space (all nodes of the Typhi phylogenetic tree shown in Figure 2.1 were detected among a set of less than 500 extant Typhi isolates (2), indicating that the Typhi population is not shaped by clonal replacement).

It has long been suspected that human carriers provide the main reservoir driving the transmission of Typhi (309, 597). Carriage occurs in the gall bladder, which facilitates shedding of Typhi into the environment enabling fecal-oral transmission to new hosts. Typhi is relatively difficult to isolate from water and the environment even in endemic regions (598, 599) and it is generally believed that the bacterium has a limited survival time outside the human host (600). If human carriers provide the main persistent reservoir for Typhi, this could account for the patterns of genetic drift and lack of recombination or gene acquisition detected in the present study, as the human reservoir is likely to be small and physiologically isolated (i.e. divided into distinct populations within the gall bladders of individual carriers, which are isolated from each other as well as from other enteric bacteria) (309, 597). Furthermore, adaptive mutations arising during symptomatic typhoid infections, may have no fitness advantage in the carrier state and may therefore be short lived in the long-term Typhi population. This sort of scenario has been described as the source-sink model of evolutionary dynamics, which distinguishes permanent "source" (Typhi carrier) and transient "sink" (typhoid patient) populations (601). The model predicts that adaptive mutations arising in the sink may be short lived in the population if they provide no fitness benefit in the long-term source (i.e. carriers), and thus clonal replacement does not occur. This model provides a plausible explanation for the patterns of variation evident within the Typhi population, including the absence of clonal replacement and general lack of evidence for adaptive selection assuming Typhi is well suited to carriage in the gall bladder. It also suggests that attention should be paid to treating and preventing Typhi carriage in addition to the relatively simpler task of treating typhoid fever. A key implication is that vaccination programmes are likely to be a highly effective strategy for long-term disease control in endemic areas, as they would not only reduce the number of typhoid infections but also the number of asymptomatic carriers of Typhi, thereby achieving a direct reduction in the size of the reservoir.

Chapter 3

Genomic sequence variation in Paratyphi A

3.1 Introduction

A single Paratyphi A isolate was sequenced at Washington University in 2004 (EMBL: CP000026) (49). The isolate, known as SARB42 or ATCC9150 is of unknown origin but is part of the SARB collection of *Salmonella* reference isolates assembled in 1993 (602) and has been used for many years as a laboratory strain. Very little analysis of genomic variation in Paratyphi A has been reported. The ATCC9150 genome sequence was used to design a microarray to screen for variation in an additional 12 Paratyphi A isolates (49). Variation was detected in the three prophage sequences harboured in the ATCC9150 genome, and in just two other regions (deletions of *hyaCDE*, or *cobB* and *ycfX*) (49).

There have been very few studies published reporting typing of Paratyphi A isolates. Those that have been reported utilised phage typing, PFGE, *IS*200 typing or ribotyping and revealed little variation among isolates (less variation than similar techniques can detect among Typhi isolates) (8, 477, 478, 479, 480). Variability has been detected among Paratyphi A strains at three VNTR loci identified in *Salmonella* Typhimurium (481), however to date (May 2009) no studies have reported using VNTR to type Paratyphi A isolates. Similarly, no studies have reported using MLST to analyse Paratyphi A isolates, however all Paratyphi A isolates recorded in the *S. enterica*

MLST database (464) so far were of the same sequence type, so this is unlikely to be a useful technique for analysing Paratyphi A populations.

All available data suggests that there is even less variation within the Paratyphi A population than within the Typhi population. The development of sequenced-based typing schemes, phylogenetic analysis and evolutionary analysis will therefore require whole genome sequence data. A second Paratyphi A genome AKU_12601, isolated from a paratyphoid patient in Karachi, Pakistan in 2004, was recently sequenced and finished at the Sanger Institute. In this chapter, the novel AKU_12601 sequence was compared to that of ATCC9150 in order to characterise all nucleotide variation between the two genomes. An additional five isolates were sequenced using the Solexa platform, allowing the construction of a sequence-based phylogenetic tree for Paratyphi A. Finally, a novel approach was developed and applied to screen for SNPs among a global collection of 160 Paratyphi A isolates. As well as providing evolutionary insights, these novel SNPs will provide the basis for development of sequence-based typing methods in the future.

3.1.1 Aims

The general aim of the work presented in this chapter was to characterise whole-genome variation in Paratyphi A. Since there was no predetermined phylogenetic structure to build upon, one of the aims was to develop a method for screening large collections of isolates for SNPs at the whole genome level. Specific aims of the analysis were to:

- define a phylogenetic tree based on seven Paratyphi A genomes;
- determine the quality and quantity of genetic differences within the Paratyphi A population;
- gain insights into the evolution of Paratyphi A, including the nature and frequency of genetic changes and any evidence of selective pressures upon individual genes; and
- identify SNPs that may be used to develop sequence-based typing methods.

3.2 Methods

3.2.1 Identification of repetitive and horizontally transferred sequences in the Paratyphi A genome

A list was assembled of all features annotated in either of the finished Paratyphi A genomes AKU_12601 or ATCC9150 as '/repeat_unit', '/repeat_region', or with the keywords 'phage', 'transposase', or 'IS'. The start and end coordinates of these features in the AKU_12601 reference genome were recorded in a table of regions to be excluded from SNP analysis. This analysis was done using Artemis; the same analysis was performed independently by Camila Mazzoni (Environmental Research Insititute, Cork, Ireland) using alternative software (Kodon, Bionumerics) and produced the same result. All bases found by Maq to be non-unique during short read mapping were also excluded. This essentially identifies any 35 bp sequences that occur more than once in the reference genome, with a maximum mismatch of 2 bp, which is particularly important as these are precisely the source of mapping and subsequent SNP calling errors using Maq.

3.2.2 SNP detection

Maq (564) was used to align 35 bp reads to the finished AKU_12601 sequence, using cut-offs determined in 2.3.1.3. For the comparison of short read data from AKU_12601 to the finished AKU_12601 sequence, capillary traces were manually inspected for the five loci at which SNPs were reported by Maq with consensus base quality $\geq$20 and read depth $\geq$5. MUMmer was used as described in 2.3.1.3 to detect SNPs in the finished sequence of ATCC9150 and 454 contig sequences from 6911 and 6912, using AKU_12601 as the reference sequence. SNPs detected among the seven genome sequences were merged and alleles checked as described in 2.3.1.5; alleles were checked in the same manner in Typhi CT18 and the genomes of other *S. enterica* serovars for use as outgroups.

3.2.3 Phylogenetic network analysis

Phylogenetic networks, or more specifically split networks, were generated using SplitsTree4 (603). A split network is a combinatorial generalisation of phylogenetic trees, designed to represent incompatibilities within the data set, which may arise through recombination, horizontal gene transfer, gene duplication/loss, etc. Parallel edges, rather than single branches, are used to represent the splits computed from the data. The length of an edge in the network is proportional to the weight of the associated split, analogous to the length of a branch in a phylogenetic tree. In this study, split networks were constructed directly from character data (as opposed to a distance matrix) using the parsimony splits method implemented in SplitsTree4.

3.2.4 Detection of insertion/deletion events and plasmid sequences

De novo assemblies were generated for each isolate using Newbler (454 data) (Roche) or Velvet (Solexa data) (567). Assembled contigs were ordered against the AKU_12601 genome using MUMmer (`nucmer` algorithm, (576)). Pairwise whole-genome sequence comparisons were generated with BLASTN and visualized using ACT (604). Contigs that did not map to the AKU_12601 or ATCC9150 genomes were analysed individually, using BLASTN to identify the sequences by comparison to the EMBL nucleotide sequence database. All such contigs matched phage or pGY1 plasmid sequences in the database. Insertions and deletions (indels) between the collinear Paratyphi A AKU_12601 and ATCC9150 genomes were identified using diffseq (part of the EMBOSS package (577)). These loci were checked in the remaining five genomes using either (i) alignments of Solexa reads visualised using Maqview (http://maq.sourceforge.net) to check indels of 1-20 bp or (ii) alignments of Solexa or 454 contigs visualised using Artemis to check larger indel events. The presence of plasmids pAKU_1 and pGY1 was assessed by (i) mapping of Solexa reads to the plasmid sequences using Maq, and (ii) BLASTN searches of the plasmid sequences within contigs assembled from 454 or Solexa data. Were novel plasmids present, they should have been identified during BLASTN searches of the EMBL database with unmapped contigs as described above (either by matches to plasmid sequences in the database or by failing to identify highly similar matches within the database).

3.2.5 Gene ontology analysis

A gene ontology annotation of the AKU_12601 genome was downloaded from EBI (http://www.ebi.ac.uk/GOA/; note this annotation was generated automatically using evidence from InterPro protein domains and did not include manual curation or experimental evidence). Lists of AKU_12601 genes were analysed using GOstat (605) (http://gostat.wehi.edu.au) to identify gene ontology terms that were statistically over-represented in the list as compared to the genome as a whole (using Benjamini and Hochberg correction to correct for multiple testing).

3.2.6 Accession codes

The AKU_12601 genome sequence and annotation, including all pseudogenes identified during comparative analysis in this study, is availabe in EMBL at accession FM200053. The genomes used for comparative analysis were Typhi strain CT18 (AL51338), Typhi strain Ty2 (AE014613), Typhimurium strain LT2 (AE006468) and Paratyphi A strain ATCC9150 (CP000026). Solexa data generated in this study is available in the European Short Read Archive at accession ERA000012 (AKU_12601 reads) and accession ERA000083 (six other single genomes and the pool of these six isolates). Accession ID for plasmid pAKU_1 is AM412236 and pGY1 is EF150947.

3.3 Results

3.3.1 Comparison of seven Paratyphi A genome sequences

3.3.1.1 Whole genome sequencing

Finished sequence data was available for two Paratyphi A genomes, isolates ATCC9150 and AKU_12601. These isolates were resequenced in the Sanger Institute Solexa sequencing pipeline (561), along with five additional isolates chosen on the basis of interesting phenotypes or their use in the lab (see Table 3.1). Reads of 35 bp were generated for each strain, to a depth of 27x - 46x.

3.3.1.2 SNP analysis

Short reads (35 bp) generated by resequencing of AKU_12601 were aligned to the finished sequence, which identified five high quality single base discrepancies between the assemblies. One was found to be an erroneous base call in the finished sequence following checking of capillary trace files and was corrected prior to comparison with other genomes in this study. The remaining four base calls (6-, 8-, 10-, and 20-fold read depth in Solexa data) may be errors in Solexa sequencing or base calling, or reflect

Strain	Source	Year	Motivation	Solexa	454
AKU_12601[1]	Karachi, Pakistan	2002	Finished sequence	22x	n/a
ATCC9150[1]	-	-	Finished sequence	41x	n/a
C1468[2]	Kolkata, India	2005	H_2S positive	43x	n/a
6911[3]	Nairobi, Kenya	2007	Cipro resistant	9x	
6912[3]	Nairobi, Kenya	2007	Cipro resistant	43x	16x
38/71[4]	Delhi, India	2006	Efflux phenotype	42x	n/a
BL8758[5]	Karachi, Pakistan	2004	Lab strain	46x	n/a

Table 3.1: Paratyphi A strains with whole genome sequence data available - Isolates were provided by [1]John Wain, Sanger Institute, UK; [2]Shanta Dutta, National Institute of Cholera and Enteric Diseases, Kolkata; [3]Sam Kariuki, Kenya Medical Research Institute, Nairobi; [4]Dr Rajni Gaind, Safdarjung Hospital, Delhi; [5]Rumina Hasan, Aga Khan University Hospital, Karachi. Year and location of isolation, and motivation for generating whole genome data is given for each isolate. Read depth is given for Solexa and 454 data. Cipro = ciprofloxacin.

genuine mutations arising during culturing in the laboratory. SNPs were detected in the remaining six genomes by comparison to AKU_12601 (see 3.2.2). In total, 227,377 bp (5.0%) of the AKU_12601 were identified as repeated or prophage sequences (see Methods 3.2.1, Table 3.2), including three prophage regions, IS elements, and duplicated genes such as the *oad* and *ccm* operons. Note that this is in line with the earlier analysis of Typhi where 7.4% of the Typhi CT18 genome, including 7 prophage regions, was excluded from SNP analysis. SNPs in these repetitive regions were excluded, resulting in a total of 550 SNPs.

For each SNP detected in any isolate, alleles were checked in all six isolates (as described in (2.3.1.5) and the Typhi CT18 sequence as a representative outgroup. There were 147 SNPs for which alleles could not be determined in all Paratyphi A strains, these were excluded from phylogenetic analysis. The remaining 403 SNPs were used to generate a maximum parsimony phylogenetic tree using Typhi CT18 as an outgroup (determined using the `mix` algorithm in the `phylip` package (573)). This produced a balanced tree, with three lineages emerging from the root (Figure 3.1). To confirm the position of the root was not inaccurately inferred by the use of Typhi as an outgroup, alleles were also determined for seven additional *S. enterica* serovars and a parsimony splits network constructed (see 3.2.3). The resulting network, shown in Figure 3.2,

(a) Genomic bp	**Excluded**	**Included**	**Total**	**% Included**
Intergenic	15476	533560	549036	97.2
rRNA	32119	0	32119	0.0
tRNA	4312	1436	5748	25.0
Protein coding	175470	3819424	3994894	95.6
All bases	227377	4354499	4581797	95.0
(b) Genes				
Total	221	4064	4285	94.8
IS elements	14	0		
Phage-like	46	0		
Phage	128	0		
Other	33	0		

Table 3.2: Repetitive Paratyphi A AKU_12601 sequences excluded from SNP detection anlaysis - Details of (a) genomic nucleotides and (b) genes in the AKU_12601 genome that were included or excluded from SNP detection analysis.

supports the positioning of the root close to the three-way split between the three major lineages. A single homoplasic SNP was identified, introducing a stop codon within the coding sequence of SSPA1928a (a component of a glutamate ABC transporter) in AKU_12601 as well as isolates BL8758, 38/71, 6911 and 6912. The distribution of SNPs per gene followed an exponential distribution (Figure 3.3) with no clustering of SNPs within genes.

3.3.1.3 Gene acquisition

The genomes of Paratyphi A ATCC9150 and AKU_12601 were collinear, with no variation in prophage content. Assemblies of the five other genomes (see 3.2.4) did however reveal some gain and loss of prophage sequences in Paratyphi A. The prophage at AKU_12601 coordinates 2.65 Mbp (SPA-2-SopE) was missing from isolates 38/71, 6911 and 6912. The latter two isolates were also lacking the prophage at AKU_12601 coordinates 2.67 Mbp (SPA-3-P2). A novel prophage sequence was inserted between SSPA3930 SSPA3931 in genomes 6911 and 6912, generating 15 bp direct flanking repeats. The phage was similar to P2-like prophages sequenced in the genomes of several *E. coli* and *Shigella* isolates, and was not detected in the other Paratyphi A isolates.

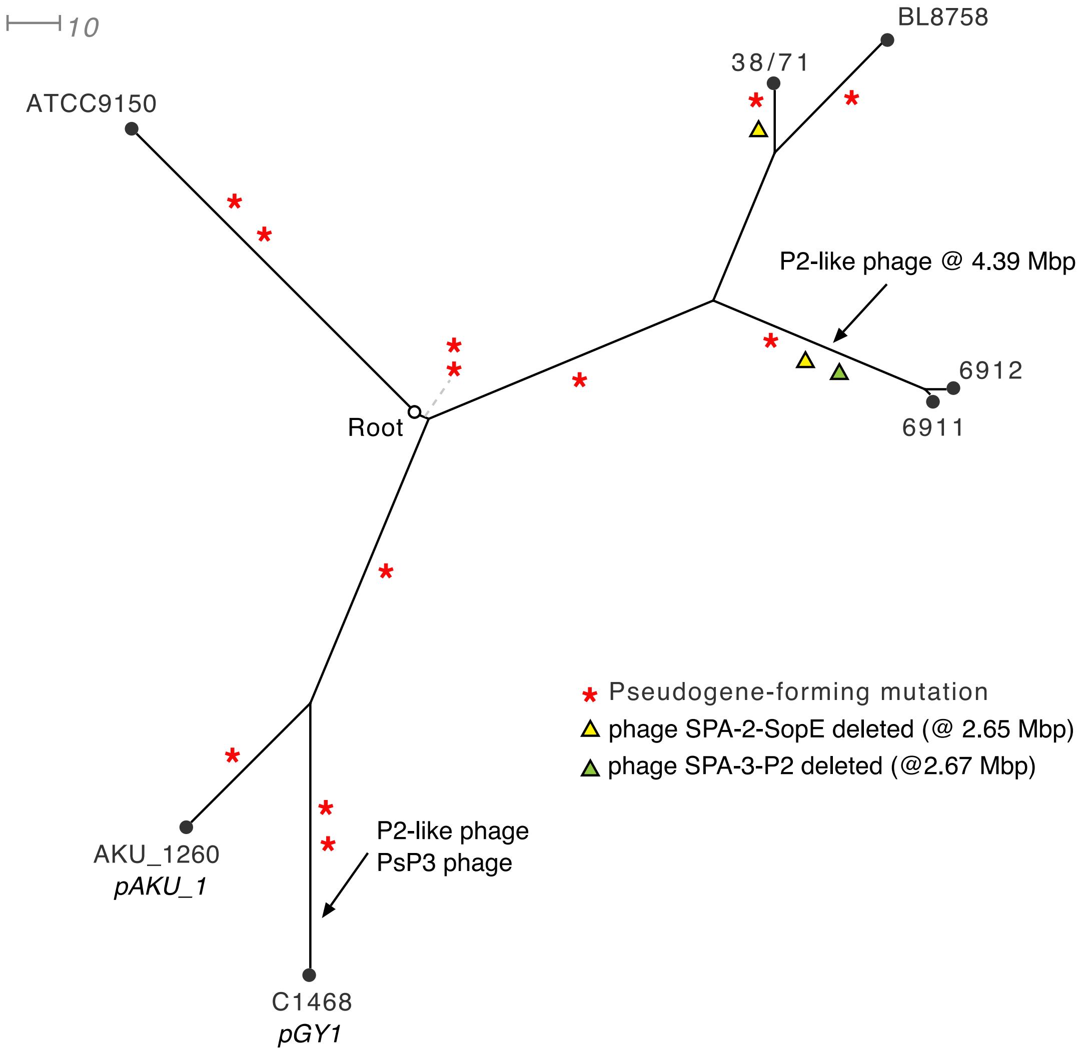

10
BL8758
38/71
ATCC9150
P2-like phage @ 4.39 Mbp
6912
Root
6911
* Pseudogene-forming mutation
△ phage SPA-2-SopE deleted (@ 2.65 Mbp)
△ phage SPA-3-P2 deleted (@2.67 Mbp)
P2-like phage
PsP3 phage
AKU_1260
pAKU_1
C1468
pGY1

Figure 3.1: Phylogenetic tree of seven Paratyphi A isolates based on genome-wide SNPs detected by sequencing - The tree was constructed using maximum parsimony methods based on 403 loci, using Typhi as an outgroup to root the tree. Scale bar = 10 SNPs. Phage insertions are labelled with arrows. Pseudogene forming mutations and phage deletions are indicated by symbols as indicated.

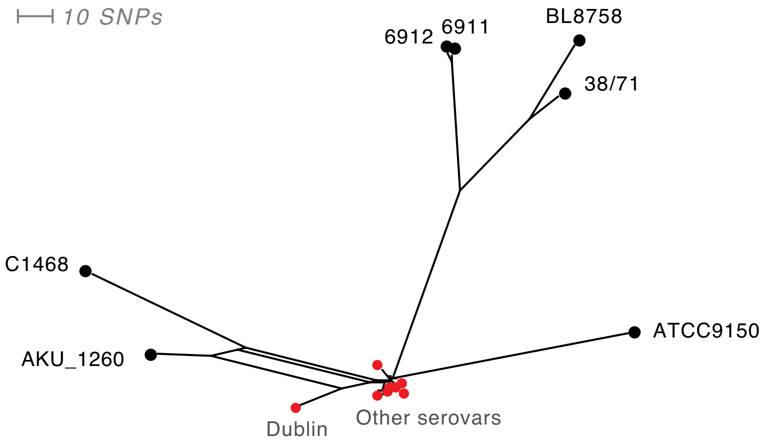


Figure 3.2: Phylogenetic network of seven Paratyphi A isolates including seven serovars as outgroups. - Serovars Typhi, Typhimurium, Enteritidis, Paratyphi B, Choleraesuis, Dublin, Galinarum and Pullorum were used as outgroups (red nodes).

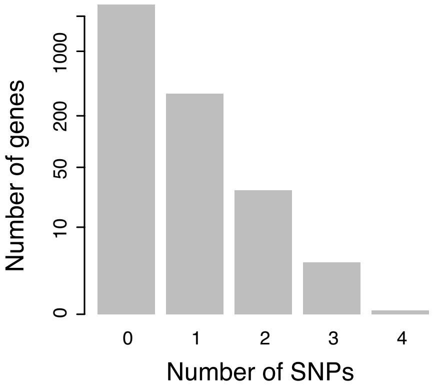


Figure 3.3: Distribution of SNPs per Paratyphi A gene - Note the y-axis, number of genes, is on a natural logarithmic scale.

Two other P2-like prophage, including one similar to PsP3 (EMBL:AY135486) were inserted in the genome of C1468, although the precise insertion sites could not be determined. These prophage sequences were not detected in the other Paratyphi A isolates.

Two plasmids have been sequenced from Paratyphi A: the MDR IncHI1 plasmid pAKU_1 (see Chapter 5) and pGY1 (287). Among the seven isolates, pAKU_1 was found only in AKU_12601 from which it was originally sequenced (Chapter 5), and pGY1 was found in C1468 (see 3.2.4). The only novel insertion sequence evident among the genomes was *IS*10, inserted into two locations in the AKU_12601 chromosome (see 3.2.4). *IS*10 is part of the *Tn*10 transposon encoded in pAKU_1 (see 5.3.1.2) and the *IS*10 sequences inserted in the AKU_12601 chromosome were 100% identical at the nucleotide level to that encoded in the plasmid pAKU_1. Thus it is highly likely that the chromosomal insertions were acquired from pAKU_1. This is similar to the situation in Typhi, where *IS*1 was only detected in the chromosomes of isolates known to contain the IncHI1 plasmid which itself carries several copies of *IS*1.

3.3.1.4 Insertion/deletion mutations

A total of 39 insertion/deletion (indel) events, including 13 differences in homopolymeric tracts, were identified between the finished sequences of AKU_12601 and ATCC9150 (see 3.2.4 and Table 3.3). Five variable number tandem repeats (VNTRs) were identified between AKU_12601 and ATCC9150, including one less tandem copy each of the tRNA-*Gly* and *rtT* RNA genes (606) in AKU_12601 (repeat numbers for these genomes are given in Table 3.3, but could not be resolved for the other five genomes using short sequencing reads). An additional 122 bp sequence was present in AKU_12601 between the *iap* and *ygbF* genes, including two additional copies of a 30 bp repeat sequence present in six copies in ATCC9150. The sequence in 454 data from isolates 6911 and 6912 matched that of AKU_12601 at this locus. Smaller VNTRs were identified within SSPA0767 and SSPA2694, resulting in repeats differing by two and four amino acids respectively in the encoded proteins. VNTRs are useful as genetic markers for typing *Salmonella enterica* serovars and variability in the SSPA2694 VNTR among Paratyphi A isolates has been reported previously (481). Isolates 6911 and 6912 matched ATCC9150 at this VNTR (N=5), but could not be resolved for the SSPA0767 VNTR.

Coding effect	Gene	Mutation	Strain	Gene function
pseudo-forming	*aidB*	217 bp del	A	Probable acyl Co-A dehydrogenase
pseudo-forming	*asnB*	1 bp del (H)	B	Asparagine synthetase B
pseudo-forming	*ccmH*	95 bp del	A	Cytochrome c-type biogenesis protein H2
pseudo-forming	*nmpC*	1338 bp ins (IS)	A	Outer membrane porin
pseudo-forming	*pduF*	1 bp del (H)	B	Propanediol diffusion facilitator
pseudo-forming	*pduG*	171 del	A	Diol/glycerol dehydratase reactivating factor, large subunit
pseudo-forming	*proQ*	7 bp del	B	ProP effector
pseudo-forming	*rbsC*	1 bp ins (H)	B	High affinity ribose transport protein
pseudo-forming	*rbsR*	1 bp ins (H)	B	Ribose operon repressor
pseudo-forming	*rhlB*	2 bp ins	B	Putative ATP-dependent RNA helicase
pseudo-forming	SSPA3202	1 bp ins (H)	A	Putative lipoprotein
pseudo-forming	*tesB*	352 bp del	B	Acyl-CoA thioesterase II
pseudo-forming	*wcaA*	1 bp ins (H)	A	Putative glycosyl transferase
pseudo-forming	*yaaJ*	1 bp del (H)	B	Putative amino-acid transport protein
pseudo-forming	*yeaG*	1 bp del	B	Conserved hypothetical protein
pseudo-forming	*yeeO*	1 bp ins (H)	B	Putative inner membrane protein
already pseudo	SSPA4008a	1338 bp ins (IS)	A	Hypothetical protein
coding change	SSPA0767	VNTR (N=1 vs 2)	-	Putative CoA-dependent proprionaldehyde dehydrogenase
coding change	SSPA2694	VNTR (N=5 vs 7)	-	Putative inner membrane protein
coding change	SSPA3369	9 bp del	B	Hypothetical protein
coding change	SSPA3558a	10 bp del	B	Possible transferase
coding change	SSPA3928	3 bp del	B	Putative exported protein
intergenic	before *rnpB*	1 bp ins (H)	A	-
intergenic	before SSPA1079	1 bp ins (H)	B	-
intergenic	before SSPA2694	3 bp ins (H)	-	-
intergenic	before *rrfH*	1 bp del	B	-
intergenic	before SSPA1464	1 bp in/del	-	-
intergenic	before *rffD*	1 bp del	B	-
intergenic	after SPA3575	1 bp del	B	-
intergenic	before SSPA3682	2 bp ins	A	-
intergenic	after *iap*	VNTR (N=6 vs 8)	-	-
RNA	*rtT*	VNTR (N=5 vs 4)	-	RNA associated with tRNA-*tyrT*
RNA	*rrlC*	1 bp del (H)	B	23S rRNA
RNA	*rrlD*	1 bp ins (H)	B	23S rRNA
RNA	*rrsB*	1 bp ins	B	16S rRNA
RNA	*csrB*	1 bp ins (H)	A	Regulation of *csrA*
RNA	tRNA-*ProL*	7 bp del	B	tRNA
RNA	tRNA-*GlyW*	VNTR (N=3 vs 2)	-	tRNA

Table 3.3: Insertion/deletion mutations detected between two Paratyphi A genomes - Strain containing mutation: A = AKU_12601, B = ATCC9150. Mutation type: H = homopolymer, IS = *IS*10 insertion.

The feasibility of detecting small indel mutations from Solexa data was tested using the short read data from ATCC9150. Two methods of detection were trialled: (i) short reads were assembled *de novo* using Velvet and compared to AKU_12601 using MUMmer to detect indels, and (ii) short reads were aligned directly to AKU_12601 using Maq and indels called using SAMtools (607) to analyse the alignments. Indels detected by either method were compared to those detected from comparison of the finished ATCC9150 and AKU_12601 genomes (Figure 3.4). Of the short (≥20 bp) indels detected in the finished sequence, only 8 (35%) were detected using both assembled and directly aligned reads. Analysis of assembled data was more sensitive, with 12 (57%) of indels successfully detected. However, this analysis also identified an additional 9 indels that are not present in the ATCC9150 finished sequence, putting specificity at just 57%. The remaining five Paratyphi A genomes were therefore not analysed for short indels, as the error rates were considered too high to allow a reliable analysis of this kind of variation. They were however checked for deletions of ≥20 bp compared to AKU_12601, which can be reliably detected because at >60% of read length they cause reads to be simply unmappable rather than producing unreliable gapped alignments. Besides the variation in phage and IS sequences described above, only five novel deletions were identified. These ranged from 20-120 bp in size, affecting pseudogene SSPA1125a and potentially inactivating four other genes listed in Table 3.4c.

3.3.1.5 Loss of gene function

Eleven nonsense SNPs were identified among the seven strains, resulting in the formation of 11 pseudogenes since their most recent common ancestor (Table 3.4a,c). These mutations were randomly distributed in the phylogenetic tree, as shown in Figure 3.1. A further 16 pseudogene-forming mutations were identified between the finished genome sequences of AKU_12601 and ATCC9150, including one *IS*10 insertion and 15 other indel events (Table 3.4b). An additional four deletions of 20-120 bp, likely resulting in disruption of coding sequences, were identified among the other five genomes, although it was not possible to detect smaller deletions (Table 3.4c). Thus the 31 pseudogene-forming mutations identified in this study (Table 3.4) likely underestimate the level of gene inactivation since the last common ancestor of these seven Paratyphi A genomes.

a. Gene	Mutation	Isolate	Gene product
gltJ	nonsense SNP	AKU_12601	Glutamate/aspartate transport system permease
SSPA1447	nonsense SNP	AKU_12601	Putative oxidoreductase
SSPA3581	nonsense SNP	AKU_12601	Conserved hypothetical protein
yhaO	nonsense SNP	ATCC9150	Putative transport system protein
yjhW	nonsense SNP	ATCC9150	Putative membrane protein
trpD	nonsense SNP	AKU_12601	Anthranilate synthase component II

b. Gene	Mutation	Isolate	Gene product
aidB	del	AKU_12601	Probable acyl Co-A dehydrogenase
asnB	1 bp del (homopol)	ATCC9150	Asparagine synthetase B
ccmH	88 bp del	AKU_12601	Cytochrome c-type biogenesis protein H2
nmpC	IS10 ins	AKU_12601	Outer membrane porin
pduF	1 bp del (homopol)	ATCC9150	Propanediol diffusion facilitator
pduG	171 bp del	AKU_12601	Propanediol dehydratase reactivation protein
proQ	7 bp del	ATCC9150	ProP effector
rbsC	1 bp ins (homopol)	ATCC9150	High affinity ribose transport protein
rbsR	1 bp ins (homopol)	ATCC9150	Ribose operon repressor
rhlB	2 bp ins	ATCC9150	Putative ATP-dependent RNA helicase
SSPA3202	1 bp ins (homopol)	AKU_12601	Putative lipoprotein
tesB	352 bp del	ATCC9150	Acyl-CoA thioesterase II
wcaA	1 bp ins (homopol)	AKU_12601	Putative glycosyl transferase
yaaJ	1 bp del	ATCC9150	Putative amino-acid transport protein
yeaG	1 bp del	ATCC9150	Conserved hypothetical protein
yeeO	1 bp ins (homopol)	ATCC9150	Putative inner membrane protein

c. Gene	Mutation	Isolate	Gene product
SSPA0470	nonsense SNP	BL8758, 38/71	Conserved hypothetical protein
SSPA0720	del	C1468	Membrane transport protein
SSPA1311	nonsense SNP	C1468	Putative HlyD-family protein
SSPA2643	del	6911, 6912	Lactaldehyde reductase
SSPA2775	nonsense SNP	C1468	Nucleoside permease
SSPA3629	del	BL8758	Two-component sensor kinase protein
SSPA3565	nonsense SNP	38/71	Molybdopterin-guanine dinucleotide biosynthesis B
SSPA4071	del	BL8758	Lipoate-protein ligase A
SSPA4083	nonsense SNP	6911, 6912	Putative two-component response regulator

Table 3.4: Pseudogene-forming mutations detected among seven Paratyphi A genomes - a,b: Nonsense SNPs and insertion/deletion mutations detected between finished genomes AKU_12601 and ATCC9150. c: Additional mutations identified in the other five genomes. Note small insertion/deletion mutations may exist in these five genomes, but could not be reliably assessed with current software.

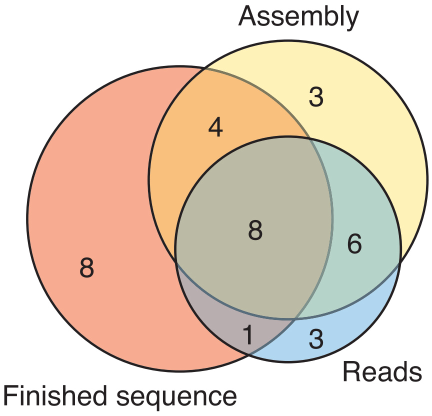


Figure 3.4: Detection of small indels from short read data - Indels of <20 bp detected from finished sequence (pink) compared to those detected from alignment of *de novo* assembled contigs (yellow) or reads (blue) to the reference.

3.3.2 Optimisation of SNP detection from pooled sequence data

In order to screen for SNPs among a large collection of >150 Paratyphi A isolates, a DNA pooling approach was used. Since there have been no reported studies using short read sequencing to detect SNPs in pooled DNA samples, it was necessary to develop and validate a method for calling SNPs from this data and estimating SNP frequency within pools. This was done using a pool containing 400 ng of DNA from each of the isolates in Table 3.1, excluding AKU_12601. The pooled DNA was sequenced in the Solexa pipeline at the Sanger Institute. A total of 5.4 million reads of 35 bp were generated, 97.77% of which were mapped to the Paratyphi A AKU_12601 reference genome sequence using Maq. This equates to an average read depth of 40x across the pool, or 6x per isolate.

3.3.2.1 SNP detection and frequency estimation

An initial mapped assembly of the reads was performed using Maq to align reads to the AKU_12601 finished genome, with the number of haplotypes set to 6 (`-N` option) and default settings for other parameters. This assembly is the basis upon which Maq calls SNPs, so in order to determine the optimal parameters, the maximum number of mismatches allowed to map a read (`maq assemble -m` option) was varied from 0-7 bp, i.e. up to 20% mismatches per 35 bp read. Potential SNPs identified by Maq from this assembly were then analysed for quality and to estimate the frequency of the SNP within the pool. This was achieved using information generated by Maq's `pileup` program, which retrieves the base call (A, C, G, T) and base call quality (a phred-like quality score) for each base that is mapped to a given SNP locus. The minimum mapping quality required for a read to be included in this output (`maq pileup -q` option) was varied from 10-50. The frequency of each SNP k in pool p containing S_p strains was estimated using data on each read i of N reads mapped to the SNP locus, including the base quality $q_{k,p,i}$. Frequencies were calculated according to the formulae below, calculations were implemented in a Perl script. Here $p_{k,p}$ is the estimated proportion of isolates in pool p containing SNP k, while $freq_{k,p}$ is the estimated frequency of (i.e. number of isolates containing) SNP k in pool p.

$$p_{k,p} = \frac{\sum_{i=1}^{N} w_{k,p,i}.x_{k,p,i}}{\sum_{i=1}^{N} w_{k,p,i}} \tag{3.1}$$

($x_{k,p,i} = 1$ if SNP allele, 0 otherwise)

$$var(p_{k,p}) = \frac{p_{k,p}.(1-p_{k,p}).\sum_{i=1}^{N} w_{k,p,i}^2}{(\sum_{i=1}^{N} w_{k,p,i})^2} \tag{3.2}$$

$$freq_{k,p} = S_p * p_{k,p} \tag{3.3}$$

$$95\%CI(freq_{k,p}) = S_p * (p_{k,p} \pm 1.96.\sqrt{var(p_{k,p})}) \tag{3.4}$$

Quality-weighted frequency estimates were calculated according to a number of different weighting schemes (used in equations 3.1 and 3.2):

$$w_{k,p,i} = 1 \tag{3.5}$$

$$w_{k,p,i} = q_{k,p,i} \tag{3.6}$$

$$w_{k,p,i} = \frac{q_{k,p,i}}{q_{max}} \tag{3.7}$$

$$w_{k,p,i} = (\frac{q_{k,p,i}}{q_{max}})^2 \tag{3.8}$$

$$w_{k,p,i} = q_{k,p,i} - q_{min} \tag{3.9}$$

$$w_{k,p,i} = \frac{q_{k,p,i} - q_{min}}{q_{max} - q_{min}} \tag{3.10}$$

$$w_{k,p,i} = (\frac{q_{k,p,i} - q_{min}}{q_{max} - q_{min}})^2 \tag{3.11}$$

$$w_{k,p,i} = 1 - \frac{1}{q_{k,p,i}} \tag{3.12}$$

$$w_{k,p,i} = 1 - 10^{\frac{-q_{k,p,i}}{10}} \tag{3.13}$$

Here q_{min} is the minimum base quality for inclusion in the analysis (set to 20 in this study), and q_{max} is the maximum possible calibrated quality score (99 for this data set).

3.3.2.2 Comparison of potential methods

SNPs previously detected between AKU_12601 and the six strains in the pool (3.3.1.2) were analysed to determine their expected frequencies in the pool. These expected SNP frequencies were compared to those estimated from the pool, using all possible combinations of assembly parameters (affecting SNP detection), pileup parameters (affecting SNP frequency estimates) and weighting measures (affecting SNP frequency estimates). For each combination of parameters, the following measures were calculated (after removing SNP calls in repetitive or phage sequences):

- sensitivity of SNP detection, i.e. proportion of the 550 known SNPs that were detected with an estimated frequency of ≥1 strain,
- false positive rate of SNP detection, i.e. proportion of the SNPs detected with estimated frequency of ≥1 strain that were not expected to be present in the pool,
- correlation (Pearson R^2) between the expected and estimated allele frequencies, and
- the proportion (among the 403 SNPs with reliable frequency estimates) of loci for which estimated and expected allele frequencies differed by ≥1 strain, i.e. the rate of incorrect frequency estimates.

The weighting measures 3.9 - 3.13 were excluded from detailed analysis as they gave highly insensitive or inaccurate results (see Figure 3.5). Analysis of variance tables for each measure are given in Table 3.5.

Using any combination of weights (equations 3.5 - 3.8), mismatches (1-7 per read) and mapping qualities (10-50), SNP detection was quite sensitive (78% - 84% of expected

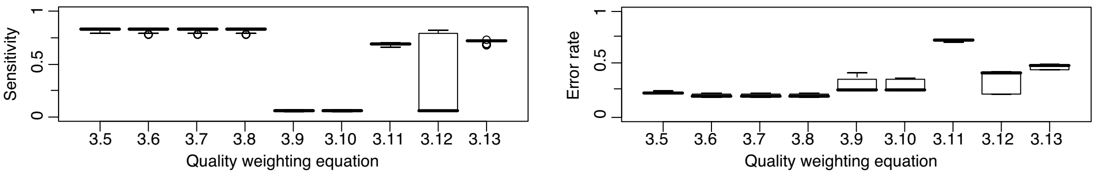


Figure 3.5: Sensitivity and error rates for different weighting measures - The distribution of each measure is shown as a box-and-whisker plot where black bars indicate median values, boxes indicate interquartile ranges, and whiskers indicate the full range of values observed using each weighting equation. Weighting equations are labelled as in the text; each measure was calculated across all combinations of mismatches and mapping qualities.

SNPs detected) and the experimentally observed frequencies were strongly correlated with the expected frequencies (Pearson R^2 0.92 - 0.95) (Figure 3.6). Detection sensitivity was highly dependent on SNP frequency, with 37% detection for SNPs present in just 1 strain, compared to 95% and 100% detection respectively for SNPs present in 2 or $\geq$3 strains. The false positive rate varied between 5 - 18% using different methods and was closely correlated with number of mismatches allowed during mapping (Figure 3.7). However setting the number of mismatches $\leq$1 reduced sensitivity too low (78%), thus the optimal setting was $\leq$2 mismatches per read (mean false positive rate 8.8%, mean sensitivity 82.7%). The proportion of incorrect frequency estimates was reduced by using any of the weighting methods 3.6 - 3.8 and was also dependent on mapping quality. The lowest rate of incorrect estimates (19%) was seen with a minimum mapping quality 40; lowering or raising the cutoff increased the rate to >20% while offering very little improvement in false positive rate or sensitivity (Figure 3.8). The most accurate measurements (low false positive rate, low error rate, high R^2) were obtained using the weighting method shown in equation 3.8, regardless of other parameters (Figure 3.9), and the difference between expected and estimated frequencies was never more than one strain.

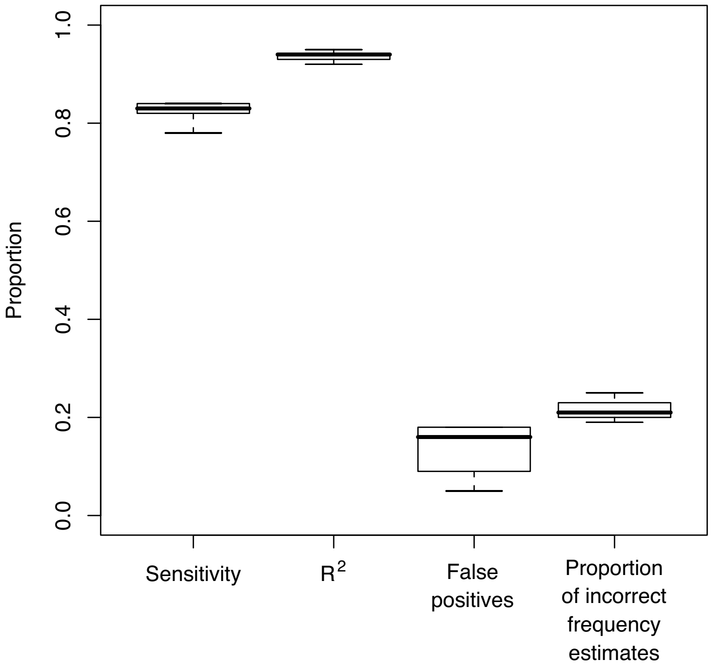


Figure 3.6: Ranges for each accuracy measure - The distribution of each measure is shown as a box-and-whisker plot where black bars indicate median values, boxes indicate interquartile ranges, and whiskers indicate the full range of values observed. Each measure was calculated for every combination of methods tested, including assembly parameters, pileup parameters and weighting methods 3.5 - 3.8.

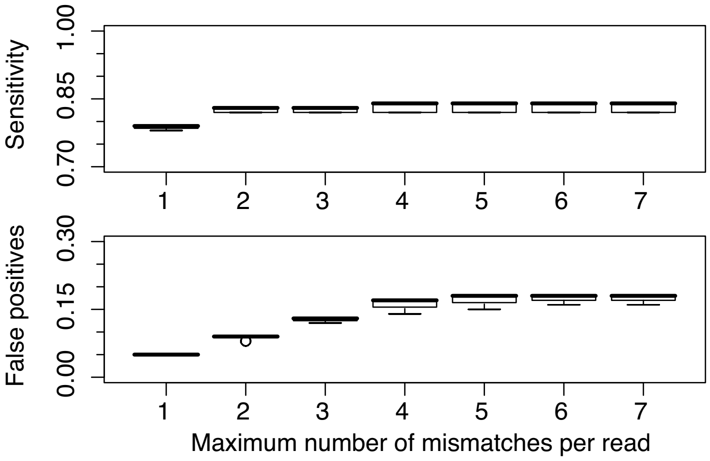


Figure 3.7: Sensitivity and false positive rates for different assembly parameters - The distribution of each measure is shown as a box-and-whisker plot where black bars indicate median values, boxes indicate interquartile ranges, and whiskers indicate the full range of values observed. Each measure was calculated for all combinations of pileup parameters and weighting methods 3.5 - 3.8.

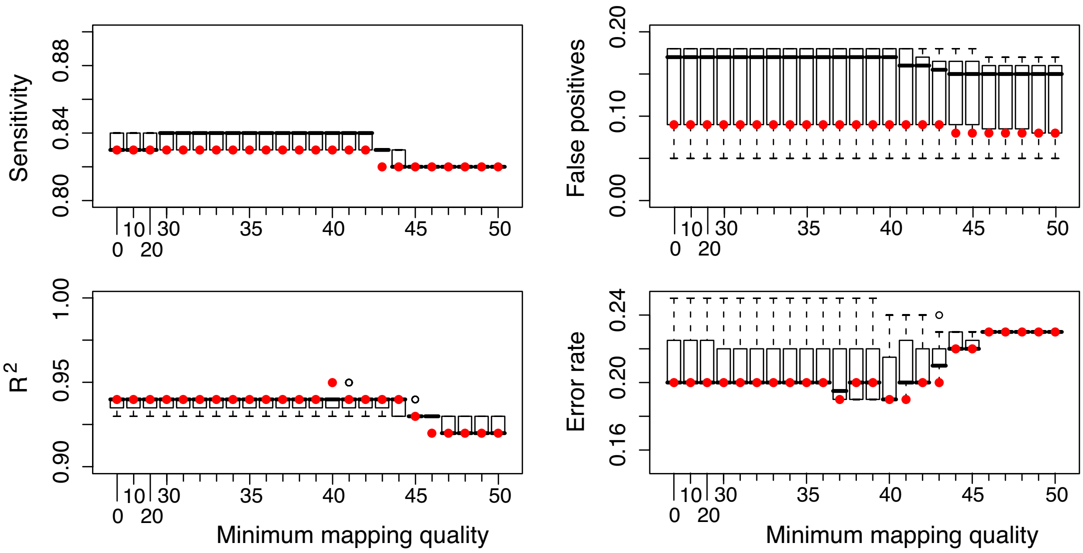


Figure 3.8: Accuracy measures for different pileup parameters - The distribution of each measure is shown as a box-and-whisker plot where black bars indicate median values, boxes indicate interquartile ranges, and whiskers indicate the full range of values observed. Each measure was calculated for all combinations of assembly parameters and weighting methods 3.5 - 3.8. Red circles show values for mismatch ≤ 2, weighting equation 3.8.

a. Parameter	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Weight	3	0.00003	0.00001	0.0647	0.9785
Mismatches	1	0.086066	0.086066	549.8773	<2.2E-16
Mapping Q	1	0.009648	0.009648	61.6447	1.65E-14
Residuals	666	0.104241	0.000157		

b. Parameter	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Weight	3	0.00132	0.000444	1.2179	0.3023
Mismatches	1	1.20531	1.20531	3341.63	<2.2E-16
Mapping Q	1	0.0092	0.0092	25.5073	5.70E-07
Residuals	666	0.24022	0.00036		

c. Parameter	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Weight	3	0.0048076	0.0016025	44.9171	<2.2E-16
Mismatches	1	0.0000405	0.0000405	1.1355	2.87E-01
Mapping Q	1	0.0056977	0.0056977	159.6986	<2.2E-16
Residuals	666	0.0237612	0.0000357		

d. Parameter	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Weight	3	0.090154	0.030051	234.092	<2.2E-16
Mismatches	1	0.002037	0.002037	15.868	7.54E-05
Mapping Q	1	0.012874	0.012874	100.389	<2.2E-16
Residuals	666	0.085497	0.000128		

Table 3.5: Analysis of variance for factors affecting accuracy of SNP detection and frequency estimation - (a) Sensitivity of SNP detection, (b) Rate of false positive SNP calls, (c) correlation (Pearson R^2) between estimated and expected SNP frequencies, (d) rate of incorrect frequency estimates.

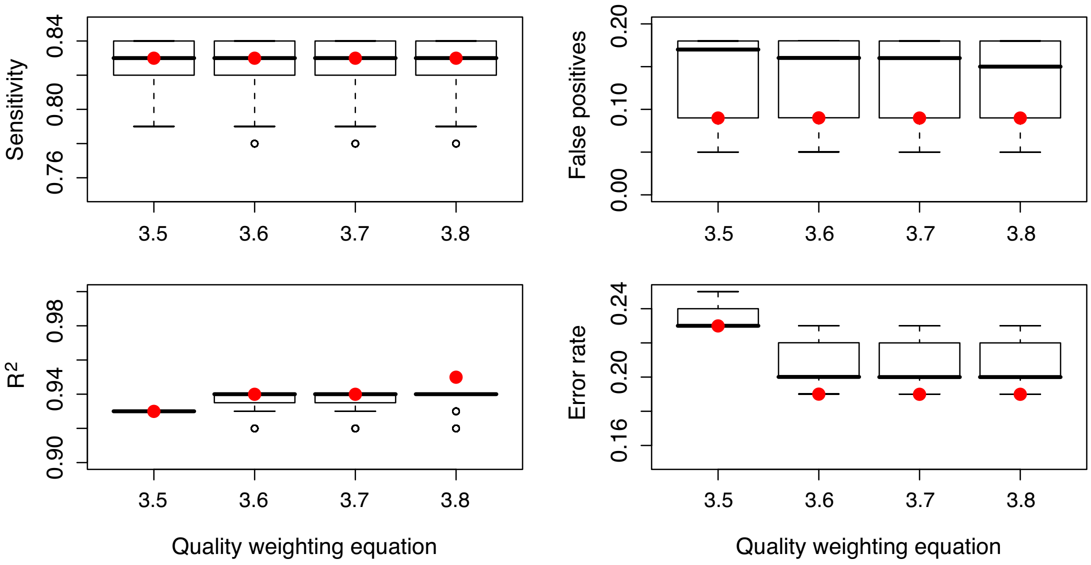


Figure 3.9: Accuracy measures for different weighting equations - The distribution of each measure is shown as a box-and-whisker plot where black bars indicate median values, boxes indicate interquartile ranges, and whiskers indicate the full range of values observed. Each measure was calculated for all combinations of assembly parameters and pileup parameters. Red circles show values for mismatch $\leq$ 2, mapping quality $\geq$40.

3.3.2.3 Performance of optimised method

The optimal combination of methods and parameters, used for subsequent SNP calling from sequence data on all Paratyphi A pools, was:

- $\leq$2 mismatches between read and reference to be included in assembly, from which SNPs are called;
- read mapping quality $\geq$40 to include data from the read in SNP frequency estimates; and
- weighting method: $w_{k,p,i} = (\frac{q_{k,p,i}}{q_{max}})^2$.

Using these parameters to analyse short read sequence data from the test pool resulted in SNP detection sensitivity of 82.7%, false positive rate of 9% and strong correlation between expected and estimated SNP frequencies (R^2=0.94, 81% of frequency estimates correct), see Figure 3.10). The sample standard deviations calculated for the frequency estimates were weakly associated with error, with slightly higher sample standard deviations observed among SNPs whose estimated frequency differed from expected (mean standard deviation 0.0688) compared to SNPs whose frequency estimates were as expected (mean standard deviation 0.0596) (p-value = 0.0001 using Welch two-sample T-test). However the ranges of standard deviations were completely overlapping for incorrect and correct estimates (Figure 3.11), so standard deviation cannot be considered a particularly useful indicator of whether a frequency estimate is likely to be correct. To confirm that specifying the expected number of haplotypes (i.e. strains) present in a pool increased sensitivity of SNP detection, the test pool data was analysed using the optimised methods described above, but without specifying the number of strains using the `-N` option. Sensitivity dropped to 70.0%, with only a minor reduction in false positive rate (to 8.1%), thus the `-N` option has been used for all subsequent analysis.

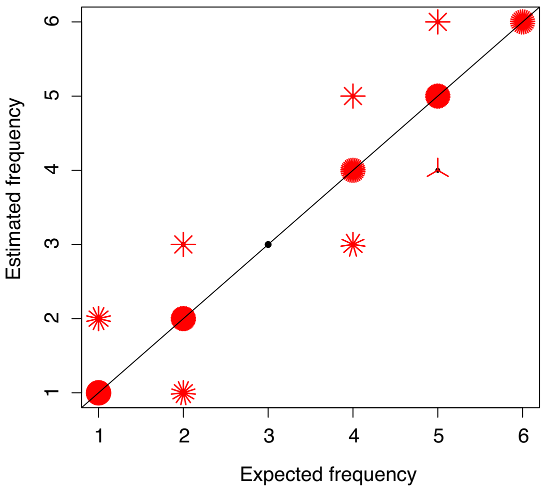


Figure 3.10: Expected frequencies vs SNP frequencies estimated from Paratyphi A test pool sequence data - The sunflower plot is designed to aid visualisation of correlations between discrete variables. It is similar to an x-y plot, but uses radial red lines to indicate the number of data points that share each combination of discrete x-y values. Here, most SNPs lie on the line y=x (black line), where estimated SNP frequency = expected SNP frequency (points with many red radial lines). Fewer SNPs (one radial line per SNP) lie outside this line, demonstrating the low rate of errors in frequency estimation.

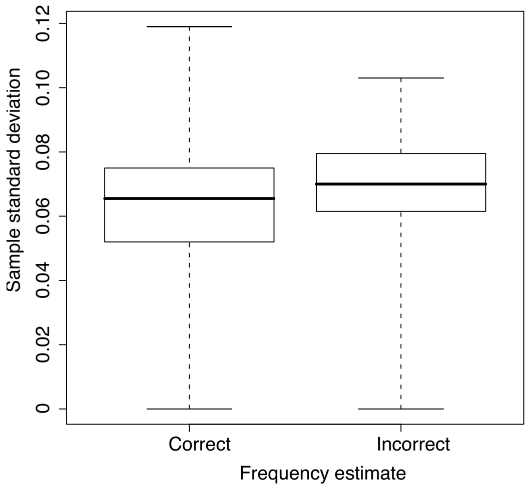


Figure 3.11: Distributions of sample standard deviations calculated among SNPs with correct and incorrect frequency estimates - Black bars indicate median values, boxes indicate interquartile ranges, and whiskers indicate the full range of values observed.

3.3.2.4 Performance of optimised method over a range of read depths

Read depth for the test pool was 40x, equal to the mean read depth obtained for the experimental Paratyphi A pools (see 3.3.3 below), so similar levels of accuracy can be expected from most of the experimental data generated in this study. However to assess performance at lower read depths, data sets were simulated by randomly sampling subsets of the Paratyphi A test pool reads. Fifty random samples of reads were generated for each level of pool-wide read depth 1x, 2x, up to 39x. The results were compared to the expected SNP frequencies as above and measures of accuracy were calculated as before. Figure 3.12 shows how accuracy of SNP detection declined with read depth. There was very little difference in performance between pool-wide read depth 35x-40x ($\geq$5.8x per strain). Sensitivity declined to 68.8% at read depth 18x (3x per strain) and false positive rate increased to 13.2%. Frequency estimation also suffered a little, with Pearson correlation dropping to R^2=0.88 and the rate of incorrect estimates increasing to 35.8% of SNPs.

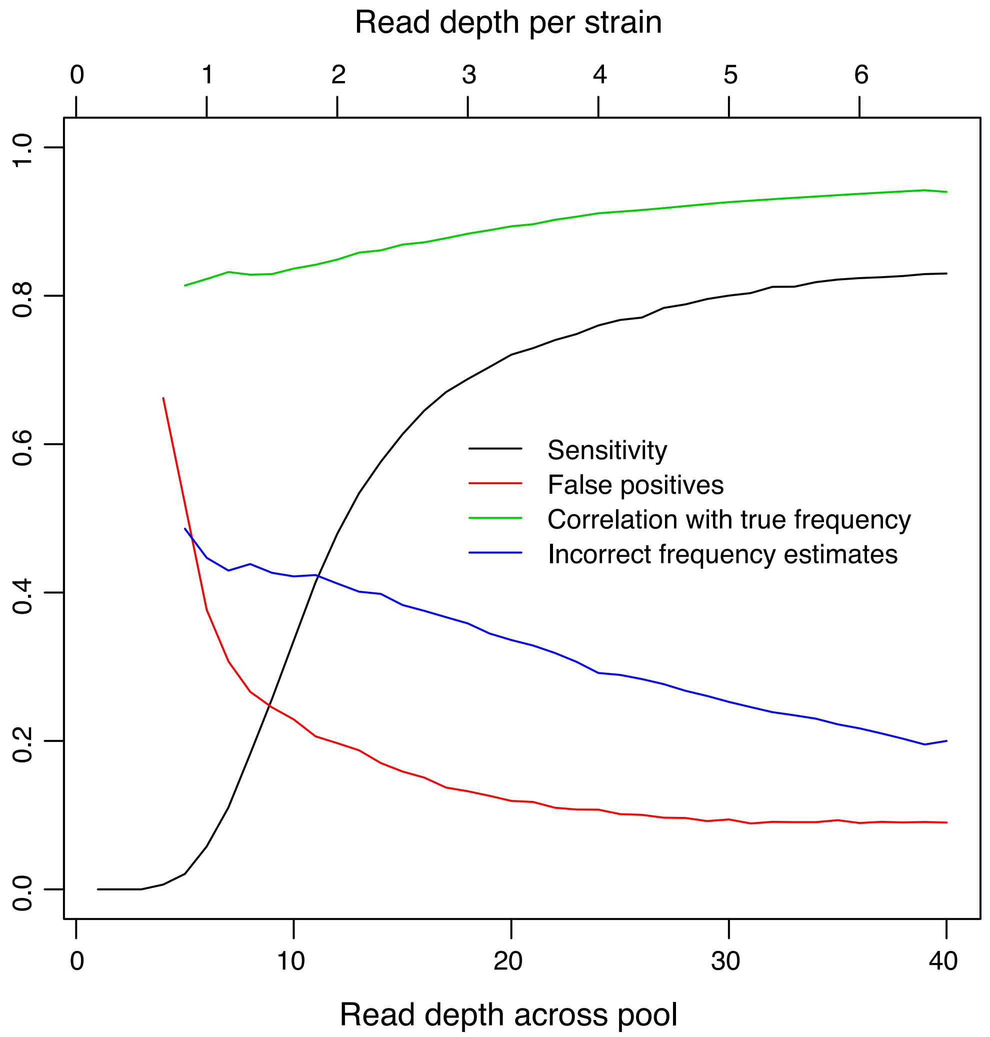


Figure 3.12: Error rates expected at different levels of read depth - Accuracy estimates were based on random sampling of reads from the test pool data at different read depths (1x, 2x, ... 39x across the pool). Data points plotted here show the mean value observed across fifty random samples at each read depth.

3.3.3 Genomic variation detected in 159 Paratyphi A isolates by pooled sequencing

A collection of genomic DNA samples from 159 Paratyphi A isolates was assembled by Mark Achtman (Environmental Research Institute, Cork, Ireland) and John Wain (Sanger Institute/Health Protection Agency, UK) from a combination of *Salmonella* reference laboratories and research laboratories around the world. A total of 107 isolates were selected from the *Salmonella* reference laboratory at the Pasteur Institute (Paris, France), which were isolated from travellers returning to France with enteric fever. These isolates were chosen to represent maximum diversity according to geography, date of isolation, phage type and any other phenotypic information available. DNA was provided by Francois-Xavier Weill (Pasteur Institute, Paris, France), and samples were grouped randomly into 17 pools of six isolates and one pool of five isolates. The 52 remaining samples were selected from recent isolates from paratyphoid patients in Delhi, Kolkata and Karachi. Isolates were provided by Shanta Dutta (National Institute for Cholera and Enteric Diseases, Kolkata, India), Rajni Gaind (Safdarjung Hospital, Delhi, India) and Rumina Hasan (Aga Khan University Hospital, Karachi, Pakistan), and DNA was extracted by Satheesh Nair (Sanger Institute). These samples were grouped into seven pools of six isolates and two pools of five isolates. A full list of isolates in each pool is give in Appendix B.

Pooled DNA samples containing 400 ng of DNA from each of 5-6 isolates were sequenced in the Sanger Institute Solexa sequencing pipeline. Reads of 35 bp were mapped to the Paratyphi A AKU_12601 reference genome sequence using Maq as described above. The mean number of reads generated per pool was 5.4 million, of which 84-98% mapped to the reference genome. The mean read depth across the genome was 40 reads per base. Details of data generated from each DNA pool are shown in Table 3.6.

	No.	**No.**	**% Reads**	**Read depth**	**Read depth**
Pool	**Isolates**	**Reads**	**Mapped**	**(pool)**	**(per isolate)**
JW1	5	3920446	92	28.1	5.6
JW2	6	2898178	84.2	18.8	3.1
JW3	6	3837653	93.5	28	4.7
JW4	6	3783920	88.8	26.2	4.4
JW5	6	2616143	91.4	18.6	3.1
JW6	6	5559046	94.8	41.3	6.9
JW7	6	5699385	94	42	7.0
JW8	5	5210321	95	38.8	7.8
JW9	6	10523926	94.6	77.5	12.9
MA1	6	5040001	92.3	36.2	6.0
MA2	6	4963803	96.5	37.3	6.2
MA3	6	6583526	98.2	50.3	8.4
MA4	6	6575035	96.5	49.3	8.2
MA5	6	6778178	97	51.1	8.5
MA6	6	6273470	96.4	47	7.8
MA7	6	4330468	96.2	32.4	5.4
MA8	6	4962394	96	37.1	6.2
MA9	6	5979484	95	44.2	7.4
MA10	6	5993929	95.4	44.5	7.4
MA11	6	5992751	97.2	45.3	7.6
MA12	6	6157760	97.2	46.5	7.8
MA13	6	5078847	90.4	35.7	6.0
MA14	6	6015837	95.1	44.5	7.4
MA15	5	4920549	92.5	35.4	7.1
MA16	6	4992161	95.1	36.9	6.2
MA17	6	4839894	96.2	36.2	6.0
MA18	6	5736325	97.6	43.6	7.3

Table 3.6: Solexa sequence data for Paratyphi A pools - Each pool contains 5 or 6 isolates as indicated. Other columns indicate the total number of reads sequenced, the percentage of reads that mapped to the AKU_12601 reference genome, and the average depth of mapped reads across the reference genome.

3.3.3.1 SNP detection

For each pool in Table 3.6, SNP detection and frequency estimation was performed as optimised above (3.3.2.3). Frequency estimates were summed across all pools p to generate, for each SNP k, the estimated frequency (and 95% confidence interval of this estimate) among all 159 isolates:

$$freq_k = \sum_{p=1}^{27} freq_{k,p} \tag{3.14}$$

$$95\%CI(freq_k) = (\sum_{p=1}^{27} lower_{k,p}, \sum_{p=1}^{27} upper_{k,p}) \tag{3.15}$$

The optimised method of SNP detection used here to identify SNPs in the Paratyphi A pools (3.3.2.3) excludes reads mapping to the reference sequence with more than two mismatching base pairs. Thus if multiple true SNPs were present in a cluster, reads covering these SNPs may not be mapped and therefore the SNPs would not be identified in the resulting sequence assembly. To check whether any clustered SNPs had been excluded from the analysis, SNP calling was repeated using Maq with default settings, which allows reads to be mapped to the reference sequence with up to seven mismatching bases. The resulting SNP calls in regions with at least 10x read depth were compared to those detected by the more stringent analysis described above. This yielded 35 SNPs that were not detected previously and were not in repetitive regions. Read alignments at these loci were examined manually to exclude SNP calls that were clearly due to poor mapping, resulting in seven SNP pairs all lying within coding sequences (see Table 3.7).

An additional 61 SNPs were identified in the comparison of seven individual genomes (3.3.1) that were not identified among the 27 pools of Paratyphi A isolates. These include 24 SNPs that were identified in isolate C1468 (not included in any experimental pools) and 24 SNPs that were detected in isolate BL8758 (present in pool JW4, sequenced at relatively low read depth (26x)); the remaining 13 SNPs had each been detected in just one isolate (3.3.1). Five of the individually sequenced genomes were included in the pools. Of the 352 SNPs previously detected between these five genomes, 315 (89.5%) were successfully detected among pooled data, with a mean estimated frequency of 29 isolates (range 1-157). Over 75% of the SNPs identified initially as unique to isolates C1468 or 38/71 were detected within the pool data, despite the absence of C1468 and 38/71 from the pools.

Position	Gene (product)	Codon	Alleles	Pool	No. Isolates
406495	*ratB*	n/a	T, C	MA13	2
406498	(pseudogene)		A, C	MA13	2
712315	SSPA0595	239,240	A, C	JW6	3
712316	(transporter)		A, C	JW6	3
766683	SSPA0643	262,264	C, A	MA15	2
766688	(lactate dehyrogenase)		A, C	MA15	2
909314	*pduT*	134,135	T, C	MA6	1
909315	(propanediol utilisation)		A, C	MA6	1
3281286	SSPA2966	2,3	G, T	JW2	2
3281290	(putative exported)		G, T	JW2	2
3716916	SSPA3354	85	T, G	JW8	2
3716918	(DNA ligase)		A, G	JW8	2
4433754	SSPA3963	287,288	A, C	MA1	2
4433758	(carbamate kinase)		G, C	MA1	2

Table 3.7: SNP clusters detected in Paratyphi A pools - Position is genomic coordinate in AKU_12601. AA residue indicates which codon(s) are affected by each SNP. No. isolates indicates the estimated frequency of the SNP pair within the given pool.

3.3.3.2 Distribution of SNPs among pools

A total of 7,364 chromosomal SNPs were detected with an estimated frequency of ≥1 isolate across all 27 experimental pools. The mean number of SNPs identified per pool was 1,152 (range 297-3,386) and the mean number of SNPs unique to each pool was 214 (range 21-2,638). The distribution of SNP calls across pools is summarised in Figure 3.13. Two pools stand out as containing exceptionally large numbers of SNPs, including large numbers of SNPs unique to the pool (MA6 and MA10). MA6 contained 2,586 unique SNPs with a frequency of one isolate, and MA10 contained 739 such SNPs. These unique SNPs were randomly distributed in the Paratyphi A genome (see Figure 3.14) and were therefore unlikely to be the result of homologous recombination with another serovar. To investigate the possibility of contaminants

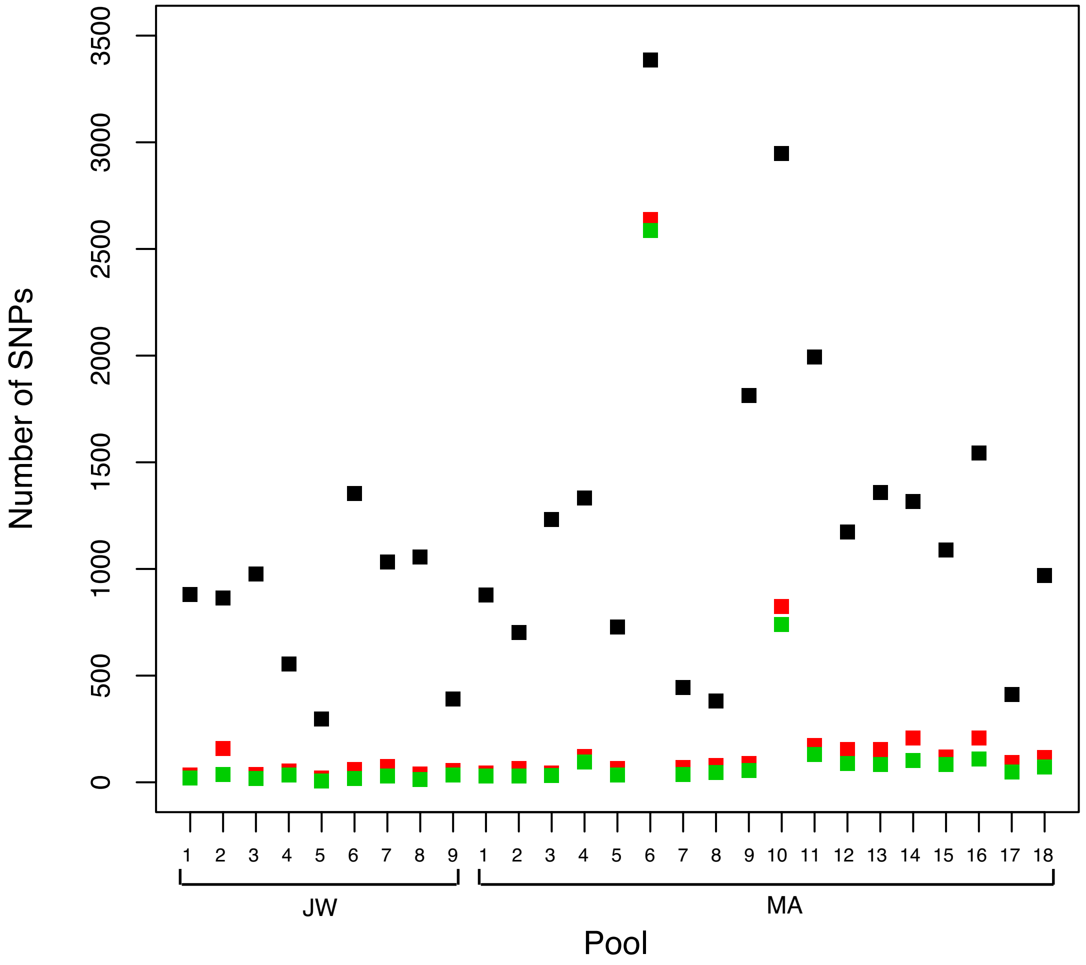


Figure 3.13: Numbers of SNPs detected in each Paratyphi A pool - Numbers are based on SNP detection analysis in all Paratyphi A pools using optimised methods. Total SNPs (black) = total number of SNPs detected at any frequency within the pool. SNPs unique to pool (red) = total number of SNPs detected within this pool but not detected in any other pool. SNPs unique to one isolate (green) = total number of SNPs detected in this pool but no other pools, with an estimated frequency of one isolate. Note that the pools known to contain contaminants (pools MA6 and MA10) have unusually high numbers of total SNPs and single-isolate unique SNPs.

within pools MA6 and MA10, isolates in these pools were re-serotyped by Francois-Xavier Weill at the Pasteur Institute, Paris, France. He discovered that isolate 9-63, part of pool MA6, was of the wrong serotype (04;Hd) and therefore not Paratyphi A (O1,2,12;Ha). Further investigation suggested the isolate 9-63 included some Paratyphi A bacteria but was contaminated with another *Salmonella* serotype. MLST analysis (performed by Dr Weill) indicated that strain WS0065 in pool MA10 had a different sequence type (ST479) compared to all other Paratyphi A strains in the pools (ST85). ST479 differs from ST85 by two SNPs, one each within the *aroC* and *hisD* loci (the *aroC* SNP was detected within pool MA10). SNPs called uniquely in pool MA6 (2,586 SNPs) or MA10 (739 SNPs) were therefore excluded from further analysis.

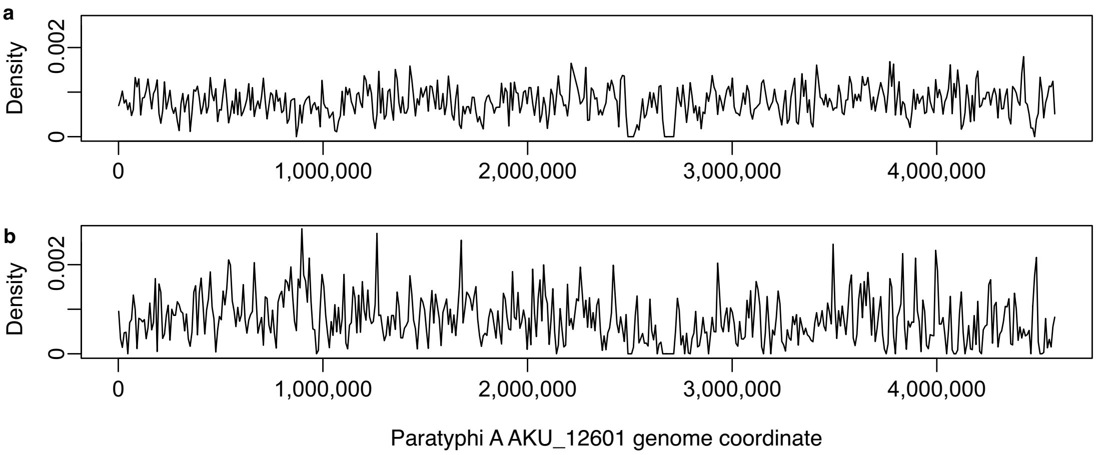


Figure 3.14: Distribution of SNPs detected uniquely in pools MA6 and MA10. - Distribution in the Paratyphi A genome of SNPs detected uniquely in MA6 (a) and MA10 (b) pools, with an estimated frequency of one strain. If the SNPs were caused by recombination from another serovar, we would expect there to be clusters of SNPs in regions where recombination has occurred. Note that SNPs within phage and repetitive sequences have been filtered out.

3.3.3.3 Distribution of SNP frequencies

Since SNPs were called in the Paratyphi A pools by comparison to the reference genome AKU_12601, it was not immediately obvious which was the ancestral allele and which was the derived allele at each SNP position. In order to determine this, alignments of all available *Salmonella* reference genomes were checked, to determine which of the two alleles was likely to be the ancestral allele at each SNP position identified among the Paratyphi A pools. This analysis was performed by Camila Mazzoni using multiple alignments produced using Kodon (Bionumerics). This information was used to convert the pool-wide detection frequencies of each SNP into pool-wide frequencies of the derived allele. The derived allele frequency is a more useful measure of frequency within the pools, as it better reflects the age of the substitution mutation responsible for the SNP. Consider for example a SNP that was detected in 150 isolates compared to AKU_12601. If the 150 isolates in which the SNP was detected actually carry the ancestral allele, while only AKU_12601 and a few other isolates have the derived allele, the frequency of the derived allele is actually nine rather than 150 isolates. The frequency of this SNP is equivalent to one that was detected in just nine isolates compared to AKU_12601, where AKU_12601 carries the ancestral allele and nine isolates carry the derived allele. The distribution of derived allele frequencies within the pools is shown in Figure 3.15. A total of 2,048 SNPs (43.0%) had an estimated frequency of just one isolate.

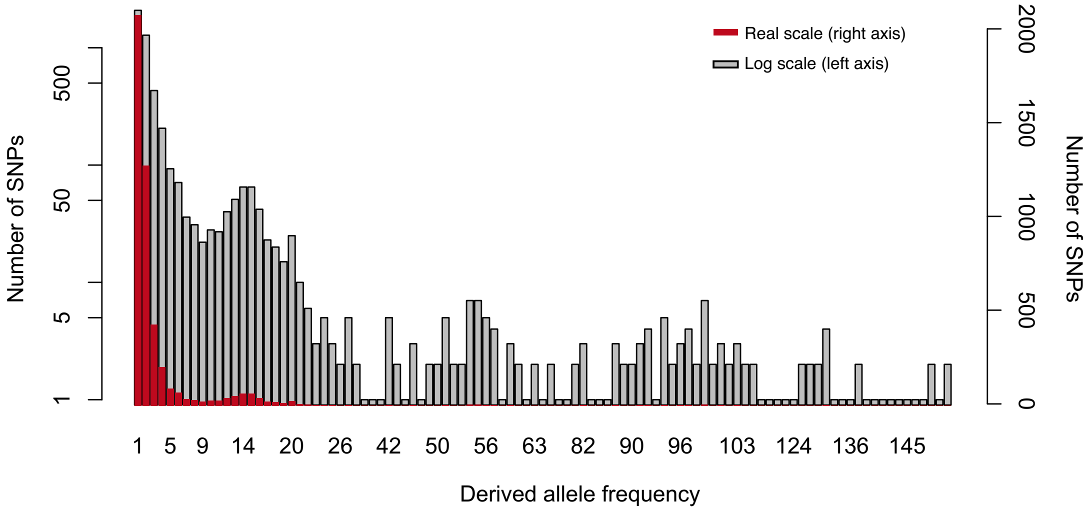


Figure 3.15: Distribution of estimated Paratyphi A SNP frequencies - Note that the same distribution is plotted on a log scale (grey bars, left axis) and non-log scale (red bars, right axis).

Figure 3.16 shows the distribution of frequencies within the pooled data, for 403 SNPs that were first identified among seven genome sequences (3.3.1) and used to determine the phylogeny in Figure 3.1). In order to investigate further, pool-wide frequencies of the SNPs defining each internal branch of the seven-genome phylogenetic tree were examined separately. Figure 3.17 shows the phylogenetic tree, and the range of frequencies for SNPs on each internal branch. SNPs defining the inner-most branches were frequent within the pools (58-83 and 39-63 isolates), while SNPs defining branches further from the root were rarer (1-24 strains, 6-28 strains and 3-20 strains).

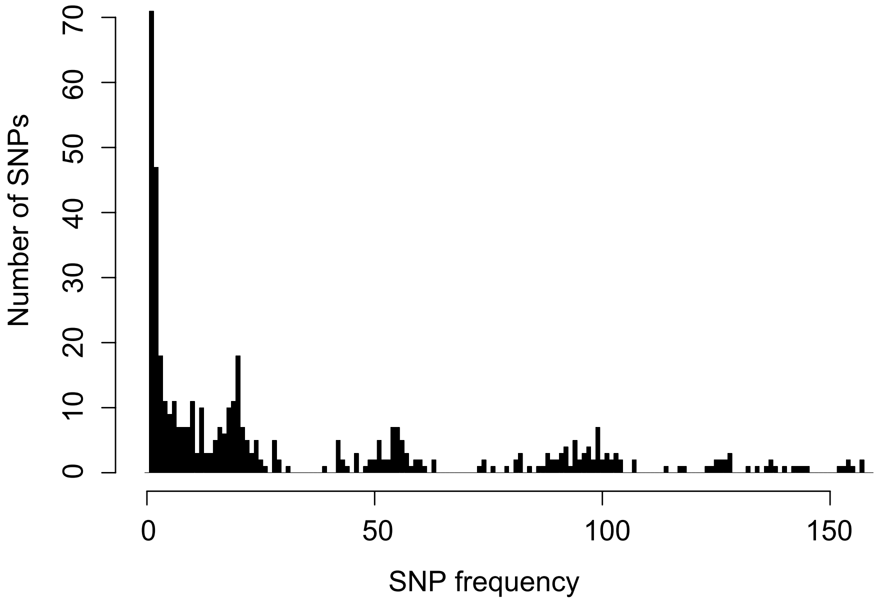


Figure 3.16: Distribution of frequencies across pools for SNPs originally detected among seven individually-sequenced Paratyphi A isolates - Frequencies shown are for 403 SNPs originally detected among seven individually-sequenced isolates.

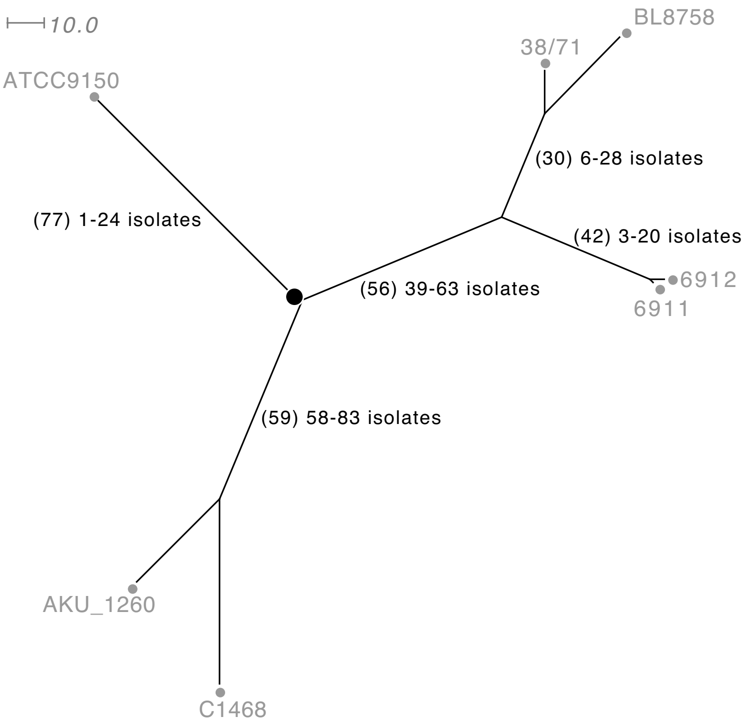


Figure 3.17: Pool-wide frequencies of SNPs defining different branches of the seven-strain phylogenetic tree of Paratyphi A - Scale bar is 10 SNPs. The number of SNPs defining each branch is shown in brackets before the pool-wide frequencies for those SNPs

3.3.3.4 Distribution of SNPs in the Paratyphi A genome

SNPs appeared to be randomly distributed in the genome, see Figure 3.18a. Among all 161 isolates sequenced either individually or in pools (excluding SNPs detected uniquely in contaminated pools MA6 and MA10), a total of 4,852 SNPs were identified, or 1 SNP per 897 bp of non-repetitive genome sequence. The distance between SNPs followed an exponential distribution with mean $\sim$ 897 bp, consistent with a random distribution of SNPs in the genome (Figure 3.18b). However SNPs were more common in non-protein-coding sequences (mean 0.135% divergence), with only 83.7% of SNPs in protein-coding sequences (mean 0.082% divergence) which make up 89.1% of the non-repetitive AKU_12601 genome (χ^2 test, p<2 x 10^{-15}).

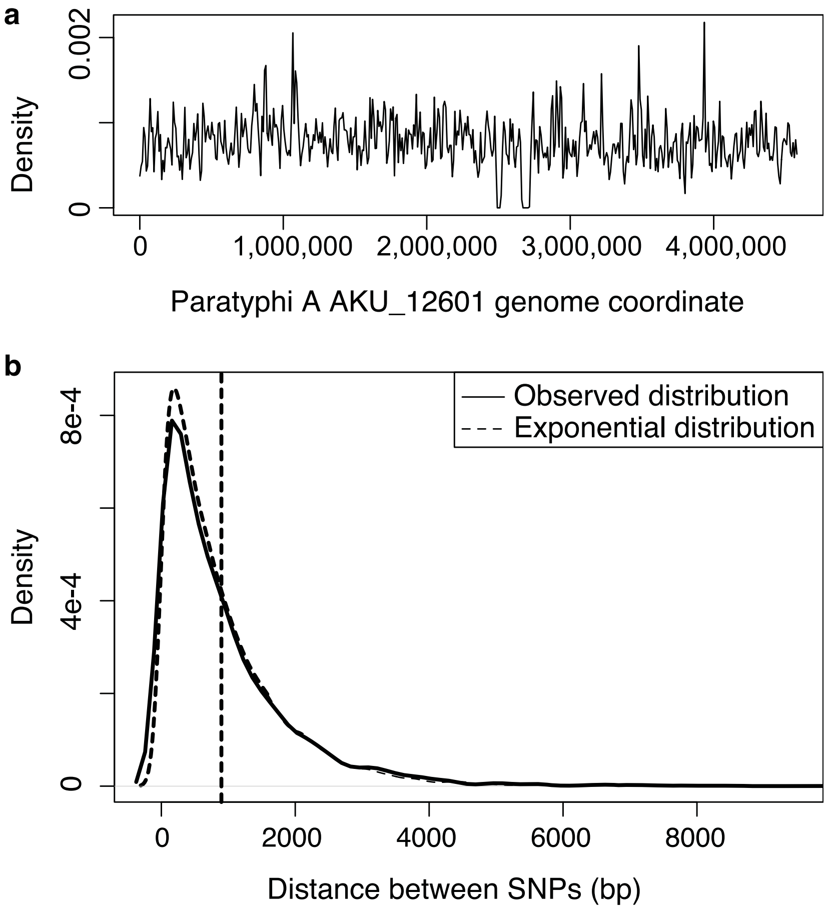


Figure 3.18: Distribution of SNPs within the Paratyphi A genome - (a) Distribution of SNPs in the genome, including those in C1468 and 38/71 and excluding SNPs detected uniquely in pools MA6 or MA10. (b) Distribution of distances between SNPs. Vertical dashed line shows the mean distance, 897 bp.

The $\frac{dN}{dS}$ across all SNPs was 0.68. Given previous observations that $\frac{dN}{dS}$ within a population tends to decrease over time (526), it might be predicted that $\frac{dN}{dS}$ would be associated with SNP frequency, since rare SNPs reflect recent mutations while frequent SNPs have presumably been conserved. However plotting $\frac{dN}{dS}$ against SNP frequency revealed no such association (Figure 3.19).

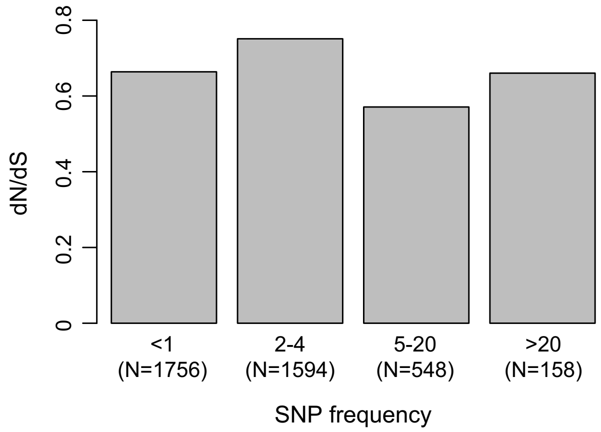


Figure 3.19: $\frac{dN}{dS}$ plotted against SNP frequency in Paratyphi A - $\frac{dN}{dS}$ values calculated for SNPs with different frequency ranges.

Figure 3.20a shows the number of SNPs per gene in the Paratyphi A genome, which suggests that SNPs are randomly distributed among genes according to an exponential distribution, with a few exceptions in the form of genes containing >10 SNPs. If SNPs were randomly distributed among genes, the expected number of SNPs for a given gene g of length l_g would be $\frac{l_g}{897}$. Figure 3.20b shows a plot of gene length vs number of SNPs per gene for all genes that contained SNPs. A total of 172 genes contained $\geq$2 SNPs more than expected (Appendix C); 11 contained $\geq$5 more than expected, some of which had high $\frac{dN}{dS}$ ratios (Table 3.8). The overrepresentation of SNPs in these genes may be a signal of diversifying selection, as they are more variable than the rest of the genome, although this could also be the result of genetic drift. Gene ontology analysis of the 172 genes containing $\geq$2 SNPs more than expected revealed an enrichment of signal transducer activity (16 genes, or 11.3% of the list vs 3.5% of genes in the genome; p=0.00277).

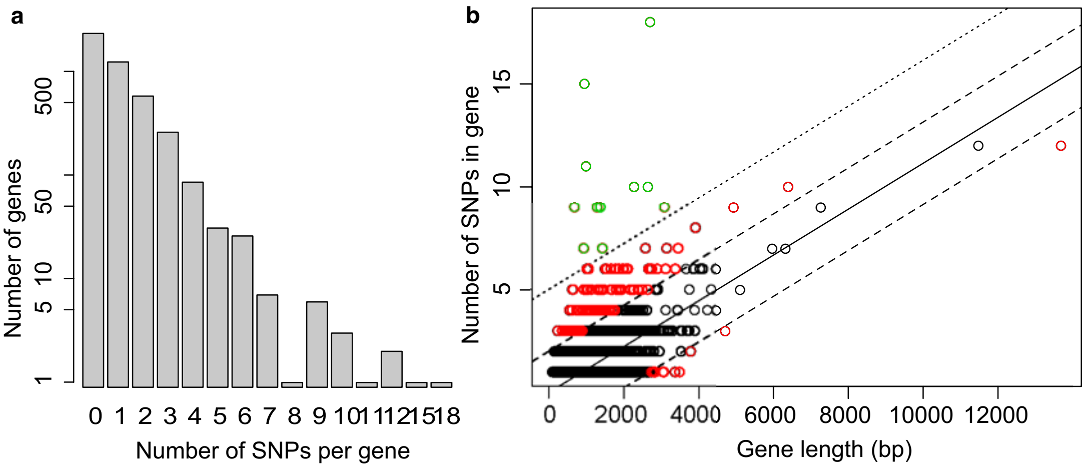


Figure 3.20: Number of SNPs per gene in Paratphi A from pools - (a) Distribution of number of SNPs per gene. (b) Gene length vs. number of SNPs. Solid line shows expected number of SNPs as a function of gene length; dashed line = 2 SNPs more or less than expected based on gene length (red data points); dotted line = 5 SNPs more than expected based on gene length (green data points).

There was also an enrichment for genes in the *wba* O-antigen biosynthesis cluster: four of the 17 genes (*wbaF, wbaX, wbaU* and *wbaP*; 23.5%) contained $\geq$2 SNPs more than expected, compared to just 4% of genes in the AKU_12601 genome (χ^2 test, $p<0.0005$). The entire *wba* cluster appears to be enriched for SNP variation. Overall, 39 SNPs were identified in 14 of the *wba* cluster genes, including 26 nonsynonymous SNPs affecting all 14 genes. Across the genome, 52.6% of genes contained at least one SNP, compared to 14 genes (82.3%) of the 17-gene *wba* cluster. This a significant enrichment according to the χ^2 test ($p=0.027$). The cluster is 18,858 bp in length, in which only 21 SNPs would be expected by chance given a random distribution with mean 1 SNP per 897 bp. Since haplotypes cannot be assigned to individual isolates using the pool data, it is difficult to test directly whether this variation is the result of horizontal transfer of genes from an external source. However, if SNPs were introduced via horizontal transfer of DNA, they should have similar patterns of distribution among the pools. The distribution of *wba* cluster SNPs among pools (Figure 3.21) highlights just two pairs of correlated SNPs which could be evidence of horizontal transfer. One pair of SNPs lay in *wbaF* (SSPA0728, nonysonymous) and *wbaM* (SSPA0737, synonymous) and were

Gene	Product	SNPs	N	dN/dS
SSPA0696	Putative RND-family transporter protein	9	4	0.27
wbaP	Undecaprenyl-phosphate galactosephosphotransferase	7	5 (0)	0.83
proQ	ProP effector	9	8 (2)	2.67
clpA	ATP-dependent Clp protease ATP-binding subunit	10	8 (1)	1.33
rpoS	RNA polymerase sigma subunit RpoS (sigma-38)	11	10 (0)	3.33
SSPA2620	Outer membrane usher protein	10	7 (0)	0.78
SSPA2639	Putative serine transporter	9	6 (0)	0.67
malT	Transcriptional regulator of maltose system	18	16 (1)	2.67
SSPA3531	Magnesium and cobalt transport protein	15	2 (0)	0.05
cpxA	Two-component sensor kinase protein	9	8 (0)	2.67
SSPA3963	Carbamate kinase	7	5 (0)	0.83

Table 3.8: Genes containing at least five more SNPs than expected by chance - These genes are highlighted in green in Figure 3.20b. SNPs = total number of SNPs detected within the gene sequence, N = number of nonsynonymous SNPs with number of nonsense SNPs given in brackets.

detected at high frequency in all pools. These SNPs are over 11 kbp apart and are consistent with recent random mutations in the AKU_12601 lineage. The other pair lay in *wbaX* (SSPA0733, nonsynonymous) and *wbaV* (SSPA0734, nonsynonymous) and were detected in nearly all pools with a frequency of roughly half the isolates. The SNPs are 861 bp apart and lie in the region *wbaXVU* that is subject to variable number tandem duplications in Paratyphi A (63). The *wbaV* SNP was present in one of the three copies in ATCC9150 and the *wbaX* SNP was not present in ATCC9150. This makes it difficult to interpret the pattern of these SNPs among the pools, as the frequency estimates are likely to be affected by tandem duplications. However, given that AKU_12601 and ATCC9150 differ at just one of these loci, these SNPs are quite unlikely to have been acquired during a horizontal transfer event. Thus the variation in the *wba* cluster of Paratyphi A is likely the result of diversifying selection by *de novo* mutation.

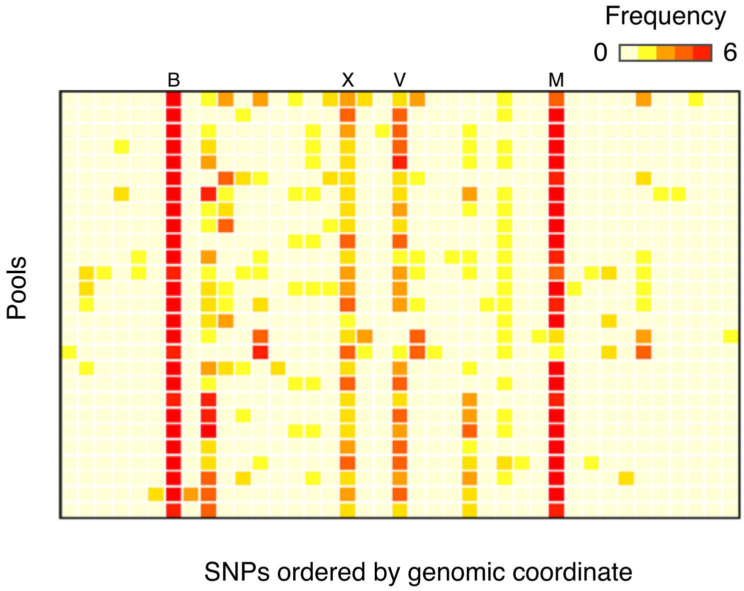


Figure 3.21: Distribution of *wba* cluster SNPs in Paratyphi A pools - B, X, V, M indicate SNPs in genes *wbaB, wbaX, wbaV, wbaM.*

3.3.3.5 Novel pseudogene-forming mutations

A total of 158 nonsense SNPs were detected, introducing stop codons within the coding sequence of 147 genes (see Appendix D). Thirteen of these genes were already pseudogenes in Paratyphi A, and the additional nonsense SNPs should be considered secondary mutations. In total, 153 genes were differentially inactivated in a subset of the Paratyphi A isolates via nonsense SNPs or deletions (see Appendix D). As described previously (2.3.4.2), these novel pseudogenes may be the result of adaptive selection by gene loss (negative selection) or simply tolerance for the loss of genes whose functions are no longer necessary in the host-restricted niche. The genes containing nonsense SNPs were generally more divergent than those without nonsense SNPs, with mean divergence 0.259% compared to 0.191% among all genes containing SNPs (T-test, p = $9.5\text{x}10^{-5}$). Again this could be explained by either negative or neutral selection pressures. Gene ontology analysis of these 'variable' pseudogenes found that genes encoding protein kinases were overrepresented (N=6, p=0.007) as were those involved more generally in two-component signal transduction systems (N=9, p=0.038). It was not possible to reliably detect frameshift mutations from the short reads data, but given that comparison of the finished AKU_12601 and ATTC9150 genomes identified twice as many frameshift mutations as nonsense SNPs (3.3.1.5), it's likely that there are many more variable pseudogenes present within the Paratyphi A pools.

3.3.3.6 Detection of IncHI1 plasmids

Maq was used to align reads from each Paratyphi A pool to the finished sequence of IncHI1 multidrug resistance plasmid pAKU_1 (described in detail in Chapter 5). The plasmid was detected in six pools, with 64%-100% coverage of the reference sequence (Table 3.9). IncHI1 plasmid sequence was identified in all of the pools known to contain multidrug resistant isolates. IncHI1 plasmids are maintained in *Salmonella* at an average copy number of one plasmid per cell, based on Solexa coverage data shown in Table 3.10 and personal communication with John Wain (Sanger Institute/Health Protection Agency). The number of isolates containing the plasmid in each pool x ($\mathrm{N}_{p,x}$) was therefore estimated from the ratio of the mean depths of reads mapping to the plasmid and chromosome sequences:

$$N_{p,x} = \frac{d_{p,x}}{d_{c,x}} * N_x \tag{3.16}$$

where N_x is the number of isolates in pool x, $d_{p,x}$ is the mean depth of reads in pool x mapping to the plasmid and $d_{c,x}$ is the mean depth of reads in pool x mapping to the chromosome. Only one or two isolates per pool were estimated to contain the plasmid, resulting in a total of nine isolates across six pools. The pool containing AKU_12601 and pAKU_1 (JW2) had 94% coverage of pAKU_1 and was estimated to contain two isolates harbouring the plasmid.

SNPs in the IncHI1 plasmid were detected using the same methods validated for the chromosomal SNPs. Here the number of plasmids estimated per pool ($\mathrm{N}_{x,p}$ in equation 3.16), rather than the total number of isolates per pool, was provided as the expected number of haplotypes (`maq assemble` option `-N`). A total of 53 SNPs were identified across the estimated nine plasmids in six pools, of which 16 were not in repetitive sequences. Each of these 16 SNPs was detected in just one pool, with estimated frequencies of 1-2 isolates. Parsimony splits analysis was used to construct a phylogenetic network, resulting in the phylogenetic tree shown in Figure 3.22 (see Methods 3.2.3). R27 and pHCM1 alleles were included to root the tree; they fell into a single node along with the pAKU_1 reference alleles, the pool JW5 which included AKU_12601

Pool	Isolates	pAKU_1 coverage	Depth ratio	Plasmid isolates
JW1	5	1%	0.01	0
JW2	6	94%	1.82	2
JW3	6	1%	0.01	0
JW4	6	98%	1.63	2
JW5	6	64%	0.41	1
JW6	6	1%	0.01	0
JW7	6	1%	0.01	0
JW8	5	1%	0.01	0
JW9	6	100%	2.25	2
MA1	6	1%	0.01	0
MA2	6	1%	0.01	0
MA3	6	1%	0.02	0
MA4	6	1%	0.02	0
MA5	6	98%	0.94	1
MA6	6	1%	0.01	0
MA7	6	3%	0.04	0
MA8	6	78%	0.32	1
MA9	6	1%	0.02	0
MA10	6	1%	0.02	0
MA11	6	1%	0.01	0
MA12	6	1%	0.02	0
MA13	6	1%	0.02	0
MA14	6	1%	0.02	0
MA15	5	5%	0.03	0
MA16	6	2%	0.04	0
MA17	6	2%	0.03	0
MA18	6	1%	0.02	0

Table 3.9: IncHI1 plasmids detected in pools - Isolates = total number of isolates in each pool; pAKU_1 coverage = coverage of the 212 kpb IncHI1 plasmid sequence from reads data; Depth ratio = ratio of mean depth of reads mapping to pAKU_1 and mean depth of reads mapping to the chromosome, multiplied by the number of isolates; Plasmid isolates = estimated number of isolates that contain a pAKU_1-like IncHI1 plasmid, based on this ratio and assuming plasmid copy number of no more than 1 per cell.

and therefore pAKU_1, and one other pool. None of the IncHI1 SNPs identified here were identified in earlier comparisons of pAKU_1 with IncHI1 plasmids found in Typhi (see 5.3.2.1). This, together with the tree structure, suggests that all of the IncHI1 plasmids identified here in Paratyphi A are closely related plasmids with a recent common ancestor and distinct from those found in Typhi.

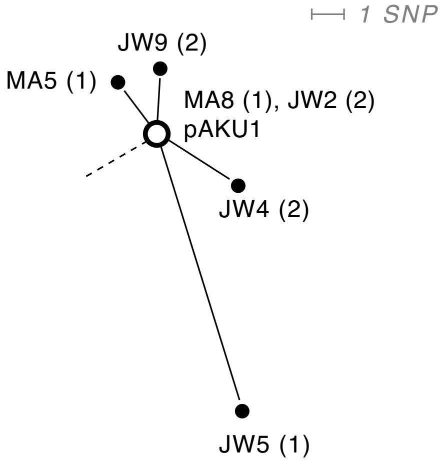


Figure 3.22: IncHI1 SNPs detected in Paratyphi A pools - Split network based on 16 IncHI1 SNPs detected in Paratyphi A pools. Plasmid sequences R27 and pHCM1 were used as outgroups to root the tree; the open circle represents this root.

Isolate	Plasmid:Chromosome Read Depth
Paratyphi A AKU$_1$2601	1.4
Typhi CT18	1.4
Typhi E03-9804	0.88
Typhi ISP-03-07467	0.90
Typhi ISP-04-06969	0.89

Table 3.10: Ratio of read depths for IncHI1 plasmids and *Salmonella* chromosomes - Ratios of the average depth of reads mapped to IncHI1 plasmid sequences and Typhi or Paratyphi A chromosomes, using Maq 0.6 with default parameters.

3.3.3.7 Detection of plasmid pGY1

Maq was used to align reads from each Paratyphi A pool to the finished sequence of plasmid pGY1 (287). The plasmid was detected in nine pools, with 100% coverage of the reference sequence (Table 3.11). The pGY1 plasmid is very small (3,592 bp) and appears to be maintained at high and variable copy number in *Salmonella* Paratyphi A cells (based on high ratio of read depths covering plasmid vs chromosome for isolate C1468, and data in Table 3.11). It was therefore not possible to estimate the number of isolates per pool which contained the plasmid. SNPs in the pGY1 plasmid were analysed in nine Paratyphi A pools containing the plasmid using Maq with default parameters. SNP calls were filtered to exclude those with read depth ≤ 10 or quality score ≤ 20 (which removed only 3 SNP calls). A total of 23 SNPs were identifed, across five pools (no SNPs were identified in pGY1 sequences within pools JW1, JW2, JW7 and MA1). The presence of SNPs in each pool was encoded as 0 (not detected) or 1 (detected) and the resulting table used to build the phylogenetic network shown in Figure 3.23.

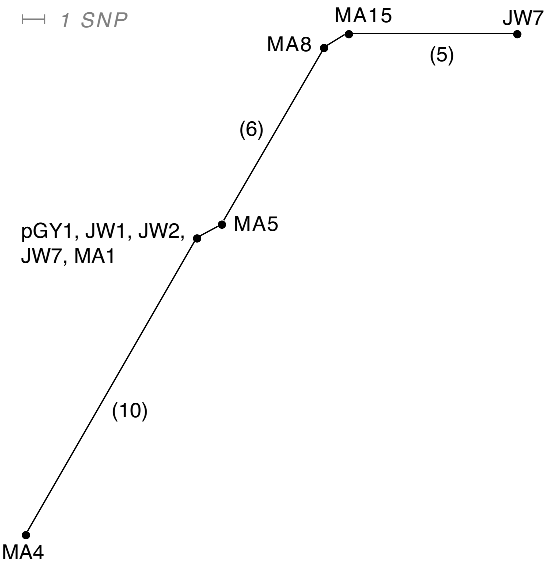


Figure 3.23: Phylogenetic network of pGY1 plasmids detected in Paratyphi A pools - Note that the pGY1 reference sequence matches identically pGY1 sequences in four pools. Branch lengths represent the number of SNPs; scale bar is one SNP; branches longer than one are labelled with the number of SNPs in brackets.

Pool	No. isolates	pGY1 coverage	pGY1 depth	Depth ratio
JW1	5	100%	352	63
JW2	6	100%	176	56
JW3	6	4%	0	0
JW4	6	100%	45.4	10
JW5	6	0%	0	0
JW6	6	2%	0	0
JW7	6	100%	299	43
JW8	5	0%	0	0
JW9	6	3%	0	0
MA1	6	100%	69	11
MA2	6	12%	5	1
MA3	6	3%	0	0
MA4	6	100%	75	9
MA5	6	100%	98	11
MA6	6	29%	19	2
MA7	6	12%	3	0.5
MA8	6	100%	68	11
MA9	6	5%	0	0
MA10	6	4%	0	0
MA11	6	1%	0	0
MA12	6	0%	0	0
MA13	6	5%	0	0
MA14	6	3%	0	0
MA15	5	100%	107	15
MA16	6	8%	0	0
MA17	6	24%	6.8	1
MA18	6	2%	0	0

Table 3.11: pGY1 plasmids detected in Paratyphi A pools - No. isolates = total number of isolates in each pool; pGY1 coverage = coverage of the pGY1 plasmid sequence from reads data; pGY1 depth = mean depth of reads mapping to pGY1; Depth ratio = ratio of mean depth of reads mapping to pGY1 and mean depth of reads mapping to the chromosome, multipled by the number of isolates.

3.4 Discussion

3.4.1 Strengths and limitations of the study

Given the lack of phylogenetic information available for Paratyphi A, it was difficult to avoid discovery bias in this study. The choice of seven isolates for whole-genome phylogenetic and comparative analysis (3.3.1) was essentially random, although isolates were chosen from four different regions (Karachi, Kolkata, Delhi and Nairobi), and exhibited some phenotypic variation (Table 3.1). Although the phylogenetic tree of these sequenced isolates was balanced (Figure 3.1), it is still possible that they reflect only a subset of the Paratyphi A population. Discovery bias is likely to be less of an issue in the global screen of genomic sequence (3.3.3), in which over 150 isolates were selected from as broad a range of geographical regions, time periods and phage types as possible.

Pool-wide frequencies of SNPs used to build the phylogenetic tree for seven isolates (Figure 3.1) were consistent with the hypothesis that the subpopulations sampled by the seven isolates and the pools were largely overlapping. The phylogenetic tree shown in Figure 3.1 splits the population into three lineages, ATCC9150 in one lineage, AKU_12601 and C1468 in a second, and the remaining isolates in a third. The SNPs defining these lineages had estimated frequencies of 1-24 strains, 58-83 strains, and 39-63 strains respectively (see Figure 3.17). The range of frequencies likely reflects diversification along each of these branches, with the higher frequency SNPs closer to the root. The numbers sum approximately to 161, the total number of isolates sampled, and suggest that ∼20 of the isolates in the pool population belong to the ATCC9150 lineage, ∼80 belong to the AKU_12601 lineage and ∼60 belong to the third lineage. If the seven individually sequenced isolates represent a biased subpopulation of Paratyphi A while the larger collection of isolates sequenced in pools represented more of the underlying population (illustrated in Figure 3.24), then the most recent common ancestor of the pooled isolates would be older than the most recent common ancestor of the seven isolates (the tree root in Figure 3.1). In this case, lineages that diverged earlier than the seven sequenced isolates would not contain any of the SNPs detected among those seven isolates (see Figure 3.24) and the pool-wide frequencies of SNPs defining the three known lineages would sum to less than the total number of isolates represented in the pools. Since the pool-wide frequencies of these SNPs do sum to the total number of isolates represented in pools, it is likely that the position of the root in Figure 3.1 approximates the most recent common ancestor of all the isolates sequenced in pools. This in turn suggests that conclusions about the scale of differences between lineages identified from the analysis of the individually sequenced isolates should be generalisable to differences between Paratyphi A lineages in general.

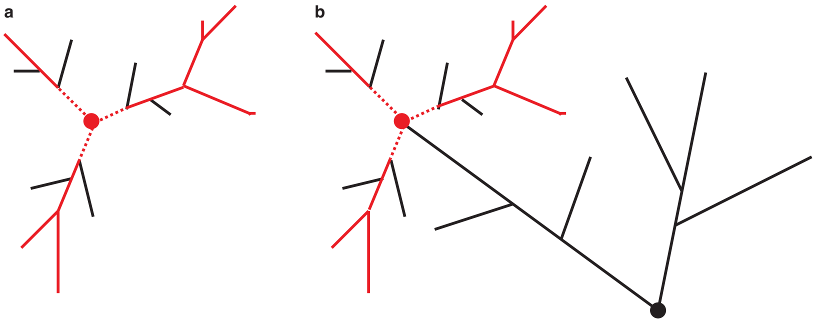


Figure 3.24: Difference in SNP frequencies given biased and unbiased sampling - Red branches indicate the phylogenetic tree of seven individually sequenced Paratyphi A isolates. Black branches indicate hypothetical branches described by the wider population of Paratyphi A sampled in the pools, under two different scenarios. (a) A scenario in which the seven genomes represent an unbiased sample of the Paratyphi A population, such that their most common ancestor is also the most common ancestor of the wider population. Under this scenario, each of the genomes in the pooled population must carry the SNPs defined by one of the dashed branches. (b) A scenario in which the seven genomes represent a biased sample of one part of the Paratyphi A population, such that their most recent common ancestor is much younger than the most recent common ancestor of the broader population. Under this scenario, many of the genomes in the pooled population would carry none of the SNPs defined by the dashed branches. Note in the real data, the frequencies of the SNPs on the dashed branches sum to the total number of pooled genomes, consistent with (a) but not (b).

A total of 352 SNPs were detected among five genomes which were also included in the pools (AKU_12601, ATCC9150, 6911, 6912 and BL8758). Of these, 315 SNPs (89.5%) were successfully detected among pooled data, with a mean estimated frequency of 29 isolates (range 1-157). This is higher than the sensitivity of detection estimated from the single test pool (83%, 3.3.2.3), likely because many SNPs were present in multiple isolates and multiple pools, giving them a higher chance of detection across the whole data set. Over 75% of the SNPs identified initially only in isolates C1468 or 38/71, which themselves were not included in the pools, were nevertheless detected within the pool data. These observations suggest that the sampling of strains for pooled SNP analysis provided good coverage of the underlying population, and that the majority of SNPs that were not detected are likely to be rare, strain-specific SNPs.

Sequencing pooled DNA dramatically reduces the cost of SNP detection, enabling a larger sample of isolates to be screened for genome-wide variation. In addition to a general increase in the number of variant loci that can be detected, the increased sample size also reduces selection bias, as larger random samples should be more representative of the population than smaller ones. While it is difficult to ensure that all pooled samples will be represented equally in the sequencing data, the results of the test pool were encouraging, with 81% accuracy for frequency estimates, which were never wrong by more than one isolate. In the present study, over 150 isolates were screened in 27 pools, resulting in detection of >4,800 SNPs from just 27 lanes (less than four full runs) of Solexa sequencing. The sensitivity of SNP detection was estimated to be >82% from the single test pool (3.3.2.3) and >89% from analysis of SNPs known to be present among the 27 pools 3.3.3.1. Therefore we would probably need to individually sequence at least 80% of these isolates to detect the same level of variation (4,800 SNPs), which would require 128 lanes of sequencing, i.e. five times as many as by pooling. Large sample sizes and large numbers of SNPs are important in this study, as they provide more data to facilitate the detection of subtle patterns of selection in the Paratyphi A population. A large number of SNPs (43%) had an estimated frequency of just one isolate (see Figure 3.15), demonstrating the need for large sample sizes in order to distinguish between conserved, informative SNPs and recent, strain-specific mutations. This facilitates the selection of appropriate loci for developing SNP typing schemes.

The obvious drawback of sequencing pooled samples is that haplotypes can not be determined for individual isolates, and therefore phylogenetic inference is not possible directly from the sequence data. However, the SNPs detected from this large-scale screen can be used to develop typing assays which yield phylogenetically informative data not only for the 161 isolates used in this study, but for much larger collections. An additional drawback of the pooling method is the problem of contamination, which in this study affected two pools (MA6 and MA10). SNPs detected uniquely in these pools and with a frequency of one isolate (>90% of those called uniquely) were excluded from further analysis, which almost certainly resulted in exclusion of novel SNPs present in other isolates within the pool. This difficulty could be minimized by carefully screening isolates prior to inclusion in the pools, e.g. re-serotyping prior to DNA extraction to ensure a clonal population of the correct serotype.

3.4.2 Genomic variation and possibilities for typing in the Paratyphi A population

Figure 3.1 shows how prophage and pseudogenes were distributed around the phylogenetic tree of sequenced isolates. Treating 6911 and 6912 as one, the six lineages differed on average by 100-200 SNPs, two prophage sequences (range 0-5), two deletions (range 0-4) and four nonsense SNPs (range 2-7). The distributions of each variant are shown in Figures 3.25 and 3.26. Five variable numer tandem repeats were identified between the two finished genomes AKU_12601 and ATCC9150 (see Table 3.3) but could not be resolved for the genomes sequenced with short reads. Similarly, indels of <20 bp could not be resolved, so the number of pseudogenes that differ between isolates is likely to vary by more than those caused by nonsense SNPs; for example 22 were identified between AKU_12601 and ATCC9150 including just six nonsense SNPs.

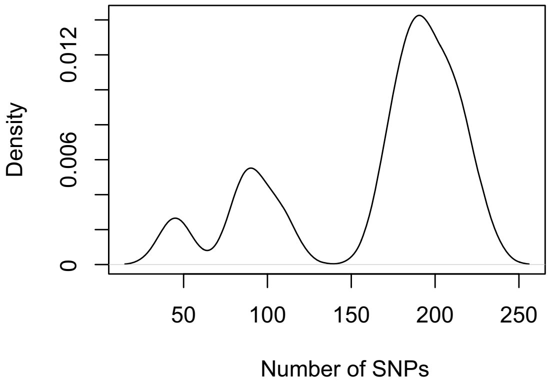


Figure 3.25: Distribution of number of SNPs between two Paratyphi A lineages - The number of SNPs between every possible pair of 6 Partayphi A lineages was calculated (treating 6911 and 6912 as a single lineage), the distribution of SNP numbers is shown.

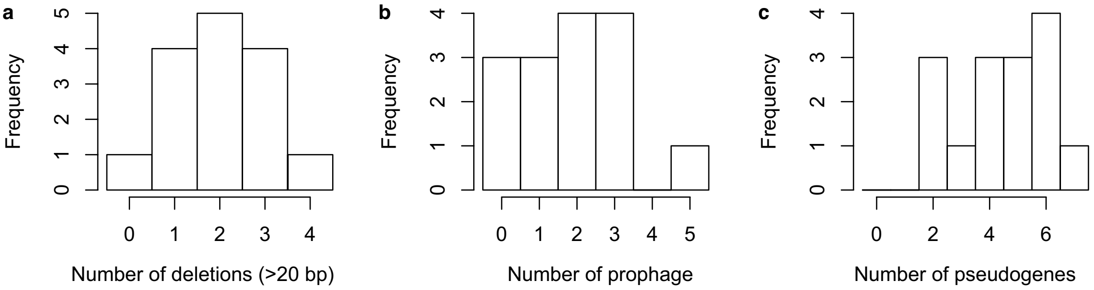


Figure 3.26: Distribution of number of deletions, prophage and pseudogenes between two Paratyphi A lineages - The number of deletions, phage insertions and pseudogenes between every possible pair of 6 Paratyphi A lineages was calculated (treating 6911 and 6912 as a single lineage). The distribution of these counts is shown for deletions, phage and pseudogenes in a-c respectively.

The Paratyphi A population was generally less diverse than the Typhi population, with lineages separated by 100-200 SNPs as opposed to 300-500 (see 3.25). This may reflect a more recent bottleneck in the Paratyphi A population, so that the most recent common ancestor of the sequenced Paratyphi A genomes is younger than the most recent common ancestor of Typhi. Given the SNP frequencies among the pooled isolates it is likely that the most recent common ancestor of the seven genomes (represented by the root in Figure 3.1) is the most recent common ancestor of the wider Paratyphi A population (see 3.4.1), so the apparent younger age of Paratyphi A is unlikely to be the effect of selection bias. Paratyphi A genomes generally differed by fewer deletions than did Typhi genomes (one vs five on average). However, although the Paratyphi A genomes contained an average of three prophage sequences (2-5 per genome) while Typhi genomes contained twice as many, the level of phage variation between isolates was equivalent (mean one prophage sequence, range 0-5). This is consistent with the hypothesis that prophages are gained and lost quite frequently in *Salmonella* genomes, so that closely related genomes can differ in their prophage complement just as much as more distantly related genomes. In any case, these observations provide further evidence that differences in prophage are not a particularly good reflection of genetic relatedness within populations of Paratyphi A or Typhi.

The low level of variation detected here among Paratyphi A genomes suggests that a highly discriminatory and phylogenetically informative typing scheme for Paratyphi

A must center around SNPs. Furthermore the SNPs detected in this study could now be used to develop the first sequence-based typing scheme for Paratyphi A. If a large number of SNPs were able to be assayed (say ≥1,000), they could be chosen randomly from those detected in this study. If a small subset of SNPs were to be assayed (say 100), they could be stratified according to frequency to ensure a mix of conserved and rarer SNPs with which to build a phylogenetic tree. In either case it would be wise to exclude SNPs with an estimated frequency of one strain, which are more likely to be phylogenetically uninformative and/or errors in SNP detection than those with higher frequency estimates, particularly those that were detected in more than one pool.

3.4.3 Adaptive selection in Paratyphi A genes

As with Typhi in Chapter 2, the very low level of nucleotide variation detected between Paratyphi A genomes makes it difficult to conclude much about selection on individual genes. Only half the genes in the Paratyphi A genome contained any SNPs at all, although it is likely that some genes harbour frameshift or other small indel mutations that could not be detected. As with the Typhi analysis, the approach of detecting selection by calculating $\frac{dN}{dS}$ or other statistics for individual genes would be inappropriate as there is not enough variation to work with. However, the distribution of SNPs per gene (Figure 3.20) suggested an overrepresentation of variation within some genes and highlighted outliers with many more SNPs than expected by chance merely as a function of gene length.

Eleven genes contained at least five more SNPs than expected by chance, suggesting they may be subject to diversifying selection (note that using the same method to analyse the Typhi SNPs presented earlier identifies only *tviE* and *yehU*). These include *rpoS*, mutations in which facilitate response to nutrient limitation (608) and *malT*, mutations in which can lead to constitutive expression of maltose metabolism genes (609). Both genes contained large numbers of SNPs and $\frac{dN}{dS}$ ratios >2.5 (Table 3.8), which may be indicative of adaptive selection in response to nutrient stress. The highly variable genes also include *proQ* (which regulates the proline transporter ProP), a serine transporter and a two-component sensor kinase protein of the OmpR family. *ProQ* and the two-component sensor had $\frac{dN}{dS}$ ratios >2.5 (Table 3.8) and variation in all three genes may be associated with adaptive selection for osmotic stress tolerance.

The 1,431 bp gene encoding WbaP, involved in the O-antigen biosynthesis pathway (610), contained seven SNPs including five that were nonsynonymous. This may contribute to variation in the O-antigen by altering the chain-length distribution of the O polysaccharide or otherwise (611, 612).

A total of 172 genes contained at least two more SNPs than expected by chance, making them potential candidates for diversifying selection, although much of this variation may be due to genetic drift. Gene ontology analysis of these genes suggested an enrichment of signal transducer activity, which may reflect subtle changes in signalling pathways, helping cells to adapt to changing environmental cues. There was also an enrichment of genes in the *wba* O-antigen biosynthesis cluster, with four of the 17 genes in the cluster containing ≥2 SNPs more than expected, and 14 containing at least one nonsynonymous SNP (3.3.3.4). These changes at the DNA level likely result in some diversification of the O-antigen polysaccharide expressed on the cell surface, which could be an adaptive response to pressure from the host immune system. This sort of variation was not observed in Typhi (Chapter 2), where only five SNPs were identified in the *wba* cluster, each in a different gene (*wbaU, wbaV, wbaH, wbaI, wbaA*) and including only two nonsynonymous SNPs (*wbaV, wbaA*).

Thus while the Typhi data showed little evidence of adaptive selection, the Paratyphi A data contained signals of diversifying selection in the O-antigen biosynthesis cluster *wba*, as well as variation in genes associated with signal transduction and stress responses which may be indicative of adaptive selection. This difference could be due to different kinds of selective pressure in Paratyphi A, or simply to the much greater sample size in the Paratyphi A study. This highlights one of the advantages of the pooled sequencing approach, which allows large isolate collections to be screened at the whole genome level, over individual sequencing of a smaller set of isolates. Although phylogenetic anlaysis is not possible using the pooled approach, it may be more sensitive to subtle variations in the population, which is particularly important in the absence of high levels of variation and recombination.

Chapter 4

Convergent evolution of Typhi and Paratyphi A

4.1 Introduction

Typhi and Paratyphi A are unusual among *Salmonella enterica* as most serovars infect a broad range of host species and cause self-limiting gastroenteritis, while Typhi and Paratyphi A infect only humans (host restricted) and cause systemic disease in the form of enteric fever which can be transmitted from person to person (host adapted) (613). While it has been claimed that Paratyphi A causes a milder disease more often associated with gastrointestinal infection than Typhi (614), there is little data to support this. Studies directly comparing Typhi and Paratyphi A infections in Egypt, Nepal and Indonesia found no significant clinical differences (279, 615, 616). There is limited evidence that the risk factors for Typhi and Paratyphi A are different, with a study in Indonesia (N=114 cases) suggesting Paratyphi A infection was independently associated with street vendor food and flooding, while Typhi infection was associated with household risk factors consistent with transmission within households (342). However studies in Nepal (N=600) and India (N=70) found no such difference (279, 615), suggesting that transmission routes are highly similar. Asymptomatic chronic carriage in the gall bladder occurs with both Typhi (309, 313, 617, 618) and Paratyphi A (311, 313). Studies reporting Paratyphi A carriage are fewer and more recent, probably associated with the rising incidence of Paratyphi A infection (280, 335, 336, 338) and increased attention on this serovar (280). Therefore, although there will certainly be differences at the molecular level, it is assumed that the mechanisms of infection and transmission in Typhi and Paratyphi A are highly similar. It is proposed that this convergence in pathogenic phenotype should be reflected to a significant degree by convergence at the genetic level, which stands in contrast to the features that each serovar shares with their host-generalist relatives.

S. enterica serovars Paratyphi B and Paratyphi C are also associated with systemic infection in humans (619, 620, 621), however they are difficult to distinguish from closely related serovars (620, 622, 623) and as a consequece their host ranges and clinical syndromes are not well characterised. However it has been reported that Paratyphi C causes a milder disease than Typhi and Paratyphi A, except in immunocompromised patients (624). Studies of Paratyphi B have been complicated by the existence of two biotypes with the same serotype, now divided into serotype Paratyphi B *sensu stricto* (unable to ferment D-tartrate: dT-) and serotype Java (able to ferment D-tartrate: dT+) (225). Paratyphi B (dT-) isolates appear to belong to a single MLEE (multilocus enzyme electrophoresis) type, contain a SopE1-phage and display a consistent pattern of expression of effector proteins, whereas Java (dT+) strains are more diverse in terms of MLEE type, phage type and expression (225). Paratyphi B *sensu stricto* (dT-) is the serotype associated with paratyphoid fever (the "systemic pathotype"), however it causes milder systemic disease than Typhi, a high rate of non-invasive gastroenteric infections in humans and has been isolated from animals (621). A fifth serovar, Sendai, was described in 1925 as the causative agent of an enteric fever outbreak in Japan (625, 626) but has been rarely reported since (one reported case in the U.S. between 1997-2004 (627)). It is of the same serotype as Paratyphi A and closely related by MLEE (628). Typhi, Paratyphi A, B and C belong to distinct serogroups (D, A, B and C, respectively (629)) and genetic lineages (628) of *S. enterica* and appear to have evolved independently, implying that their similar human-adapted pathogenic phenotypes are the result of convergent rather than divergent evolution.

While host restriction and host adaptation are well known in *S. enterica*, the mechanisms underlying these phenomena are far from clear. The concept of "host adaptation" can be understood as the ability to circulate within a specific host population, transmitted from one member of the population to another (629). For example, Typhi is transmitted directly between humans, whereas infections with Typhimurium and other *S. enterica* in humans are generally associated with transmission via the food chain (630, 631, 632), although human-to-human transmission has been documented (633, 634). Restriction to a specific host implies the lack of ability to infect other hosts (629), for example Typhi does not appear able to infect any hosts besides simians (higher primates) (33). Therefore Typhi is both human adapted and human restricted, as is Paratyphi A. It is possible to be host adapted but not host restricted, for example serovar Choleraesuis is swine adapted as it causes systemic infection in pigs that can be transmitted directly between individuals, but it is also quite virulent in humans (31, 32). On the other hand, to be host restricted but not host adapted would imply an evolutionary dead end. The relationship between host adaptation and host restriction is not entirely clear, but it is likely that adaptation to a particular host may come at the expense of virulence traits required for infection of a wider range of hosts, ultimately resulting in restriction to a very narrow range of hosts.

Host adaptation in *S. enterica* is associated with host specificity of macrophages and other cell types which the bacteria is able to invade and survive within (635, 636, 637, 638). This involves host specificity of cells to which the bacteria is able to adhere, which is associated with fimbriae (639), thus host specificity in cell adhesion may result from the the particular combination of fimbrial genes present in a given serovar (120, 640). Host adaptation is also associated with differences in host responses to invasion with particular serotypes, for example infection of chicken cells with host generalist serovar Enteritidis or Typhimurium results in expression of the pro-inflammatory host cytokine IL-6, whereas infection with the fowl adapted serovar Gallinarum does not (641). Host adapted serovars are also often associated with much higher rates of chronic carriage than host generalist serovars. For example, Typhi and Paratyphi A are able to establish chronic infection of the gall bladder in humans (309, 311, 313), while cattle adapted serovar Dublin frequently results in chronic carriage following infection of cows (642) but not humans. Host adapted serovars also tend to display increased virulence, associated with systemic infection and bacteraemia (31, 32, 290). Systemic infection results in higher rates of morbidity and mortality, which could be disadvantageous for the pathogen. However for enteric pathogens, systemic infection is much more likely to lead to chronic carriage in the gall bladder, which may open a new route to increased transmissibility as carriers remain infected and therefore infectious for a long time. The mechanisms by which these traits evolve are unclear, but presumably involve selection for acquisition, deletion and mutation of specific genes.

The availability of multiple complete genome sequences for both Typhi and Paratyphi A provides the opportunity to study the genetic basis for their pathogenic convergence in detail. The Typhi and Paratyphi A genomes are much more closely related at the DNA level than other *S. enterica* serovars. Didelot *et al.* showed that this was due to a relatively recent recombination between a quarter of their genomes (56). They found a quarter of genomic sequences exhibited low nucleotide divergence (mean 0.18%) between Paratyphi A and Typhi, while the rest of the genome sequences were as divergent as any other pair of *S. enterica* serovars analysed (mean 1.2%) (56). Model-based simulations indicated that this was most likely due to relatively recent convergence via recombination between 23% of the Paratyphi A and Typhi genomes, which occurred in a rapid burst long after their initial divergence around the same time as other *S. enterica* serovars. The direction of recombination could not be determined, and may have been uni- or bi-directional. Furthermore the role of this recombination in each serovar's restriction and/or adaptation to the human systemic niche is unknown.

There are no known virulence genes unique to Typhi and Paratyphi A. Until recently SPI8 was thought to be unique to these serovars, but it is present in the genome sequence of serovar Agona (EMBL: NC_011149) and was recently detected in other serovars using a PCR screen (156). The Paratyphi A genome does not carry SPI7 and does not produce Vi, but it does carry a SopE-prophage similar to that present within Typhi SPI7 and other serovars. The Typhi and Paratyphi A genomes harbour a large number of pseudogenes (>4% of coding sequences in each genome) (46, 47, 49) compared to many host-generalist relatives such as *S. enterica* serovar Typhimurium (0.9%) or *E. coli* K-12 (1.2%). As discussed above, loss of gene function through pseudogene formation and gene deletion appears to be a hallmark of host restricted pathogenic bacteria compared to their host-generalist relatives (46, 49, 263, 264, 265, 266). The reason is uncertain, but is likely to be a combination of (a) adaptation, whereby the loss of certain proteins has a selective advantage in the new host, and (b) genetic drift, due to population bottlenecks following host restriction and/or the absence of selective pressure to maintain certain functions that are no longer required in the new host. It has been reported that Paratyphi A and Typhi share some of their pseudogenes (49), resulting in convergent loss of protein functions which may be associated with adaptation to their shared niche. The first Paratyphi C genome sequence (strain RKS4594/SARB49) was recently published and showed similarly high levels of pseudogenes (3.3%), a few of which were shared with Paratyphi A and Typhi (93). A single Paratyphi B genome sequence is available, however the annotation is incomplete and does not include many pseudogenes, which are likely to be missed by automated annotation (EMBL:CP000886 as of June 1, 2009). Furthermore it is unclear whether the sequenced strain, SPB7, is of the systemic pathotype (negative for tartrate fermentation, but also *sopE*-negative (225)).

4.1.1 Aims

The aim of this chapter was to investigate convergent evolution of Typhi and Paratyphi A by developing a comparative annotation of genetic features unique to these serovars. Specific aims were to:

- identify gene acquisitions and losses unique to Typhi and Paratyphi A in the context of *S. enterica*;
- produce a comparative annotation of pseudogenes across all Paratyphi A and Typhi genomes, including identifying genes that are pseudogenes in both genomes and comparing the inactivating mutation(s) in these pseudogenes;
- determine how many pseudogenes, acquired genes and gene deletions were shared between serovars via recombination or otherwise; and
- determine the relative timing of recombination and pseudogene accumulation in Paratyphi A and Typhi.

4.2 Methods

4.2.1 Whole genome comparisons

Mauve (algorithm `progressiveMauve`, default parameters) (578) was used to align the Typhi CT18 and Paratyphi A AKU_12601 genomes with those of the eleven other *S. enterica* serovars with whole genome sequences available in EMBL/GenBank as at June 1, 2009 (listed in Table 4.1). In addition to a whole genome alignment, Mauve reports regions that were not conserved in all the input genomes. This was used to identify sequences that were present in Typhi and/or Paratyphi A that were absent from all other genomes. As an independent method of identifying unique sequences, pairwise whole-genome BLASTN searches were performed for Typhi and Paratyphi A genomes against genomes of the other 11 serovars. The distributions of sequences identified by Mauve and/or BLASTN as potentially unique among *S. enterica* to Typhi and/or Paratyphi A were manually checked using BLAST nucleotide and protein searches of the GenBank database. Genome comparisons were visualised in ACT (604) and circular representations of the genomes were drawn using DNAplotter (643). A whole-genome maximum likelihood phylogenetic tree was constructed using RAxML (GTR+Γ model) with 100 bootstraps (644).

4.2.2 Phylogenetic network analysis

Phylogenetic networks were generated for multi-locus sequence data using SplitsTree4 (603). For analysis of the *S. enterica* MLST database (464), representative sequences were generated for each unique sequence type (ST) by concatenating sequences for each of the seven locus variants defining that ST. The assignment of locus variants to STs, and the sequences themselves, were downloaded from the database on December 10, 2008. The multiple alignment of concatenated sequences was used to construct a distance matrix and define a neighbour-joining network in SplitsTree4. The alignment of sequences from recombined and non-recombined genes, described in 4.2.3 was analysed in the same way.

4.2.3 Bayesian analysis of recombined and non-recombined genes

Nucleotide sequences for the seven genes used in the *S. enterica* MLST scheme (*aroC, dnaN, hemD, hisD, purE, sucA, thrA*; complete gene sequences) were used to build a

Serovar	**Pseudogenes**	**Host range**	**Accession**
Typhi	4.57%	Human (systemic)	AL513382
Paratyphi A	4.22%	Human (systemic)	FM200053
Paratyphi C	3.17%	Human (systemic/GI)	NC_012125
Gallinarum	7.23%	Avian (systemic)	NC_011274
Agona	3.26%	Mammalian (GI)	NC_011149
Choleraesuis	3.29%	Mammalian (GI) Porcine (systemic)	NC_006905
Dublin	5.83%	Mammalian (GI) Bovine (systemic)	NC_011205
Enteritidis	2.62%	Mammalian (GI) Avian (reproductive tract)	NC_011294
Heidelberg	3.48%	Mammalian (GI)	NC_011083
Newport	2.97%	Mammalian (GI)	NC_011080
Paratyphi B dT-	>0.45%	Possibly human (systemic/GI) Possibly mammalian (GI)	NC_010102
Schwarzengrund	3.17%	Mammalian (GI)	NC_011094
Typhimurium	0.56%	Mammalian (GI) Murine (systemic)	AE006468

Table 4.1: *S. enterica* serovar genomes - Details of all *S. enterica* serovars with finished genome sequences available in public databases, as at 1 June 2009. Accessions are for EMBL/Genbank.

phylogenetic tree of the *S. enterica* serovars in Table 4.1, using *S. arizonae*, *S. bongori* and *E. coli* sequences as outgroups. A separate analysis was done for a random sample of seven genes lying in regions that have undergone recombination between Typhi and Paratyphi A (56) and are conserved in *Salmonella* and *E. coli* (*moaC, rluC, oppB, accB, ilvE, atpF, uspA*; complete gene sequences). Homologous sequences in each genome were identified by BLASTN search with the Typhi CT18 nucleotide sequence as the query. Multiple alignments for each gene were constructed using ClustalX (574), and concatenated into two codon alignments, one for the recombined genes and one for the non-recombined genes.

Analysis with ModelTest (645) (implemented in FindModel (646)) suggested the GTR+Γ substitution model provided the best fit to both data sets. The Bayesian estimation package BEAST (647) was used to fit a GTR+Γ two-site codon substitution model to the data using MCMC analysis. Ten million iterations were run for each combination of tree priors (coalescent with constant population size; coalescent with exponential population growth; or speciation) and molecular clocks (strict or relaxed (648)). Models were compared via the Bayes factor, the ratio of the marginal likelihoods of each model (estimated by calculating the harmonic mean of the marginal likelihoods for the output of each model (649) in BEAST). The coalescent prior with a relaxed uncorrelated log-normal clock (648) gave the best fit to both data sets (Bayes factor 13-15 compared to strict clocks, 25-46 compared to speciation; note that Bayes factor >10 is considered strong support). The coefficients of variation for the relaxed clock models were significantly greater than zero (95% confidence intervals [0.20,0.72] and [0.13,0.47]), providing further support for a relaxed clock model over a fixed clock (648). The covariance of parent and child branches under the log-normal model of rate variation was essentially zero (95% confidence intervals [-0.27,0.33] and [-0.34,0.32]), confirming that the uncorrelated relaxed clock is appropriate for these data sets (648). The estimated growth rate under the exponential growth model was negative and close to zero, suggesting that a constant population size provides a better fit for this data.

Thus the final analysis was performed using the coalescent prior with constant population size, and relaxed clock with uncorrelated log-normal rate distribution. An additional 10 million iterations were run using this combination of settings for both recombined and non-recombined gene sets, and results from 20 million iterations combined for each data set. Dates were calibrated using two reported ages of the split between *E. coli* and *Salmonella*: 140 million years and 70 million years. The 140 million year estimate was based on comparison of DNA encoding 16S RNA between *E. coli* and *S. enterica* serovar Typhimurium (42). Later studies, based on protein alignments for glutamine synthetase (44) and other proteins (45) across the Bacterial and even other kingdoms, have yielded much lower estimates in the range 70-114 years.

4.2.4 Time estimation using dS

At the time of the study, model-based estimates of divergence time were not possible using whole-genome data, as the appropriate software packages were unable to handle such large data sets (e.g. BEAST (647), which was used previously to generate estimates of the age of Typhi (2), and above (4.2.3) to estimate divergence times based on small sets of genes). However, a simple approximation can be used to estimate divergence time from SNP data. Divergence time between a set of genomes can be approximated by the mean divergence since their most recent common ancestor (mrca) divided by the annual mutation rate per site (molecular clock rate). Using the number of synonymous SNPs per available site as a measure of divergence, the time t_{mrca} since the most recent common ancestor can be expressed as:

$$t_{mrca} = \frac{1}{n}\sum_{i=1}^{n}\frac{s_i}{S * \mu}, \tag{4.1}$$

where n is the number of genomes, s_i is the number of synonymous SNPs accumulated in genome i since the mrca, S is the number of synonymous SNP sites in the genome and μ is the synonymous substitution rate per site per year.

The numbers of synonymous SNP sites (S) included in the analyses of Paratyphi A (Chapter 3) and Typhi (Chapter 2) were 1.27 million and 1.35 million respectively. The mean number of synonymous SNPs accumulated in each genome since the most recent common ancestor ($\frac{1}{n}\sum_{i=1}^{n} s_i$) were 31 for Paratyphi A (s_i range 21-37) and 89 for Typhi (s_i range 71-102). The mutation rate for synonymous sites was estimated previously at 3.4x10^{-9}, based on a divergence date for *Salmonella* and *E. coli* of 140 million years ago (2, 41, 42, 43). Using the lower estimate of 70 million years, the upper bound on this rate would be 6.8x10^{-9}. The alignments of recombined and non-recombined genes described above were used to generate a novel rate estimate and to provide an alternative measure of divergence times. Pairwise synonymous site divergence ($dS_{i,j}$ for genomes i and j) was calculated using the method of Yang and Nielsen (650) implemented in the `yn00` algorithm of the software package PAML (651). Since pairwise dS incorporates evolution on both branches since the mrca ($dS_{i,j} = dS_i + dS_j$), estimates were divided by 2 before use in equation 4.1. The mean $dS_{i,j}$ between *E. coli* and *Salmonella* sequences was 0.8, providing a novel molecular clock rate estimate of 0.8/2/140,000,000 = 2.8x10^{-9} per site per year, or 5.6x10^{-9} using the alternative calibration time of 70 million years.

Assuming the silent substitution rate is equivalent for both serovars, the ratio of t_{mrca} between Paratyphi A and Typhi was approximated (using the mean and range of s_i) as:

$$\frac{t_{mrca}(ParatyphiA)}{t_{mrca}(Typhi)} = \frac{\frac{1}{n}\sum_{i=1}^{n} s_i}{\frac{1}{m}\sum_{j=1}^{m} s_j} / \frac{1.27}{1.35} \tag{4.2}$$

4.2.5 Comparison and annotation of pseudogenes

In order to compare annotated genomes of Paratyphi A AKU_12601 and ATCC9150, Typhi CT18 and Ty2 with Typhimurium LT2, pairwise whole-genome sequence comparisons were generated with BLASTN and visualised using ACT (604). Every gene annotated as a pseudogene in any Typhi or Paratyphi A genome was manually inspected in all five genomes and its pseudogene status in each genome reassessed. All pseudogenes identified in this way were included in the AKU_12601 genome annotation, although many such genes are not annotated in all of ATCC9150, CT18 and Ty2. For coding sequences found to be a pseudogene in more than one serovar, multiple alignments were constructed and viewed using ClustalX (574) to determine whether the same or independent inactivating mutation(s) were present in the different serovars. Pseudogenes annotated in the novel Paratyphi C genome were compared to those in the table and inactivating mutations were compared for all shared pseudogenes using ClustalX (574) and ACT (604).

4.2.6 Data simulation

An initial set of 40 genes were selected at random to represent ancestral pseudogenes. Additional sets of 20 and 150 genes were selected at random for each of two serovars, to represent pseudogenes that accumulated after initial divergence of the serovars (sampling with replacement). The same random sets of pseudogenes were used to simulate both scenarios, with only the timing varying (set of 150 pseudogenes arising before or after recombination). To simulate uni-directional recombination events depicted in Figure 4.6, serovar 2 pseudogenes lying in recombined regions were replaced with serovar 1 pseudogenes lying in recombined regions. Note the effect would be the same using replacement with randomly distributed directionality. All genes were selected at random from the 4,600 annotated in Typhi CT18 and their status as recombined or non-recombined was taken directly from the table of Typhi genes provided in (56).

4.3 Results

4.3.1 Evolution of Typhi and Paratyphi A

It is estimated that *Salmonella* diverged from *E. coli* 70-140 million years ago (42, 44, 45) and *S. enterica* diverged from the rest of *Salmonella* some time later. The diversification of *S. enterica* subspecies *enterica* into thousands of serovars (19) is generally thought of as a radiation or "star-burst" punctuated by recombination between lineages, rather than a series of bifurcations resulting in clear phylogenetic relationships (23, 56, 446, 448, 602, 652, 653). In order to test these assumptions with the most recent sequence data and attempt to date significant points in the evolution of Typhi and Paratyphi A, all publicly available whole genome and MLST data for *S. enterica* serovars was analysed. Thirteen whole genome sequences were available in EMBL/GenBank (as at June 1, 2009), listed in Table 4.1. These were aligned with Mauve and used to build a phylogenetic tree using maximum likelihood to fit a GTR+Γ model (see 4.2.1). The resulting unrooted tree (Figure 4.1a) showed some structure among the serovars, including very close relationships between Choleraesuis and Paratyphi C, and between Enteritidis and Gallinarum. Typhi and Paratyphi A were more closely related to each other than to other serovars, due at least in part to the reported recombination (56). In order to capture information about the broader

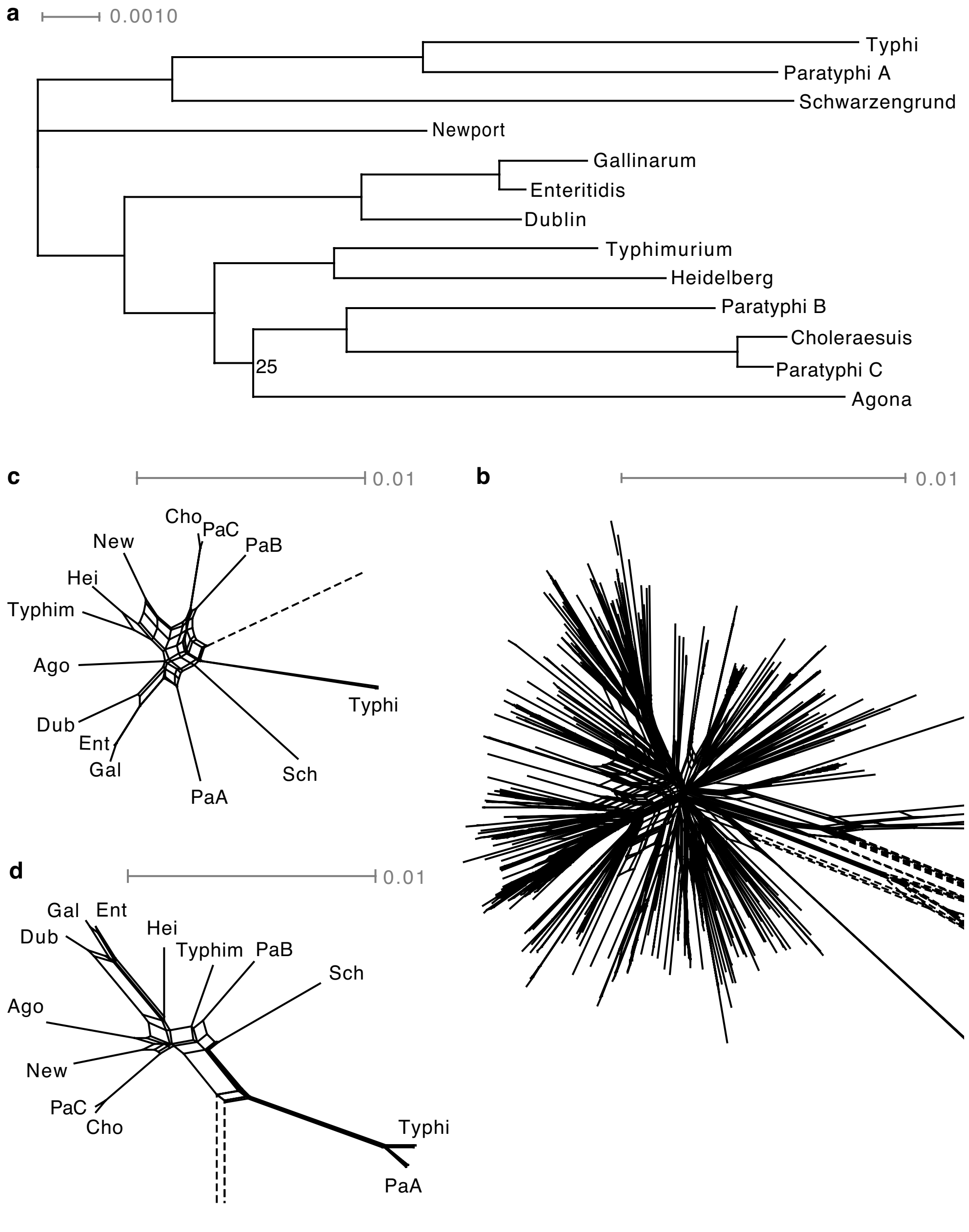


Figure 4.1: Phylogenetic trees for *Salmonella enterica* - Scale bars show substitutions per nucleotide. Dashed lines indicate where other *Salmonella* species join the networks. (a) Maximum likelihood phylogenetic tree of *S. enterica* based on whole genome alignment. The tree shown is the best fit (maximum likelihood) from 100 bootstraps; all nodes had 100% bootstrap support except the divergence of serovar Agona which had 25% support as shown. (b) Neighbour-joining phylogenetic network based on concatenated MLST sequences for all *S. enterica* available in the *S. enterica* MLST database. (c-d) Neighbour-joining phylogenetic networks based on seven non-recombined genes and seven recombined genes, respectively (recombined between Typhi and Paratyphi A). Ago = Agona, Cho = Choleraesuis, Dub = Dublin, Ent = Enteritidis, Gal = Gallinarum, Hei = Heidelberg, New = Newport, PaA = Paratyphi A, PaB = Paratyphi B, PaC = Paratyphi C, Sch = Schwarzengrund, Typhim = Typhimurium.

range of *S. enterica* serovars, all MLST sequences currently available in the *S. enterica* MLST database were used to build a phylogenetic network (see 4.2.2). The network (Figure 4.1b) is consistent with a radial pattern of diversification, with most serovars more or less equally closely related.

In order to examine the relative timing of the recombination event more closely, multiple alignments of *S. enterica* sequences were constructed for seven recombined and seven non-recombined genes, using *S. arizonae*, *S. bongori* and *E. coli* as outgroups (see 4.2.3). Phylogenetic networks of these sequences (4.2.2) are shown in Figure 4.1c-d, which again support a radial model of diversification in *S. enterica* (Figure 4.1c) but highlight a different pattern among the subset of genes assumed to be recombined between Typhi and Paratyphi A (Figure 4.1d). The alignments were analysed using Bayesian estimation (implemented in BEAST (647)) to fit an appropriate model (GTR+Γ codon model, relaxed molecular clock, coalescent prior with constant population size, see 4.2.3) and estimate the divergence date of *S. enterica* and also the divergence date for Paratyphi A and Typhi recombined sequences (see 4.2.3). The date of divergence between *E. coli* and *Salmonella* (70-140 million years (42, 44, 45)) was used to calibrate the time scale. Figures 4.2 and 4.3 show the resulting phylogenetic trees and estimates for the ages of key nodes. In the analysis of recombined data, the divergence time for Typhi and Paratyphi A was 0.75-1.5 Mya (million years ago), compared to ~2.5-5 Mya divergence time for Typhi and Paratyphi A using the non-recombined sequences and 3.5-10 Mya for the divergence of *S. enterica.* For both recombined and non-recombined data sets, the mean mutation rate (molecular clock) was ~1.3x10^{-9} substitutions per nucleotide per year.

The 'age' of Typhi, that is the time since divergence of extant strains from their most recent common ancestor, has previously been estimated using first 3.3 kbp of sequence in 26 strains (MLST (1)) and then 89 kbp of sequence in 105 strains (SNP detection by dHPLC (2)). The resulting estimates were 15,000-150,000 and 10,000-43,000 years respectively. These estimates themselves rely on an estimate of the molecular clock rate in *Salmonella* (see 4.2.4), which come from comparisons of *S. enterica* and *E. coli* (2, 41, 42, 43). No estimate has been reported for Paratyphi A since sequence information has been scarce, although it has been reported that the Paratyphi A population

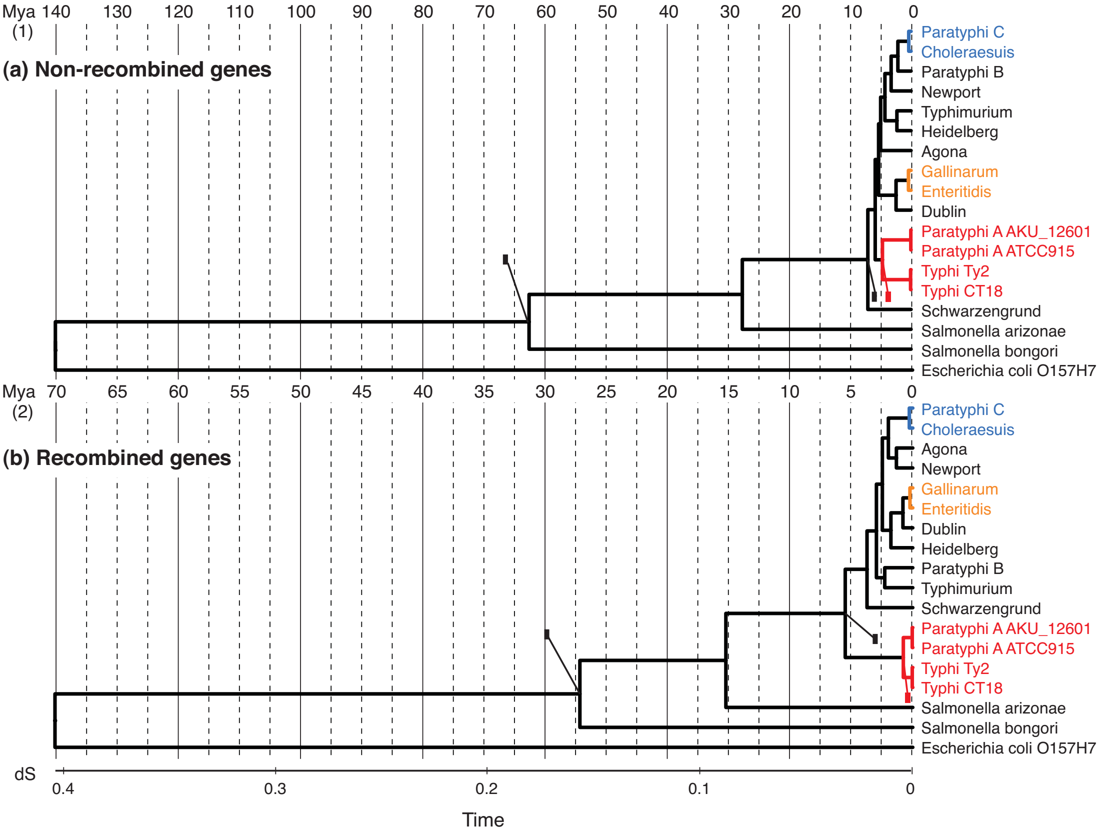


Figure 4.2: Phylogenetic trees for *Salmonella* and *E. coli* with divergence time estimates - Bayesian analysis was used to construct phylogenetic trees using (a) seven genes that were not recombined between Typhi and Paratyphi A and (b) seven genes that were recombined between them. Bayesian phylogenetic analysis was conducted in BEAST using a GTR+Γ 2-site codon substitution model, a relaxed molecular clock with log-normally distributed rates, a coalescent prior with constant population size and estimates of 70-140 million years for the date of divergence between *Salmonella* and *E. coli*. The trees shown are the marginal trees from 20 million iterations on each data set. Time along the x-axis is labelled in 3 ways: Mya (1) and Mya (2) = time in millions of years before present with the root calibrated to 140 and 70 million years, vertical lines correspond to these divisions, branch lengths and node positions were fit to this scale; dS = synonymous substitution rate, estimates for specific nodes are also given on this scale, indicated by rectangles (at the correct time point on this scale) joined to tree nodes by thin lines.

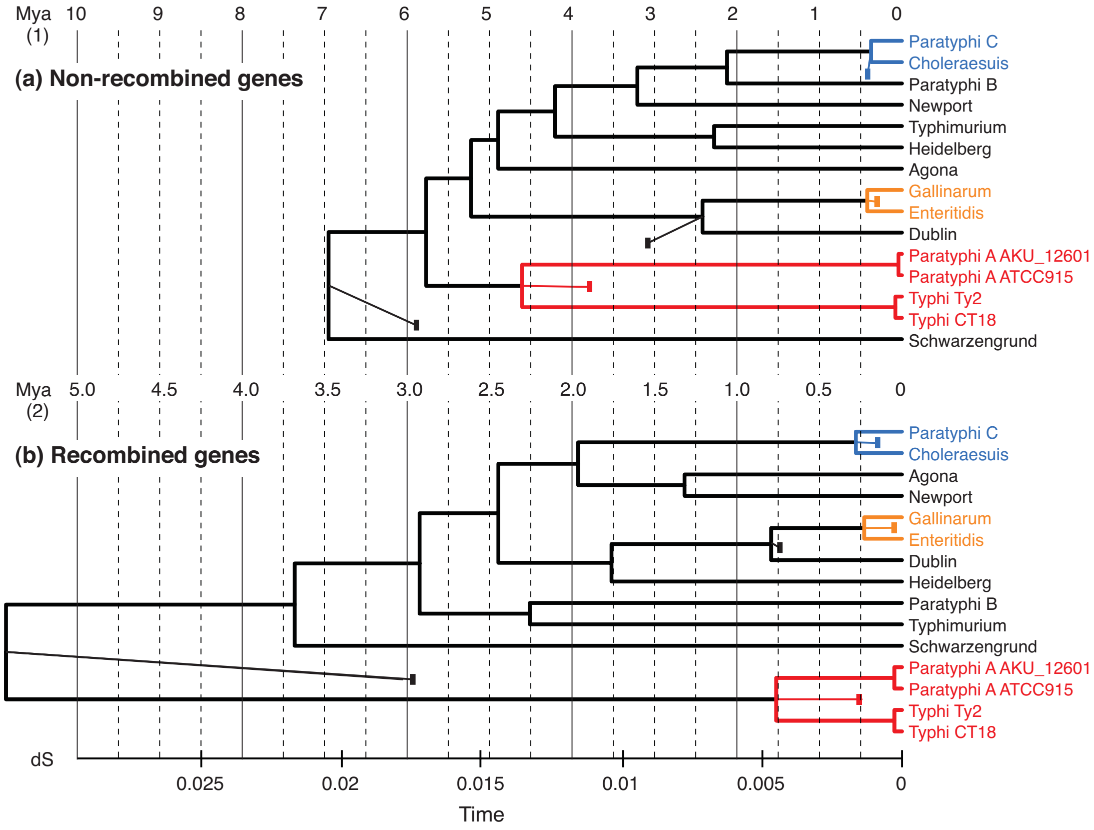


Figure 4.3: Phylogenetic trees for *S. enterica* with divergence time estimates - Zoom in on *Salmonella enterica* from the trees shown in Figure 4.2. Bayesian analysis was used to construct phylogenetic trees using (a) seven genes that were not recombined between Typhi and Paratyphi A and (b) seven genes that were recombined between them. Bayesian phylogenetic analysis was conducted in BEAST using a GTR+Γ 2-site codon substitution model, a relaxed molecular clock with log-normally distributed rates, a coalescent prior with constant population size and estimates of 70-140 million years for the date of divergence between *Salmonella* and *E. coli.* The trees shown are the marginal trees from 20 million iterations on each data set. Time along the x-axis is labelled in 3 ways, exactly as in Figure 4.2: Mya (1) and Mya (2) = time in millions of years before present using two alternative calibration times, vertical lines correspond to these divisions, branch lengths and node positions were fit to this scale; dS = synonymous substitution rate, estimates for specific nodes are also given on this scale, indicated by rectangles (at the correct time point on this scale) joined to tree nodes by thin lines.

is less diverse than Typhi (single MLST type (464); fewer PFGE profiles (479)). Since the Bayesian estimation sofware (BEAST (647)) could not handle whole genome data, it was not possible to use this method to estimate the divergence times for Typhi or Paratyphi A using the SNPs identified in Chapters 2 and 3. However a simple estimation was used based on the rate of synonymous substitutions (dS) detected in each population, as outlined in 4.2.4. Calculations were made using the previously described molecular clock rate of 3.4x10^{-9} synonymous substitutions per site per year (42, 43) and a new clock rate of 2.8x10^{-9} based on dS between *E. coli* and *Salmonella* sequences analysed in this study (4.2.3 and 4.2.4). Both rates are based on an *E. coli*-*Salmonella* divergence time of 140 Mya, so should be doubled to incorporate the lower estimate of 70 Mya for this divergence (see 4.2.3). The resulting estimate for the age of Typhi was 19,000-24,000 years using slow rates based on 140 Mya calibration, and 10,000-12,000 using fast rates based on 70 Mya calibration. For Paratyphi A the estimates were 7,000-9,000 using slow rates, and 3,600-4,400 using fast rates, suggesting Typhi is significantly older than Paratyphi A.

To allow direct comparison to the other ages estimated above, the pairwise synonymous substitution rate (dS) was calculated between each pair of sequences included in the analysis of recombined and non-recombined genes. The resulting estimates, shown as rectangles in Figures 4.2 and 4.3 were generally smaller than those estimated with Bayesian analysis. The dS method resulted in similar estimates for the divergence of *S. enterica* serovars using the recombined and non-recombined genes (dS 0.0173-0.0175, see Figure 4.3a-b). However Bayesian analysis gave less consistent estimates for *S. enterica* divergence using the two data sets: 11 Mya (95% confidence interval [5-19]) with recombined genes vs 7 Mya [4-10] for non-recombined genes (using the 140 Mya root calibration). All methods gave compatible estimates for the divergence of *S. bongori* from other *Salmonella* lineages, see Figure 4.2. While mutation rates are uncertain, the direct comparison of dS for the Typhi and Paratyphi A populations may give a reliable indication of their relative ages, assuming both populations have been subject to similar short-term substitution rates that have not varied too much since the last common ancestor of the oldest serovar (see 4.2.4). The ratio of dS among Paratyphi A to dS among Typhi was 0.36 (range 0.29-0.47), suggesting that Paratyphi A is approximately one third the age of Typhi.

4.3.2 Convergent features of the Typhi and Paratyphi A genomes

4.3.2.1 Shared genes

The Typhi CT18 and Paratyphi A AKU_12601 genomes were compared to each other and to all other *S. enterica* genomes available at the time of the study (11 serovars in EMBL/GenBank, June 1, 2009). The genomes used in the comparison are listed in Table 4.1, along with their known host ranges. They include several host-generalist serovars that cause gastroenteritis in humans and the host-adapted serovars Paratyphi C (human), Gallinarum (chicken), Dublin (cattle) and Choleraesuis (swine). Pairwise nucleotide BLAST searches and a multiple whole genome alignment were used to identify genes that were present in Typhi and/or Paratyphi A but absent from the other *S. enterica* genomes (see 4.2.1). The genes are listed in Table 4.2 and their distribution in the Typhi and Paratyphi A genomes is shown in Figure 4.4.

The majority of genes unique to either the Typhi or Paratyphi A genomes were prophage genes (Table 4.2a,b), many of which were found to be absent from other Typhi or Paratyphi A strains (2.3.3.1, 3.3.1.3). The CT18 genome contained two fimbrial operons (*sta, stg*) not found in other *S. enterica* genomes. These were present in all of the Typhi genomes resequenced in Chapter 2 but *stgC* was always a pseudogene (although functional analysis suggests that this operon still encodes a functional fimbria in Typhi (640)). The SPI15 region described in 2.3.3.2 was only found in Typhi. SPI7 was present only in Typhi and Paratyphi C, although several pieces were missing from the Paratyphi C genome (92). Two predicted coding sequences STY4074 and STY4075 were found only in Typhi, between *waaB* and *waaP*. No variation was detected at this locus among Typhi genome sequences. The sequence between *waaB* and *waaP* has been studied in detail in a collection of *S. enterica* serovars during which the Typhi sequence was also found in serovar Stanleyville (66), and the authors of that study expressed doubt as to whether STY4074 and STY4075 really encode proteins.

Only five regions of the AKU_12601 genome were unique to Paratyphi A, including two prophage regions (see Table 4.2b). The locus SSPA3985-3987, encoding a restriction/modification system, was present in all seven Paratyphi A genomes sequenced to date but was not identified in any other serovars. The CDS SSPA2364 was present in

(a) Gene IDs	**Region**	**Function**
STY0201-07	*staABCDEFG*	fimbrial operon
STY1014-33;50-77	ST10	phage
STY1591-1643	ST15	phage
STY2012-77	ST18	phage
STY2879-89	ST27	phage
STY3188-93	SPI15	unknown
STY3658-3703	ST35	phage
STY3918-22	*stgABCD*	fimbrial operon
STY4074-5	-	polysaccharide pyruvyl transferase family domain
STY4547-52	*pilSTUV*, *rci*	type IVB pilus (in SPI7)
STY4667-80	SPI7	unknown
STY4822-24,27	ST46	phage (in SPI10)

(b) Gene IDs	**Region**	**Function**
SSPA2233-69	SPA-1	phage
SSPA2306,8-9	SPI6	unknown
SSPA2364	-	unknown
SSPA2424-34,45-47	SPA-3-P2	phage
SSPA3985-7	-	restriction/modification system

(c) Typhi	**Paratyphi A**	**Details**
STY2747-49	SSPA0337-35a	unknown
STY4629-32	SSPA2407-09	*unknown
STY3091	SSPA2625	*insertion in *ste* fimbrial operon
STY4217-22	SSPA3215-11	*unknown
STY4037,39	SSPA3365a,65	*unknown
STY4881	SSPA4034	*restriction/modification system gene

Table 4.2: Genes unique to Typhi and/or Paratyphi A - The Typhi CT18 and Paratyphi A AKU_12601 genomes were compared to each other and to those of serovars Paratyphi B, Paratyphi C, Choleraesuis, Typhimurium, Enteritidis, Gallinarum, Schwarzegrund, Agona, Dublin, Heidelberg and Newport. (a) Genes present only in Typhi. (b) Genes present only in Paratyphi A. (c) Genes present in Typhi and Paratyphi A but absent from the other genomes. *=divergence <0.3%, consistent with sharing via recombination.

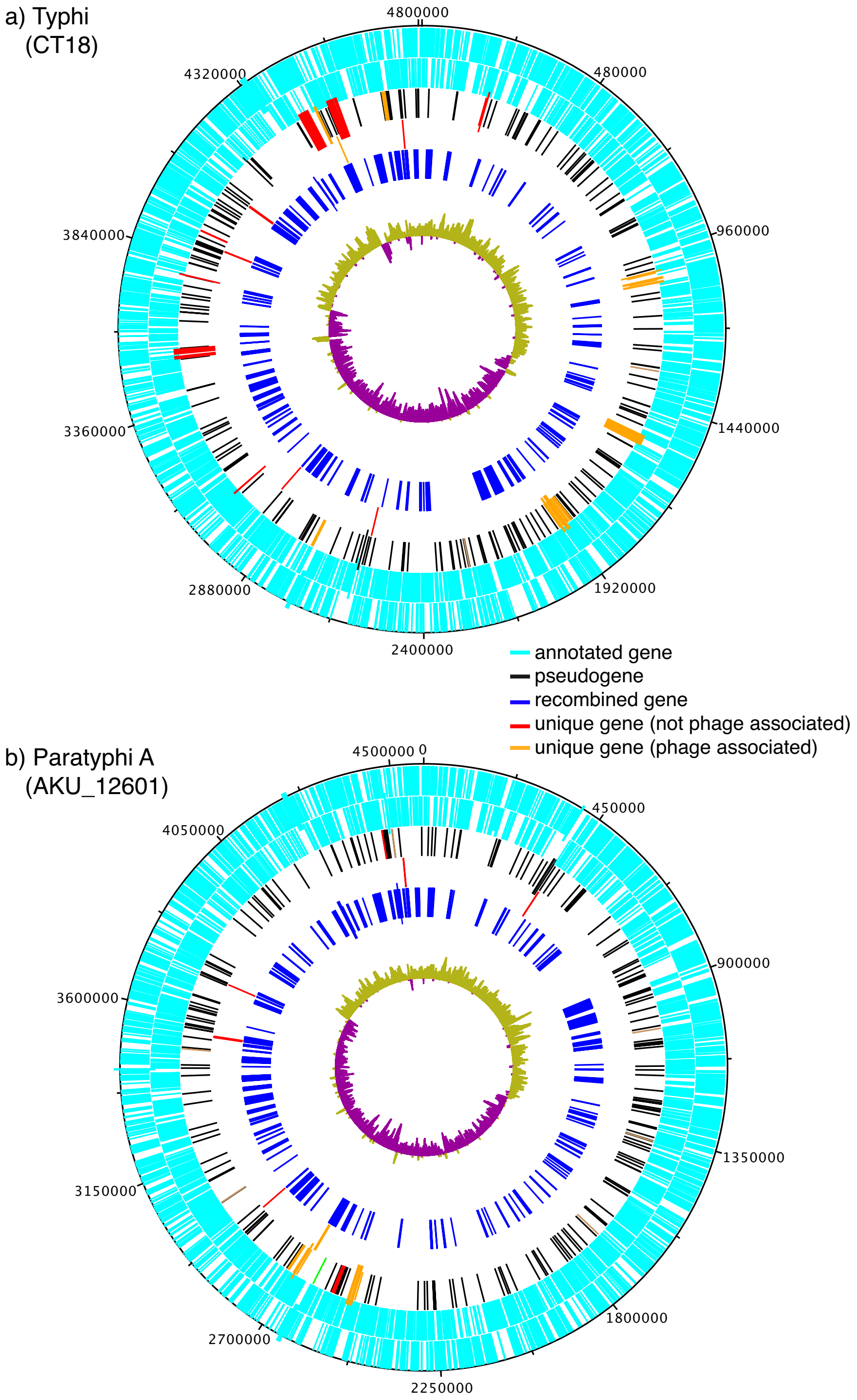


Figure 4.4: Pseudogenes, recombined genes, and unique genes in the Typhi and Paratyphi A genomes - Rings from outside: 1, 2 = coding sequences on forward, reverse strands; 3 = pseudogenes and genes unique to the serovar; 4 = genes present in both Typhi and Paratyphi A but absent from 11 other sequenced serovars; 5 = genes recombined between Typhi and Paratyphi A. Central plot shows GC deviation ((G-C)/(G+C), i.e. the difference in G content between the forward and reverse strands). Outer labels show genome sequence coordinates (bp).

Paratyphi A at a locus occupied in other serovars by a different sequence of similar size, including STY2867 in Typhi. A BLAST search of the SSPA2364 translated protein sequence revealed similar proteins in serovars Saint Paul and Javiana (which do not cause systemic infection in humans), however the sequence contained no protein domains of known function. In Paratyphi A, SPI6 contained a region of unique sequence compared to other serovars, including three CDSs of unknown function SSPA2306, SSPA2308 and SSPA2309. A BLAST search of the translated protein sequences revealed similar proteins in serovar Saint Paul, but no protein domains could be identified. The sequences did not vary between Paratyphi A isolates. Typhi and Paratyphi A shared just 17 CDSs that were not detected in the other *Salmonella* genomes (Table 4.2c), the functions of which are unknown.

4.3.2.2 Comparison of pseudogenes in Typhi and Paratyphi A

Typhi and Paratyphi A each carry >200 pseudogenes, distributed around their genomes (see Figure 4.4). This constitutes 4-5% of coding sequences, a higher rate than most host-generalist serovars of *S. enterica* (see Table 4.1). In order to comprehensively investigate the mechanisms of convergent gene loss in Paratyphi A and Typhi, a comparative table of pseudogenes present in all four finished genomes was assembled. The table was based initially on a list of pseudogenes annotated in ATCC9150, CT18 and Ty2. To this were added some additional Typhi pseudogenes previously noted during the annotation of Paratyphi A ATCC9150 (49), pseudogenes annotated in AKU_12601 and some novel pseudogenes identified by manually inspecting Typhi and Paratyphi A sequences for all genes annotated as pseudogenes in any of the AKU_12601, ATCC9150, CT18 or Ty2 genomes.

The resulting table included 66 pseudogenes common to Typhi genomes CT18 and Ty2 and Paratyphi A genomes AKU_12601 and ATCC9150 (Table 4.3 and see Figure 4.5). This was almost double the figure reported previously (49), although many of the additional pseudogenes were remnants of transposase or bacteriophage genes. By aligning the Typhi and Paratyphi A DNA sequences for the shared pseudogenes, inactivating mutations were identified and classified as shared or independent mutations (Table 4.3). Contrary to previous reports (49), many of the shared pseudogenes harboured identical inactivating mutations in each serovar. Twenty of the shared pseudogenes (54%

of non-phage/transposase shared pseudogenes) encode secreted or surface-exposed proteins (Table 4.3), the loss of which may have contributed to convergence upon similar patterns of host interactions.

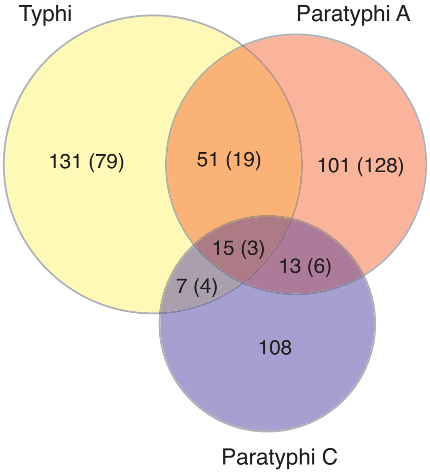


Figure 4.5: Overlap of pseudogenes in Typhi, Paratyphi A and Paratyphi C - Figures indicate the number of pseudogenes that are present in each serovar or combination of serovars; strain-specific pseudogenes are shown in brackets.

4.3.2.3 Genes missing from Typhi and Paratyphi A

A total of 38 genes were identified as absent from Typhi and Paratyphi A but present in the genomes of the eleven other serovars listed in Table 4.1. The genes, listed in Table 4.4, were mostly associated with energy use and metabolism, including anaerobic metabolism. They include a phosphotransferase system (STM4534-40) and a gene (*ydiD*/STM1350) involved in an anaerobic oxidation pathway associated with growth on fatty acids (654). Also missing were a putative efflux pump (STM0350-53) and a cluster of genes encoding an anaerobic C4-dicarboxylate transporter, L-asparaginase and a ribokinase (STM3598-600). In addition, the SPI2-secreted effector *sseJ* and chemotaxis receptor *trg* were deleted along with several neighbouring genes. In Typhimurium, SseJ is targeted to the *Salmonella*-containing vacuole, and deletion mutants show attenuated replication in mice (655). Trg is one of five chemotaxis receptors present in Typhimurium, which enable the bacterial cell to direct movement in response to attractant or repellent chemical stimuli; Trg is the glucose-specific receptor. The loss of *sseJ* and *trg* was noted in the publication reporting the Paratyphi A ATCC9150 genome

Class	SSPA	STY	Gene	Gene product	Div.
i$^+$	0062a	n/a	-	putative viral protein	-
i$^+$	0255a	n/a	-	putative uncharacterized protein	-
i	1103	1362	-	Pertussis toxin subunit S1 related protein	1.22%
i$^+$	1699a	0971	*sopD2*	*secreted effector protein SopD homolog	1.73%
i$^+$	2014	0610	*silA*	*putative inner membrane proton/cation antiporter	1.08%
i$^+$	2014a	0609a	*cusS*	*putative copper-ion sensor protein	0.18%
i$^+$	3229	4202	-	putative phosphosugar-binding protein	0.14%
i	3640	3800	*cdh*	CDP-diacylglycerol pyrophosphatase	2.32%
i$^+$	3888	4728a	-	putative uncharacterized protein	1.35%
i				*30 transposase/phage genes and gene remnants*	
ii	0097	0113	-	*putative secreted protein	0.25%
ii	0431b	2631	-	putative IS transposase	0.24%
ii	**0754a**	**2275**	***sopA***	***secreted effector protein**	**0.23%**
ii	3228	4203	-	putative L-asparaginase	0.14%
ii	3365a	4037	*sugR*	putative uncharacterized protein (SPI3)	0.14%
iii	0192a	0218	*fhuA*	*ferrichrome-iron receptor precursor	23.95%
iii	0317a	2775	-	putative anaerobic dimethylsulfoxide reductase component	1.79%
iii	0329a	2762	*sivH*	*putative invasin (CS54)	1.17%
iii	**0331a**	**2758**	***ratB***	***putative lipoprotein (CS54)**	**1.67%**
iii	**0331b**	**2755**	***shdA***	***putative uncharacterized protein (CS54)**	**2.11%**
iii	**0621a**	**2422**	***mglA***	***galactoside transport ATP-binding protein**	**1.09%**
iii	0720a	2311	*wcaK*	*putative extracellular polysaccharide biosynthesis protein	1.82%
iii	0756a	2268	*yeeC*	penicillin-binding protein	2.19%
iii	**0850a**	**2166**	***fliB***	***lysine-N-methylase**	**3.11%**
iii	**0943a**	**1995**	**-**	**transposase**	**4.77%**
iii	1014a	1913	*hyaA*	hydrogenase-1 small subunit	0.33%
iii	1220a	1508	-	*putative transport protein	1.31%
iii	1367a	1739	-	putative ribokinase (SPI2)	1.42%
iii	**1531a**	**1244**	***fhuE***	***FhuE receptor precursor**	**0.96%**
iii	1642a	1104	-	*putative secreted protein	1.54%
iii	**1820a**	**0833**	***slrP***	***secreted effector protein**	**1.95%**
iii	2045a	0569	*ybbW*	*putative allantoin transporter	1.19%
iii	2301a	0333	*safE*	*probable lipoprotein (SPI6 fimbrial cluster)	1.52%
iii	3388a	4007	-	putative cytoplasmic protein	1.12%
iii	**3636a**	**3805**	-	***permease, Na+:galactoside symporter family**	**2.42%**
iii	3828b	4503	*dmsA*	anaerobic dimethyl sulfoxide reductase chain A	0.22%
iii	3998a	4839	*sefD*	*putative fimbrial protein (SPI10)	0.18%

Table 4.3: Pseudogenes shared between Paratyphi A and Typhi - (i) Ancestral pseudogenes; '+' intact in Typhimurium. (ii) Pseudogenes shared by recombination. (iii) Recent conserved pseudogenes (independent inactivating mutations in each serovar). SSPA and STY - systematic identifiers in Paratyphi A AKU_12601 and Typhi CT18 respectively; n/a - not annotated. For genes lying in SPIs the island is indicated in brackets after the gene product. Div. - nucleotide divergence reported in (56). *Secreted or surface-exposed proteins; bold - Paratyphi C pseudogene.

sequence (49). The authors also pointed out that Tsr, the receptor specific for serine, was interrupted in Typhi while Tar, specific for aspartate and maltose, contained an inframe deletion in Paratyphi A. These mutations were conserved in all Typhi and Paratyphi A isolates analysed in Chapters 2 and 3.

ID	**Deletion**	**Functions**
STM0350-53	identical	Putative efflux pump
STM0538-39	identical	Putative membrane proteins
STM1188	identical*	Putative inner membrane lipoprotein
STM1350-62	identical	Proton-driven metabolite uptake system (*ydiLMN*,*aroED*) Anaerobic growth on fatty acids (*ydiFOPQRSTD*)
STM1625-31	different	Chemotaxis protein *trg*, secreted effector *sseJ*
STM2508-09	identical	Putative protein
STM3598-600	identical*	Anaerobic C4-dicarboxylate transporter, L-asparaginase, ribokinase
STM4534-40	identical	Phosphotransferase system

Table 4.4: Genes absent from Typhi and Paratyphi A but present in 11 other serovars - Identifiers in Typhimurium LT2 for deleted genes are given in column one. Column two indicates whether the deletion boundaries are identical in Typhi CT18 and Paratyphi A AKU_12601, *=deleted region flanked by recently recombined genes (divergence <0.3%).

4.3.2.4 Features shared with Paratyphi C

No genes were identified as present in Typhi, Paratyphi A and Paratyphi C but missing from all other serovars. Paratyphi C carries most of SPI7 and is capable of producing Vi (83, 92), however Paratyphi A does not share these features, indicating that they are not required for causing enteric fever in humans. A total of 15 pseudogenes shared by Typhi and Paratyphi A were also pseudogenes in Paratyphi C (highlighted in Table 4.3, see Figure 4.5). These include the secreted effectors *sopA* and *slrP*. Paratyphi C shared an additional 13 pseudogenes with Paratyphi A, including a putative chemotaxis receptor protein (SSPA1138a/SPC_2077), and an additional seven with Typhi, including putative chemotaxis receptor protein *yeaJ* (STY1834/SPC_2458) (Figure 4.5). No coding sequences were identified that were absent from Typhi, Paratyphi A and Paratyphi C but present in all other serovars.

4.3.3 The role of recombination

4.3.3.1 Sharing of unique genes and deletions by recombination

Regions containing genes that were shared uniquely by Typhi and Paratyphi A (Table 4.2c) were analysed for evidence of recombination. In each case, the insertion sites in both serovars appeared to be identical relative to Typhimurium LT2. Most of the shared genes had low divergence (<0.3%) between Paratyphi A and Typhi, and were flanked by other genes of low divergence (starred in Table 4.2c). These genes were therefore likely to have been shared via recombination. The only exception was a cluster of three genes, STY2747-49, which were 0.6% divergent between Typhi and Paratyphi A, and were not flanked by genes of low divergence. This insertion is therefore less likely to be the result of recombination between Typhi and Paratyphi A, at least not as recently as the other shared genes. A protein BLAST search with the translated amino acid sequences of these genes yielded hits in regions sequenced from serovars Weltevedren, Kentucky and Javiana (which do not cause systemic infection in humans), suggesting that this region is not actually unique to the enteric fever agents Typhi and Paratyphi A.

To investigate whether recombination played a role in shared deletions in Typhi and Paratyphi A, each deleted locus (Table 4.4) was examined in Typhimurium and the flanking sequences compared in Typhi and Paratyphi A. In seven of the eight regions, the deletion boundaries were identical in Typhi and Paratyphi A compared to Typhimurium. However, only two of these deletion sites were flanked by sequences of low divergence (<0.3%) (starred in Table 4.4), which would be expected if the deletion had been shared during recombination between homologous flanking sequences. For another two loci the deletion was flanked by identical repeats in Typhimurium, which may have facilitated the independent occurrence of the same deletion in Typhi and Paratyphi A via homologous recombination between the repeats. One of the regions with identical deletion boundaries involved the replacement of two genes (STM2508-09) with three genes (STY2747-49). As mentioned above, these three genes were 0.6% divergent between Typhi and Paratyphi A, and similar sequences have been reported in other serovars. Thus this region is unlikely to be shared via recombination. One region, including *trg* and *sseJ*, was almost certainly deleted independently in each genome, as the deleted region was larger in Paratyphi A than Typhi. It therefore appears that while recombination contributed to the sharing of unique genes between Typhi and Paratyphi A, most of the genes deleted from both genomes were the result of independent events.

4.3.3.2 Sharing of pseudogenes by recombination

More than 30% of the pseudogene complements of Typhi and Paratyphi A were shared (Figure 4.5), consistent with the possibility that recombination of 23% of the genomes resulted in direct sharing of many of their pseudogenes. It was determined whether each pseudogene was likely to have undergone relatively recent recombination between Paratyphi A and Typhi (sequence divergence <0.3% between serovars according to (56)). Of all the pseudogenes present in both Paratyphi A AKU_12601 and ATCC9150, 24.3% were recently recombined; of the pseudogenes present in both Typhi CT18 and Ty2, 25.0% were recently recombined. According to the original study by Didelot *et al.* (56), more than 20% of all genes in Typhi CT18 lie in the recently recombined regions.

These observations are consistent with two scenarios, illustrated in Figure 4.6: (1) most pseudogenes were inactivated prior to recombination, and recombination was random with respect to the location of pseudogenes (Figure 4.6b); or (2) most pseudogenes were inactivated after recombination, and these pseudogene-forming mutations were random with respect to recombined regions (Figure 4.6c). If (1) were true, we would expect that (i) genes that are pseudogenes in one serovar but intact in the other (i.e. serovar-specific pseudogenes) would not lie in recombined regions, and (ii) most pseudogenes in recombined regions would have been shared during recombination, i.e. they would be pseudogenes in both Paratyphi A and Typhi and share common inactivating mutations in both genomes (red circles in Figure 4.6b). If (2) were true, we would expect that (i) serovar-specific pseudogenes would be distributed randomly with respect to recombined and non-recombined regions, and (ii) very few pseudogenes would have been shared during recombination, i.e. very few pseudogenes in recombined regions would share inactivating mutations (red circles in Figure 4.6c).

The distribution of serovar-specific and shared pseudogenes in recombined and non-recombined regions is shown in Figure 4.6a and summarised in Table 4.5. Pearson χ^2 tests for each serovar based on this data gave non-significant results (p>0.2, Table

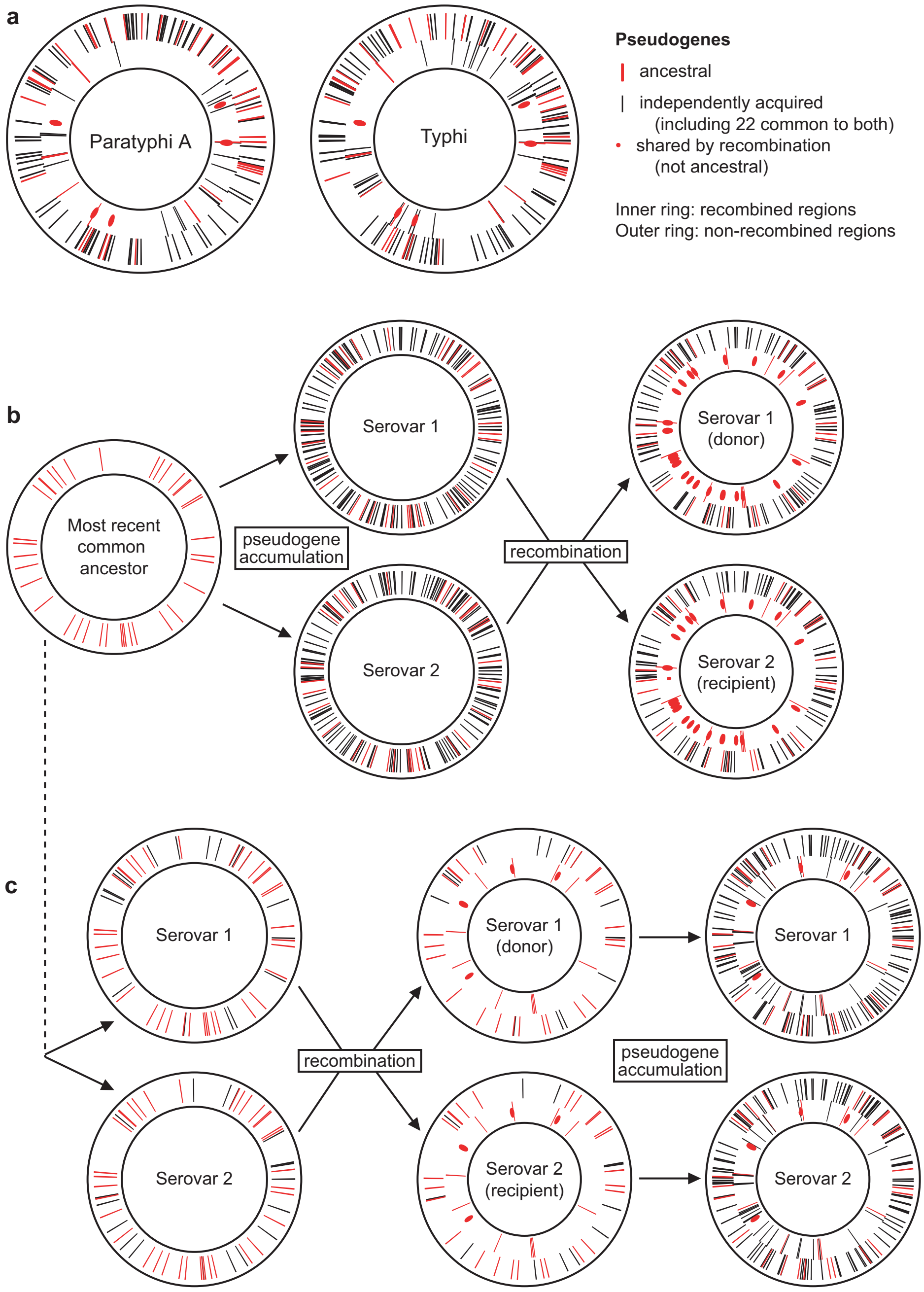


Figure 4.6: Scenarios of recombination and pseudogene formation in Paratyphi A and Typhi - (a) True distribution of pseudogenes in the Paratyphi A AKU_12601 and Typhi CT18 genomes (gene order based on gene coordinates in Typhi CT18). (b-c) Distribution of pseudogenes resulting from data simulated under two scenarios, under both of which 40 pseudogenes are inherited from the most recent common ancestor of Paratyphi A and Typhi, and extensive accumulation of pseudogenes occurs before or after recombination of 25% of genes. For ease of simulation, the recombination shown is uni-directional, but bi-directional exchange would result in similar patterns. (b) Scenario 1: 150 additional pseudogenes accumulate in each serovar, followed by recombination. (c) Scenario 2: only 20 additional pseudogenes arise before recombination, after which a further 150 pseudogenes accumulate in each serovar.

4.5), thus there was no evidence of association between shared or serovar-specific pseudogenes and regions of recombination, consistent with scenario (2). More than 20% of serovar-specific pseudogenes lie in recombined regions of each genome (Figure 4.6a, black lines in inner ring), consistent with scenario (2) whereby serovar-specific pseudogenes are expected to be randomly distributed in the genome of which 23% has been recombined (Figure 4.6c, black lines in inner ring). These observations are extremely unlikely under scenario (1), which would predict recombination to result in shared but not serovar-specific pseudogenes being present in recombined regions (Figure 4.6b, inner ring).

Distribution	Recombined	Non-recombined	χ^2 test, specific vs. shared
Typhi-specific	114	39	0.33 (p=0.57)
Paratyphi A-specific	92	24	1.63 (p=0.20)
Shared	46	20	

Table 4.5: Distribution of serovar-specific and shared pseudogenes in recombined regions - Pearson χ^2 tests were performed separately for each serovar based on the two-way contingency table obtained from the respective serovar-specific row and shared row.

Only 18 pseudogenes in recombined regions harboured the same inactivating mutations (red lines and circles in inner rings, Figure 4.6a), less than 20% of pseudogenes in the recombined regions of each genome. As illustrated in Figure 4.6, this is consistent with scenario (2) but not scenario (1), which would predict that most pseudogenes lying in recombined regions would be shared by virtue of recombination and therefore carry the same inactivating mutations (red circles in Figure 4.6). The observed patterns of pseudogene distribution therefore suggest that the majority of pseudogenes present in the extant genomes of Paratyphi A and Typhi accumulated after the recombination of 23% of their genomes.

4.3.4 Pseudogene formation in the evolutionary histories of Typhi and Paratyphi A

4.3.4.1 Pseudogene formation over time

The recombination described between Paratyphi A and Typhi provides a rare marker of relative time in the evolutionary histories of these organisms. The time estimates shown in Figures 4.2 and 4.3 should be treated with care, however it is clear that the recombination occurred well before the most recent common ancestors of each serovar, representing the last population bottlenecks in the Paratyphi A and Typhi populations.

The pseudogenes were divided into distinct categories with different relative ages (Table 4.3): (i) ancestral pseudogenes (shared pseudogenes inactivated prior to the divergence of Paratyphi A and Typhi), (ii) recombined pseudogenes (shared pseudogenes in recombined regions, with shared inactivating mutations assumed to have arisen after initial divergence), and (iii) recent conserved pseudogenes (including serovar-specific pseudogenes, and shared pseudogenes containing different inactivating mutations in Paratyphi A and Typhi; the majority of these are expected to have become pseudogenes after recombination). An additional category (iv) was defined (Table 4.6), containing 21 recent strain-specific pseudogenes that were shared by both serovars (i.e. containing inactivating mutations in some but not all strains belonging to their respective serovar). Tables 4.3 and Table 4.6 summarise the shared pseudogenes (excluding ancestral transposase/phage gene remants) and Figure 4.7 shows their approximate timing overlaid on a phylogenetic tree of *S. enterica* serovars. Note that some serovar-specific pseudogenes (group iii) will probably be strain-specific (group iv) as more strains are sequenced. It is clear from Figure 4.7 that the rate of accumulation of pseudogenes in both serovars increased dramatically at some point after the recombination event.

STY	SSPA	Ty	Pa	Gene	Product
1167	1599a	1	63	*nanM*	Conserved hypothetical protein
1486	1197a	9	all	*narW*	Respiratory nitrate reductase 2 delta chain
1574	1268a	1	all	*clcB*	Voltage-gated ClC-type chloride channel
1648	**1282a**	**3**	**all**		**Putative uncharacterized protein**
0026	**0021**	**all**	**15**	***bcfC***	**Fimbrial usher**
2229	0791	all	1	*cbiK*	Synthesis of vitamin B12 adenosyl cobalamide
2231	0790	all	1	*cbiJ*	Synthesis of vitamin B12 adenosyl cobalamide
2747	0337	all	1		Putative outer membrane lipoprotein
3421	2900a	all	19	*yhaO*	Putative transport system protein
3657	3484	all	2	*yifB*	Putative magnesium chelatase, subunit ChlI
3828	3616	all	3	*rhaD*	Rhamnulose-1-phosphate aldolase
4030	3370	all	2	*misL*	Putative autotransported protein
4162	3259a	18	18	*yhjW*	Putative membrane protein
4820	3979	all	2		Hypothetical fused protein
4876	**4030**	**all**	**3**		**Putative aldehyde dehydrogenase**
2328	0708	1	1	*wcaA*	Putative uncharacterized protein
1503	1217	1	4	*glgX*	Putative hydrolase
1572	1267	1	1		Putative ABC transporter membrane protein
2877	2375	15	4		Putative type I secretion protein, ATP-binding protein
3049	2592	2	2	*rpoS*	RNA polymerase sigma subunit RpoS (Sigma-38)
4849	4005	3	1		Putative uncharacterized protein

Table 4.6: Strain-specific pseudogenes shared between Paratyphi A and Typhi - Genes that contained inactivating mutations in both Typhi and Paratyphi, but not in all isolates tested. SSPA and STY - systematic identifiers in Paratyphi A AKU_12601 and Typhi CT18 respectively. Ty - number of Typhi isolates (out of 19) that contain the inactivating mutation(s). Pa - number of Typhi isolates (out of 7 genomes plus 155 more in pools) that contain the inactivating mutation(s). Paratyphi C pseudogenes are highlighted in bold.

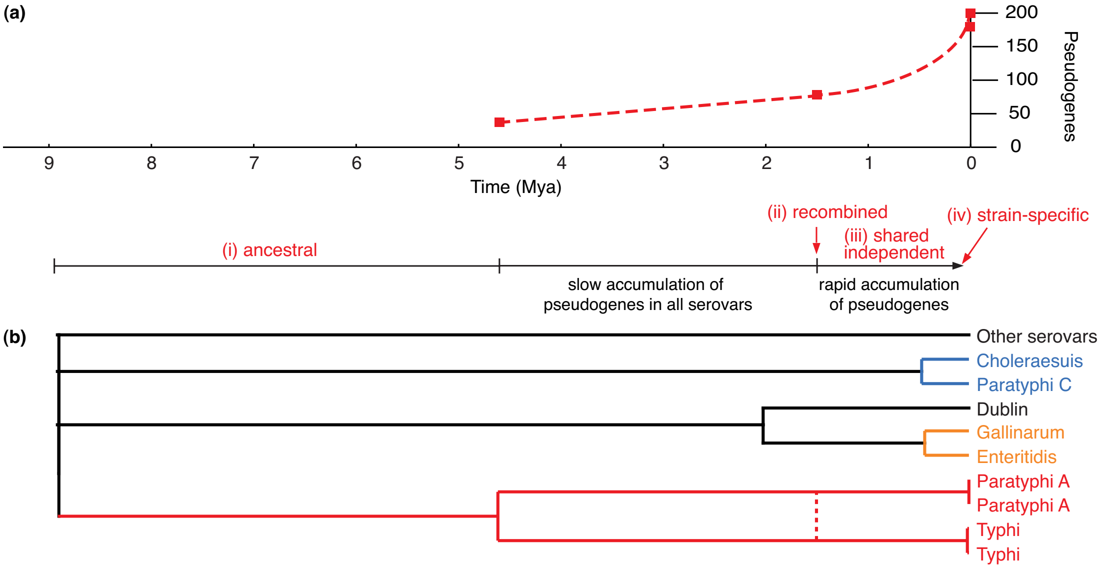


Figure 4.7: Pseudogene accumulation in Typhi and Paratyphi A over time - (a) Estimated number of pseudogenes in Typhi and Paratyphi A over time, with points corresponding to nodes in the phylogenetic tree. The number at the root is unknown; the number at the point of divergence, 39, is based on the analysis of ancestral pseudogenes inherited by Typhi and Paratyphi A from a common ancestor; the number at the point of recombination, 78, is estimated from the number shared during recombination (18 including 13 ancestral) and the scale of the recombination (23% of the genome): 18/0.23=78. Group (i) pseudogenes were inactivated prior to the divergence of Paratyphi A and Typhi, some are also inactivated in Typhimurium and other serovars; following their divergence Paratyphi A and Typhi likely accumulated few additional pseudogenes; during the recombination of 23% of their genomes (direction of transfer unknown) 18 pseudogene sequences were shared between Paratyphi A and Typhi, including five non-ancestral pseudogenes (group ii); many pseudogenes were formed during a period of accelerated pseudogene accumulation in both serovars, including most group (iii) pseudogenes; pseudogenes continue to accumulate in individual sub-lineages after the most recent common ancestor of each serovar (group iv). (b) Simplified representation of the phylogenetic trees shown in Figure 4.3. Scale is the same as in (a). The root represents the mean position calculated from Bayesian analysis of recombined and non-recombined gene sets, for the most recent common ancestor of *S. enterica* serovars, as shown in Figure 4.3. Positions of internal nodes represent the mean position calculated from Bayesian analysis of these gene sets, with the exception of red branches. The position of the most recent common ancestor for Typhi and Paratyphi A is that from Bayesian analysis of non-recombined genes as shown in Figure 4.3a; the dashed line represents the most recent common ancestor for Typhi and Paratyphi A estimated from Bayesian analysis of recombined genes as shown in Figure 4.3b, i.e. the estimated time of the recombination event between Typhi and Paratyphi A.

4.3.4.2 Pseudogenes potentially involved in host adaptation

Adaptation to the human host is most likely to be affected by mutations in genes that are directly involved in interactions between *Salmonella* and host. The most obvious candidates for such genes are secreted effector proteins, which are injected into host cells via the type III secretion system (656). Other natural candidates are genes associated with the production of cell-surface structures like flagella and fimbriae, and genes encoding proteins with transmembrane domains. The distribution of these functions among pseudogenes found in Typhi and/or Paratyphi A is shown in Table 4.7. Gene ontology analysis of pseudogenes in each group revealed no enrichment of particular biological processes or cell compartments among the inactivated genes.

Pseudogenes	**Effectors**	**Fimbriae**	**Transmembrane**	**Total**
(i) Ancestral	*sopD2*	0	1	39
(ii) Recombined	***sopA***	0	0	5
(iii) Shared independent	***slrP***	2	1	22
(iii) Paratyphi A specific	*sifB*	3	33	114
(iii) Typhi specific	*sopE2*	6	5	138
(iv) Paratyphi A and Typhi strains	0	1	7	22
(iv) Paratyphi A strains	0	1	32	130
(iv) Typhi strains	0	2	20	80
Paratyphi C only	0	3	19	108

Table 4.7: Pseudogenes in Typhi and Paratyphi A associated with secreted effectors, fimbriae or transmembrane domains - The number of pseudogenes in each group that fall into one of three functional categories: secreted effectors, fimbriae-associated, or contain transmembrane domains. The total number of pseudogenes in each group is given in the final column. Bold indicates genes also inactivated in Paratyphi C.

Three known secreted effector proteins were inactivated in both Typhi and Paratyphi A: *sopD2*, *sopA* and *slrP* (Table 4.7). *SopD2* is assumed to have been inherited in inactive form by both Typhi and Paratyphi A, as the same mutation (a 2 bp insertion) is present in both serovars yet the gene sequence was most likely not shared by recombination (divergence 1.7%). The inactive form of *sopA* was most likely shared via recombination, as the two gene sequences are only 0.2% divergent and contain the same nonsense SNP. *SlrP* on the other hand was inactivated by multiple independent mutations in Paratyphi A and Typhi, and is unlikely to have been recombined (divergence 1.9%). Only two other secreted effector proteins were inactivated in either Typhi (*sopE2*) or Paratyphi A (*sifB*). Interestingly, *sopA* and *slrP* were also inactivated in the Paratyphi C genome, by independent mutations to those found in Typhi or Paratyphi A. No other secreted effectors were identified as pseudogenes in Paratyphi C. Thus the inactivation of *sopA* and *slrP* may be a prequisite for systemic infection of the human host, whereas intact products of most other secreted effector genes may be essential for this kind of infection.

In total, six fimbrial genes were inactivated in some or all Paratyphi A strains, while eleven were inactivated in the Typhi population. Only three of the genes were overlapping (*safE, sefD, bcfC*) and were not considered to have been shared by recombination or ancestry, rather the mutations were relatively recent and independent (Table 4.7). (Note while the *sef* operon in SPI-10 was likely shared by recombination, *sefD* carries different frameshift mutations in Typhi and Paratyphi A.) Fimbriae play a role in adhesion to host cells, with genetic variants able to infect different hosts and even cell types (120, 640). The accumulation of fimbriae-associated pseudogenes in both Typhi and Paratyphi A may be associated with a narrowing of the range of host cells that the bacterial cells need to interact with. Alternatively there may be some adaptive advantages associated with preferential invasion of a different range of cell types.

Almost 20% of Typhi and Paratyphi A genes contain transmembrane domains (according to Hidden Markov modelling of transmembrane domains in all encoded CDS (657)), including many encoding transporters or receptor kinases. These genes were underrepresented among shared pseudogenes, making up only 4.5% of shared group (iii) pseudogenes and even fewer ancestral or recombined pseudogenes. Transmembrane-domain genes were overrepresented among Paratyphi A-specific pseudogenes (29%) but not Typhi-specific pseudogenes (4%). Many of the more recent, strain-specific inactivating mutations involved transmembrane-domain genes, including 32% of genes that were inactivated in members of both the Typhi and Paratyphi A populations and 25% of genes inactivated my members of either population (Table 4.7). This could be because inactivating mutations occur in transmembrane-domain genes at the same rate as others, but are not maintained in the population. Alternatively, this could indicate a recent acceleration in loss of function of membrane-spanning proteins. Most of these membrane-associated genes were either transporters or receptor kinases, which may have been needed to respond to particular environmental conditions or stimuli that are not encountered in the human restricted niche.

4.4 Discussion

4.4.1 Strengths and limitations of the study

Comparative re-annotation of the Typhi and Paratyphi A genomes revealed a number of pseudogenes that had been missed in previous annotations. This is to be expected, as it is hard to identify an incomplete or disrupted gene sequence without reference to a complete one. Gene finding software looks for open reading frames, but pseudogenes carrying multiple frameshift mutations can end up with multiple small open reading frames which are not easily recognizable as genes. Truncated CDSs may not be recognized as truncated without reference to full-length CDSs, and the 'right' version is difficult to determine in the absence of multiple sequences for comparison. Thus by comparing multiple Typhi and Paratyphi A genomes to Typhimurium, with the benefit of many other *Salmonella* and *E. coli* genomes for additional reference, a few additional pseudogenes would be identified. These include *safE* and *sefD*, fimbrial genes that contained independent inactivating mutations in Typhi and Paratyphi A but were missed from previous comparisons of Typhi and Paratyphi A pseudogenes (49).

Interpretation of the functional impact of genetic convergence in the form of shared pseudogenes, deletions and unique genes is difficult for a number of reasons. It would be optimal to study shared features of all human adapted serovars relative to other genomes, and to correlate features that are shared within subsets of human adapted serovars with particular pathogenic features. However, the availability of genome sequences is still lacking, as is characterisation of clinical and molecular features of the various serovars. There are currently no genome sequences available for systemic infection-associated Paratyphi B or Sendai, and neither of these have been extensively studied in terms of disease progression and clinical outcomes. Although a Paratyphi C genome was recently published (93), the disease syndrome caused by this serovar is not well understood. This is partly due to the difficulty of differentiating Paratyphi C from other serotype O6,7:c:1,5 *S. enterica*, including Choleraesuis, which can only be done definitively using a range of biochemical tests (622) or a combination of *IS*200 typing and ribotyping (658). For example, isolation of Paratyphi C from animals has been reported (659) and several isolates from pigs, typed as Paratyphi C, have been submitted to the *S. enterica* MLST database (464). However biochemical typing of Paratyphi C isolates in the database found that the pig isolates were not Paratyphi C, and indeed all 'real' Paratyphi C isolates in the database came from human infections and formed a single clonal group by MLST (Satheesh Nair, Sanger Institute/Health Protection Agency, personal communication, May 2009). The difficulty of accurately typing Paratyphi C has probably had a limiting effect on the reporting of infection with this serovar. There are currently very few reports of Paratyphi C infection in the literature, with most studies dating back to the 1930s-1980s in British Guyana (624) or Africa (619, 660). A 1933 study reported that Paratyphi C infection in humans was milder than infection with Typhi and Paratyphi A, and caused severe disease only in patients also infected with malaria (624). However confirmation of this, and studies of the clinical features of Paratyphi C infection, is lacking. Functional interpretations from comparative genomic analyses are also hampered by the lack of information on gene functions and pathways in *Salmonella* and in particular during human infection, which is experimentally intractable. However, genomic studies should help to develop hypotheses regarding gene function which can be experimentally tested in animals and human tissue.

This study benefits from historical clues about the relationship between Typhi and Paratyphi A, in particular the rapid burst of recombination 0.25-1.5 million years ago revealed by comparative genome analysis (56). This event serves as a historical marker (see Figures 4.2, 4.3 and 4.7), affording an additional insight into the temporal dynamics of gene degradation in Typhi and Paratyphi A. The population variation data, first presented in Chapters 2 and 3 of this thesis, add an additional historic marker in the form of the most recent common ancestor of each serovar. This allows the distinction to be made between gene degradation events that occurred prior to the last population bottleneck of each serovar and are therefore fixed in the populations, and those that have occurred much more recently. Together, these historical markers reveal that the rate of pseudogene accumulation in both Typhi and Paratyphi A increased dramatically following recombination between them (Figure 4.7). The majority of these became fixed in the last <25,000 years, although pseudogenes continue to accumulate in both serovars. Note though that the rate of this recent, strain-specific accumulation is likely to be underestimated by available variation data, which does not include small indel mutations (frameshifts) which are a common cause of gene inactivation.

The time estimates in this study are based on rates of either synonymous mutations (dS) or synonymous and nonsynonymous mutations (Bayesian estimation). The estimates based on synonymous mutations avoid inaccuracies associated with obvious selective pressures on nonsynonymous mutations, as well as mutations in intergenic regions which may be associated with regulatory or other functions under selection. However, synonymous mutations cannot be considered entirely neutral, as they may be subject to codon bias, transition bias or selective pressures on G+C content (580, 661). This may be problematic for the simple estimates of the age of Typhi and Paratyphi A (4.2.4) but should be accounted for in the calculation of dS for comparisons between serovars, which utilised a maximum-likelihood model incorporating estimates of transition bias and codon bias (650). Similarly, Bayesian analysis should account for much of these pressures using a GTR+Γ 2-site substitution model, which estimates separate substitution rates for codon positions one/two and three. Most importantly, there will be errors associated with the dating of divergence events, due to (a) the lack of reliable calibration dates and (b) the problem of substitution rate heterogeneity. The fossil record offers few clues for the ages of specific bacteria, providing calibration points only for extremely ancient events such as the oxidation of the Earth's atmosphere (45), the evolution of different Phyla (44, 45), or where there is good evidence for co-evolution with a eukaryotic host with a reliable fossil record (for example the endosymbiont *Buchnera aphidicola* (662) or the human pathogen *Helicobacter pylori* (663)). It is possible to calibrate the timing of evolutionary events using DNA sampled from different known time points. While this has proven successful in some cases, for example the analysis of viral RNA sequences subject to high mutation rates (664, 665), it is not helpful for bacterial DNA which evolves much more slowly.

In the absence of external calibration, bacterial divergence estimates usually rely on the idea of the 'molecular clock', which assumes that DNA substitutions become fixed at a constant or clock-like rate. However it is clear that different genes and different lineages evolve at different rates (666, 667, 668, 669). Furthermore there are numerous studies pointing to time-dependency of substitution rates (526, 670, 671, 672), suggesting that the rate over short time scales is dramatically higher (at least an order of magnitude) than that over long time scales. It has been suggested that this phenomenon is associated with a number of factors including purifying selection, genetic drift and effective population size (526, 670, 671, 672). The Bayesian analysis presented here incorporates a relaxed clock model, which allows substitution rates to vary on different branches of the phylogenetic tree. However, in the absence of independent information for dating coalescent events within the phylogenetic tree of *Salmonella*, this is unlikely to result in adequate correction for time-dependent rate variation. A rough *post hoc* adjustment for the effect of time dependency might be to compress the most recent events into a shorter period, which would make the accumulation of pseudogenes in Typhi and Paratyphi A more, rather than less, dramatic. It would also result in a downwards revision of the age of Typhi and Paratyphi A, so that the estimates given here should be considered upper bounds.

4.4.2 Implications for host restriction and adaptation

The sharing of DNA sequences via recombination must have resulted in increased similarity between Typhi and Paratyphi A at the DNA level. At the very least, the replacement of divergent alleles with identical ones brought these serovars closer together than other *S. enterica* genomes (56). However, this recombination likely resulted in convergence in gene content and function as well, via the sharing of insertions, deletions and pseudogenes between Typhi and Paratyphi A (4.3.3). Given their current convergence upon highly similar pathogenic phenotypes, it is tempting to suppose this led to shared features associated with host adaptation and systemic infection of humans and other simians. According to the dates estimated above (4.3.1), the recombination event occurred approximately 0.25-1.5 Mya, before the emergence of modern *Homo sapiens* ($\sim$0.2 Mya) (673, 674). Modern Typhi is able to cause typhoid-like disease in other simians including chimpanzees (*Pan troglodytes*) but not in prosimians such as rhesus macaques (*Macaca mulatta*) (33), which diverged $\sim$30 million years ago (674). Thus the process of adaptation to simians could have begun before the time of the recombination event, despite the absence of humans at this point.

The distribution of pseudogenes observed in this study suggests that the majority of pseudogenes present in the extant genomes of Paratyphi A and Typhi accumulated after recombination between a quarter of their genomes (Figures 4.6 and 4.7). How this relates to host adaptation and restriction, however, remains unclear. One possibility is that the recombination event directly contributed to host adaptation of one or both serovars, by generating a novel combination of genes and alleles. An alternative hypothesis is that both serovars were already host adapted, and recombination contributed to host restriction. A more tempting hypothesis might be that both serovars were already somewhat host adapted at the time of recombination, but learnt new tricks from each other by generating novel combinations of genes and alleles via recombination. The sharing (via recombination) of an inactive form of the secreted effector gene *sopA*, which remarkably is also a pseudogene in host adapted serovars Paratyphi B, Paratyphi C, Choleraesuis and Gallinarum (see 4.4.2.2 below), may be a clue that adaptation had already begun in Typhi and/or Paratyphi A at the time of their recombination. This would provide an opportunity for the recombination to occur, with both serovars circulating in a shared niche, perhaps in higher primates. At some point after the recombination, each serovar continued along a path to host adaptation, which has left a trail of pseudogenes scattered around their genomes. This was likely driven by a combination of adaptive selection for the loss of some functions, lack of selection against the loss of functions no longer needed and genetic drift associated with a narrowing host range. The accumulation of pseudogenes may have accelerated as each serovar became more adapted to systemic infection of simians and less capable of surviving in other niches, culminating in host restriction, the ultimate population bottleneck during which almost 200 pseudogenes became fixed in each population. The last such bottlenecks in Typhi and Paratyphi A almost certainly occurred less than 25,000 years ago, and possibly a lot more recently (see 4.3.1 above). At the time of these bottlenecks each serovar appears to have been restricted to the human population, as there have been no reports of isolation from animal or environmental sources (despite the finding that deliberate infection with Typhi can cause typhoid-like disease in chimpanzees (33)). The accumulations of pseudogenes since the most recent bottlenecks are most likely explained by continued loss of gene functions that are not required in the human systemic niche.

4.4.2.1 Ancestral pseudogenes

The inactivating mutations in group (i) pseudogenes are assumed to have been inherited by Paratyphi A and Typhi from a common ancestor (Figure 4.7). Alternatively some may have been exchanged between Paratyphi A and Typhi soon after their divergence from other *S. enterica*. Either way, these pseudogenes were among the earliest to arise in the evolutionary history of Paratyphi A and Typhi, thus their inactivation has been well tolerated in these serovars (most have also accumulated secondary mutations). This is unsurprising for the majority of ancestral pseudogenes which are IS, transposase or phage genes/fragments. However the inactivation of seven genes known to be functional in Typhimurium and other *Salmonella*, in particular those that are secreted or surface exposed (Table 4.3) may have had some functional impact including potential modulations of host interactions. The best described of these seven co-inherited pseudogenes is the secreted effector protein *sopD2*. *SopD2* is broadly conserved among *Salmonella* (including intact coding sequences in host adapted serovars Paratyphi B, Paratyphi C, Choleraesuis, Dublin, Gallinarum) and shares a high degree of sequence similarity with *sopD*, likely resulting from a gene duplication event (510, 675). However their functions are complementary rather than redundant in Typhimurium (675), so it is likely that the inactivation of *sopD2* in Typhi and Paratyphi A has functional consequences although their *sopD* sequences remain intact. Studies in Typhimurium have shown *sopD2* is secreted via the SPI2 type III secretion system and is associated with *Salmonella*-induced filaments on the surface of infected host cells (675, 676). It is involved in the formation of these filaments (675) and is also an important factor in inhibition of antigen presentation by murine dendritic cells (677). Furthermore, bacterial replication of a Typhimurium *sopD2* knockout mutant was impaired in murine macrophages but not in human epithelial cells (675). *SopD2* therefore constitutes a plausible candidate for an early modulator of host interactions in Paratyphi A and Typhi, although it should be noted that both *sopD2* and *sopD* are intact in the Paratyphi C genome. Interestingly, Paratyphi B *sensu stricto* isolates causing systemic infection in humans lack expression of SopD protein (225), although expression of SopD2 has not been studied.

4.4.2.2 Pseudogenes and novel genes shared by recombination

Of the 17 genes shared by Typhi and Paratyphi A but absent from all other available serovar genomes, 14 were consistent with sharing via recombination (<0.3% divergence, see Table 4.2c). Of the 39 genes deleted from both Typhi and Paratyphi A relative to other serovars, only four were consistent with shared deletion via recombination (identical deletion boundaries, flanked by genes of <0.3% divergence, see Table 4.4). Group (ii) contains five pseudogenes shared by recombination (Table 4.3). Therefore recombination between Typhi and Paratyphi A must have resulted in novel combinations of genes and allelic variants, and therefore novel combinations of protein functions, in one or both serovars (depending on the directionality of DNA transfer). This could have contributed to host adaptation or restriction in the recipient serovar(s), and certainly led to genetic convergence between Typhi and Paratyphi A which must have played at least a minor role in their pathogenic convergence. A mutation in *sopA* (one of just five secreted effector proteins inactivated in Typhi or Paratyphi A (Table 4.7)) appears to have been shared by recombination. The *sopA* gene carries different inactivating mutations in the host adapted serovars Paratyphi C, Choleraesuis and Gallinarum, and the Paratyphi B dT- genome of SPB7, but was intact in the other sequenced serovars (Table 4.1). The SopA effector mimics mammalian ubiquitin ligase and can target bacterial and host proteins for degradation by the human ubiquitination pathway, (678, 679, 680). SopA preferentially uses inflammation-associated host E2 enzymes for the ubiquitination reaction (679), which may indicate a role in bacterial regulation of host inflammation. The *sopA* gene is necessary for virulence in both murine systemic infections and bovine gastrointestinal infections by Typhimurium (681, 682), thus is clearly important for interactions between *Salmonella* and mammalian hosts. It is plausible therefore that the loss of this gene in Paratyphi A, Typhi and other host adapted serovars has been important for their evolution, perhaps by facilitating systemic infection or the establishment of long-term carriage.

4.4.2.3 Recent pseudogenes: convergence after recombination

In addition to >100 pseudogenes specific to each serovar, group (iii) includes 22 shared pseudogenes containing different inactivating mutations in Paratyphi A and Typhi (Table 4.3). While it is possible that some of those lying outside recombined regions may have been present prior to recombination, it is likely that most of these mutations arose in the period of rapid pseudogene accumulation after recombination. These pseudogenes are examples of convergent gene loss through independent mutation, and are therefore good candidates for involvement in adaptation to the human host. They include only one transposase gene, the remainder being genes of known or putative function, many of which have been implicated in host interactions in serovar Typhimurium (e.g. *fhuA, fhuE, shdA, ratB, sivH, slrP*) (49, 683).

It is not possible to distinguish whether there has been adaptive selection against the activity of these genes in Paratyphi A and Typhi, or simply shared tolerance for their inactivation. For example, it has been noted (49) that three of these genes (*shdA, ratB* and *sivH*, part of the 25 kbp pathogenicity island CS54 (683)) are involved in intestinal colonisation and persistence, which does not occur in typhoid or paratyphoid infection. However we cannot distinguish whether the independent inactivation of these genes in each serovar is due to selection against colonisation of the intestine (which may stimulate host immune responses), or genetic drift since intestinal colonisation is not required to sustain a systemic infection. These genes were not annotated as pseudogenes in the Paratyphi C genome, but do appear to be disrupted compared to the coding sequences present in the closely related swine adapted serovar Choleraesuis and in Typhimurium. This is consistent with either selection or tolerance for loss of function, although selection seems a more plausible explanation for the independent inactivation of both genes in three human-adapated serovars. The genes are not present in Gallinarum or Enteritidis, but are intact in Choleraesuis and Paratyphi B SPB7. A similar pattern was observed for the secreted effector protein *slrP*, which carries independent inactivating mutations in Typhi, Paratyphi A and Paratyphi C, is missing from Enteritidis and Gallinarum, but is present intact in Paratyphi B, Choleraesuis and many other serovars. A better understanding of the functions of these pseudogenes and their distribution among serovars with different pathogenic phenotypes may be able to distinguish negative selection from tolerance of gene loss. Regardless, it is clear that a rapid accumulation of pseudogenes occurred at some point after the recombination between Typhi and Paratyphi A, and became fixed during subsequent population bottlenecks.

It should be noted that inactivation of different genes in the same pathway will often result in similar loss of function, thus the true contribution of pseudogene formation to phenotypic convergence between Typhi and Paratyphi A is likely underestimated by considering only those pseudogenes or deletions that are shared. For example, it was noted previously that different members of the *cbi* cluster were inactivated in Typhi CT18, Ty2 and Paratyphi A ATCC9150, which may result in similar inactivation of the cobalamin synthesis pathway (49). The variation study presented in Chapter 3 detected nonsense SNPs within the Paratyphi A population in two of the four *cbi* genes that were inactivated in Typhi (*cbiJ, cbiK*) and in an additional gene *cbiQ*. Another of the four genes, *cbiC* was found in Chapter 2 to be intact in two Typhi isolates (E02-1180, E01-7866), making it a strain-specific pseudogene in the Typhi population. These findings provide further evidence of ongoing degradation of the cobalamin synthesis pathway in both Typhi and Paratyphi A, although the operon appears to be intact in Paratyphi C.

4.4.2.4 Ongoing accumulation of strain-specific pseudogenes

The comparative analysis of whole genome variation in 19 Typhi strains inferred that their last common ancestor harboured only 180 pseudogenes, while individual isolates had each accumulated at least 10-28 additional pseudogenes since their divergence from that ancestor (2.3.4.2). Comparative analysis of the AKU_12601 and ATCC9150 genomes identified 22 mutations resulting in strain-specific pseudogene formation (10-12 per strain, Table 3.4), while analysis of additional Paratyphi A samples identified inactivating mutations in a further 131 genes. The numbers for both Typhi and Paratyphi A were predicted to be an underestimate, as these studies did not take into account pseudogene formation via insertion/deletion of one or two nucleotides which would introduce frameshifts. These strain-specific pseudogenes must have arisen since the most recent common ancestors of the respective Paratyphi A and Typhi populations and are therefore more recent than the serovar-specific pseudogenes which have become fixed in each population (see Figure 4.7). This ongoing gene loss is likely to be associated with tolerance for loss of functions not required in the human systemic niche, rather than adaptive selection.

Chapter 5

IncHI1 multidrug resistance plasmids in Paratyphi A and Typhi

5.1 Introduction

Antibiotic treatment is central to the control of enteric fever. Although vaccines against Typhi have been available for a long time, they are used mainly in travellers rather than inhabitants of endemic areas (15, 425, 429) and there is currently no vaccine that provides strong protection against Paratyphi A (280, 431). Chloramphenicol resistant Typhi was first reported in 1950 (363), just two years after the introduction of chloramphenicol for the treatment of typhoid fever (362). By the early 1970s, Typhi resistant to both chloramphenicol and ampicillin had been observed (364), and multidrug resistant (MDR) Typhi emerged soon after (268, 684), see Figure 1.7. MDR has been broadly defined as resistance to three or more first-line antibiotics, but in most studies of enteric fever refers to resistance to chloramphenicol, ampicillin and co-trimoxazole (e.g. (16, 327)). Since the 1990s the recommended treatment for enteric fever has changed to quinolones and more recently fluoroquinolones (297, 685), although susceptibility to these drugs has been declining among Typhi and Paratyphi A isolates since this time (16). Rates of MDR Typhi fluctuate over time and geographical locations (see for example (16, 327)), but MDR is still a problem in many areas despite the switch to fluoroquinolones (16). For example, in southern Vietnam, over 80% of Typhi isolates tested in 2005 were MDR (16). Unlike Typhi, Paratyphi A isolates have predominantly been susceptible to antibiotics (384, 385). However, in recent years the incidence of MDR Paratyphi A has increased, particularly in Pakistan and India where rates as high as 45% of Paratyphi A isolates have been reported (386, 387, 388). Higher rates of fluoroquinolone resistance among Paratyphi A isolates compared to Typhi isolates have been reported (279, 336, 686), and a recent study in Nepal found MDR was more common among Paratyphi A than Typhi isolates (687). The situation is perhaps most extreme in China, where in some regions more than 50% of enteric fever is now caused by Paratyphi A and is largely drug resistant (285, 335). Among the 159 Paratyphi A isolates examined in pools in Chapter 3, the GyrA-Phe83 SNP which confers increased fluoroquinolone resistance (391, 393) was detected in 17 pools, with an estimated frequency of 55 isolates (35%).

MDR in Typhi is almost exclusively plasmid-mediated, with the majority of plasmids analysed being of the HI1 incompatibility type (IncHI1) (268, 269, 270, 271, 272, 273, 274, 275, 276), although other plasmids have been reported (277). The few studies which have reported MDR in Paratyphi A have pointed to a key role for plasmids in mediating resistance although few molecular studies have been undertaken (688). A large transferable plasmid of 140 MDa (∼230 kb, similar in size to the 120-200 kbp IncHI1 MDR plasmids found in Typhi) was found in 73% of MDR strains in Bangladesh in 1992 to 1993 (284). A similarly sized plasmid was reported in recent Chinese Paratyphi A isolates (285). However, in Calcutta, India, a smaller plasmid (∼55 kb) was responsible for conferring MDR in Paratyphi A isolates (286). Recent work by Satheesh Nair at the Sanger Institute demonstrated that in Paratyphi A isolated from Pakistan between 2002-2004, MDR was associated with IncHI1 plasmids of approximately 220 kbp (283).

The prototype IncHI1 plasmid is R27, isolated from Typhimurium in 1961 (269) (note that some manuscripts have incorrectly reported that R27 was isolated from Typhi). R27 is self-transmissible and can transfer between *Enterobacteriacae* and other gram-negative bacteria (689). Conjugal transfer is encoded in two regions, *tra1* and *tra2* (690). *Tra2* contains genes for pilus production and mating pair formation (691), while *tra1* contains genes required for translocation of DNA across the bacterial cell membrane and initial replication upon entering the recipient cell (692). Conjugal transfer of IncHI1 plasmids is thermo-sensitive, with maximum transfer efficiency at ambient temperatures (14-27°C) and highly impaired transfer rates at 37° (689, 692). The finished sequence of R27 was published in 2000 (692), followed one year later by that of pHCM1, the IncHI1 plasmid of the sequenced Typhi isolate CT18 isolated in 1993 (46). The two plasmids shared 168 kbp with 99% sequence identity, including the *tra1* and *tra2* regions, but differed in their resistance gene insertions. R27 contained just one resistance locus - *Tn*10, encoding resistance to tetracycline. The Typhi plasmid pHCM1 contained the same *Tn*10 transposon, along with several other mobile elements including integrons and the transposons *Tn*21 and *Tn*9, carrying several drug resistance genes: *dhfR14* (trimethoprim resistance), *sul2* (sulfonamide resistance), *catI* (chloramphenicol resistance), *bla* (TEM-1; ampicillin resistance) and *strAB* (streptomycin resistance). These transposons coincide with those found in IncHI1 Typhi plasmids in the early 1970s, including *Tn*3 (of which *Tn*21 is a subtype) encoding ampicillin resistance, *Tn*9 encoding chloramphenicol resistance and *Tn*10 encoding tetracycline resistance (269).

The Paratyphi A isolate AKU_12601, isolated in Pakistan in 2002 and presented in Chapter 3, was multidrug resistant and contained a 212 kbp IncHI1 plasmid. The plasmid, pAKU_1, was also sequenced and finished at the Sanger Institute. This Chapter begins with the annotation of the pAKU_1 sequence and comparison of this plasmid sequence to those of R27 and pHCM1. The remainder of the chapter focuses on phylogenetic relationships between these and more recently sequenced IncHI1 plasmids.

5.1.1 Aims

The aims of this chapter were to annotate the recently completed sequence of the Paratyphi A MDR IncHI1 plasmid pAKU_1 and compare the sequence to that of other available IncHI1 plasmids. Specific aims of the analysis were to:

- identify the conserved backbone of the IncHI1 plasmid and determine the phylogenetic relationships between available plasmid sequences;
- determine the mobile genetic elements encoding MDR in pAKU_1 and other IncHI1 plasmids, and examine their distribution among the population of IncHI1 plasmids; and

- gain insight into the spread of drug resistance via IncHI1 plasmids in Typhi and Paratyphi A.

5.2 Methods

5.2.1 Annotation

The 212,711 bp plasmid pAKU_1 was sequenced and assembled by members of the Pathogen Sequencing Unit at the Sanger Institute using capillary-based Sanger sequencing. Gene prediction was performed by Nick Thomson at the Sanger Institute using Glimmer. Predicted genes or CDSs were assigned systematic identifiers with the prefix 'SPAP'. The nucleotide sequences and translated protein sequences of CDSs were used to batch query several databases and motif recognition programs to assist with annotation: BLASTN search of EMBL database (nucleotide similarity), fasta3 search of EMBL (protein similarity), Pfam search (protein domains (693)), TMHMM analysis (transmembrane hidden Markov model to identify protein transmembrane domains (657)), SignalP (to identify signal peptides, indicative of protein export (694, 695)), HTH (to identify helix-turn-helix DNA-binding motifs within proteins (577, 696)) and tandem repeats (to identify tandem repeats within coding sequences). Each CDS was annotated manually in EMBL format, using Artemis (697) to coordinate the display of results from the aforementioned database searches for consideration during annotation. Start and stop codons were determined on the basis of (a) similarity with known proteins, (b) presence of Shine-Dalgarno sequence (AGGAGG) (698) and (c) GC frame plots across the CDS and flanking sequence. Annotation included assignment to a functional category listed in Table 5.1, using the '/colour' identifer in the EMBL format.

These functional assignments took into account evidence from the database searches described above (e.g. genes with transmembrane domains identified using TMHMM would be assigned to category 3). Annotation of IS elements was done using the IS finder database (699) to identify the coding sequence for the transposase gene and flanking inverted repeat sequences. Transposons were annotated with reference to previously described transposon sequences and structures identified using sequence homology and literature searches. Each CDS was assigned a confidence value (identifier '/confidence_level') for the annotation: 1=experimental evidence of function in this

Category	Function
0	Pathogenicity/Adaptation/Chaperones
1	Energy metabolism (glycolysis, electron transport etc.)
2	Information transfer (transcription, translation, DNA/RNA modification)
3	Surface (inner or outer membrane, secreted, surface structures)
4	Stable RNA
5	Degradation of large molecules
6	Degradation of small molecules
7	Central, intermediary or miscellaneous metabolism
8	Unknown
9	Regulators
10	Conserved hypothetical
11	Pseudogenes and partial genes (remnants)
12	Phage, IS elements
13	Some misc. infomation e.g. Prosite, but no function

Table 5.1: Functional categories for genome annotation - Each open reading frame is assigned to a category using the /colour identifier in the EMBL-style annotation.

species, 2=experimental evidence of function in related species, 3=presence of functional domains, 4=no real evidence of function.

5.2.2 Sequence comparison and SNP detection

Plasmid sequences were compared using BLASTN and the comparisons viewed in the Artemis Comparison Tool (ACT) (604), which aids visualisation of synteny and similarity between pairs of sequences. SNPs between finished plasmid sequences were identified using MUMmer 3.1 (576). MUMmer's `nucmer` algorithm was used to align the nucleotide sequences of R27 and pHCM1 to pAKU_1, and its `show-snps` algorithm was used to detect SNPs based on this alignment. SNPs were identified from Solexa-sequenced Typhi IncHI1 plasmids by mapping reads generated from the three Typhi isolates to the finished pAKU_1 sequence, using Maq (564) and quality filters described in 2.3.1.3.

5.2.3 Phylogenetic analysis

SNPs called in repetitive regions or inserted sequences were excluded from phylogenetic analysis, so that phylogenetic trees were based only on the conserved IncHI1 backbone sequence. SNP alleles were concatenated to generate a multiple alignment of SNP alleles. A maximimum likelihood phylogenetic tree was fit using RAxML (644) with a GTR+Γ model and 1,000 bootstraps. A phylogenetic network was also constructed, using the parsimony splits method implemented in SplitsTree4 (603) as described earlier (3.2.3).

5.2.4 PCR

PCR primers were designed using Primer3 (700) according to the following criteria: melting temperature 56°C, no hairpins or dimers affecting $3'$ ends, no cross-dimers between forward and reverse primers. Primer sequences are given in Table 5.2. Primers were designed by myself, PCR assays were performed by Minh Duy Phan (Sanger Institute).

5.2.5 Accession codes

The annotated plasmid sequence was submitted to the EMBL database with accession number AM412236. Plasmid seqeunces used for comparative analysis were: R27 - AF250878; pHCM1 - AL513383; R478 - BX66401; pRSB107 - AJ851089); pU302L - AY333434; DT193 - AY524415; IncI plasmid of Enteritidis - AJ628353; pMAK1 - AB366440.

Label	Forward and reverse primer	Length pAKU_1	Length pHCM1	Target feature
IncHI1	CGAAATCGGTCCAACCCATTG CGACAACTCATCAGAAGCGTCAAC	110	-	*repHI1A*
A	GAAAGGAATCATCCACCTTCA AACTGTCGCTACGCCTGACT	419	990	del on pAKU_1
B	ATCCAGCGTGCAAAGATTTC TGGGGGAGAACACCACTTTA	407	2589	del on pAKU_1
C	AAAGATGCAATGGGAGGAGA GCGCAGCTGCTTCAATTA	289	4399	insert on pHCM1
D	TAGGGTTTGTGCGGCTTC CCTTCTTGTCGCCTTTGC	3138	none	insert on pAKU_1
E	TCAAGGCAGATGGCATTCCC CGACGAGTTTGGCAGATGATTTC	156	none	*sul1*
F	GTGTCGAGGAAAGGAATTTCAAGCTC TCACCTTCAACCTCAACGTGAACAG	191	none	*dhfR7*
G	GATGGAGAAGAGGAGCAACG TTCGTTCCTGGTCGATTTTC	989	989	*bla/sul/str-Tn*21 (*strB/tniA*Δ)
H	GTGCTGTGGAACACGGTCTA TCATCAACGCTTCCTGAATG	271	1598	*Tn*21-*Tn*9 (*Tn*21 *tnpA*/*Tn*9 acetyltransferase)
I	ACGAAAGGGGAATGTTTCCT CGAGTGGGAATCCATGGTAG	163	1490	*Tn*21-*Tn*9 (*merR*/*Tn*9)
J	CAAAATGTTCTTTACGATGCC CCAGACAGGAAAACGCTCA	2219	none	*Tn*9-*tra2* (*cat*/*trhN*)
K	CTGTGCCGAGCTAATCAACA ACGAAAGGGGAATGTTTCCT	1314	none	*Tn*21-*Tn*9-*tra2* (*merR*/*trhI*)
L	TTTTAAATGGCGGAAAATCG GCCAGTCTTGCCAACGTTAT	none	1872	*Tn*9-*Tn*10 (*insA*/*tetA*)
M	GGGCGAAGAAGTTGTCCATA ATTCGAGCAAAACCATGGAA	none	2195	*Tn*9-backbone (pHCM1 site) *cat*/HCM1.203
N	CGGGATGAAAAATGATGCTT GGTCGGTGCCTTTATTGTTG	none	2180	*Tn*10-backbone (pHCM1 site) *Tn10*/HCM1.247
O	GCGTACAAAAGGCAGGTTTG GCTTGATGATGTGGCGAATA	1823	none	*Tn*10/backbone (pAKU_1 site) *tetD*/SPAP0276
P	TGGTCGGTGCCTTTATTGTT GGGCGTCAGAGACTTTGTTC	1899	none	*Tn*10/backbone (pAKU_1 site) SPAP0261/*Tn10*
Q	TTCGCCCGATATAGTGAAGG CTAACGCCGAAGAGAACTGG	1923	none	*strAB*/backbone (pAKU_1 2nd copy) *strA*/SPAP0228

Table 5.2: PCR primers for analysis of IncHI1 plasmids - Used to detect features of the IncHI1 backbone and resistance gene insertions. The locations of A-Q in pAKU_1 and/or pHCM1 are shown in Figures 5.2 and 5.4.

5.3 Results

5.3.1 Characterisation of IncHI1 plasmid backbone and resistance gene insertions

Coding sequences in the pAKU_1 plasmid were annotated as described in 5.2.1. The 212,711 bp plasmid contained 237 coding sequences, the majority of which are of unknown function. Figure 5.1 shows the distribution of annotated functional groups, as well as a more specific breakdown of plasmid and resistance functions.

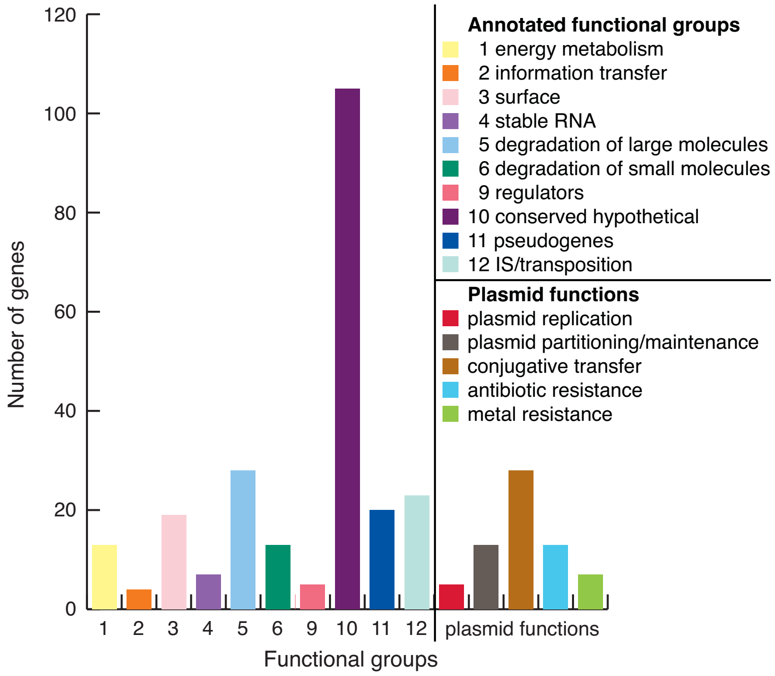


Figure 5.1: Functions of genes annotated in the IncHI1 plasmid pAKU_1 from Paratyphi A - A total of 237 coding sequences were annotated and assigned to functional groups 0-13 as defined in Table 5.1; these are labelled with the corresponding group numbers on the x-axis. A subset of these genes (66) have known functions associated with multidrug resistance plasmids, these are further divided into specific plasmid and resistance functions (labelled 'plasmid functions' on the x-axis).

5.3.1.1 The conserved IncHI1 backbone

Comparison with EMBL/GenBank sequence databases (January 2007) revealed high DNA sequence similarity between pAKU_1 and the IncHI1 plasmids R27 (692, 701) and pHCM1 (46). Detailed comparative analysis of the three plasmid sequences revealed a 164.4 kb shared IncHI1-associated backbone, with 99.7% nucleotide identity among the three plasmids. This shared backbone constitutes 83% of pAKU_1 sequence and includes the IncHI1 incompatibility locus, the *tra1* and *tra2* conjugative transfer regions (690, 691) and three potential replicon elements (RepHI1A, RepHI1B, RepFIA) characteristic of IncHI1 plasmids (702), the locations of which are indicated in Figure 5.2a. The rest of the shared backbone harbours genes involved in the core plasmid functions of replication, maintenance and conjugative transfer, as well as many hypothetical genes with no database matches to sequences outside IncHI1 and the *Serratia marcescens* IncHI2 plasmid R478 (703) (EMBL/GenBank, June 2009).

The shared IncHI1 backbone sequences of pAKU_1, R27 and pHCM1 were aligned and nucleotide differences determined using MUMmer (5.2.2). This analysis found pAKU_1 shared 99.71% nucleotide identity with pHCM1 and 99.89% with R27. Figure 5.2b shows an unrooted phylogenetic tree based on the number of single nucleotide changes found between the three IncHI1 backbones. As the tree shows, the plasmid backbone of pAKU_1 was closer to that of R27 than pHCM1. This is supported by the presence of shared variations in the backbones of pAKU_1 and R27 relative to pHCM1. These are marked (**) in Figure 5.2a and include a small inversion near the $5'$ end, two deletions downstream of this inversion and a gene (annotated as R0107 in R27 and SPAP0320 in pAKU_1) inserted at the $3'$ end of the shared backbone. A large region was inverted on pAKU_1 relative to R27 and pHCM1, however this occurred on pAKU_1 or a similar precursor plasmid rather than a common ancestor of R27 and pHCM1 (see Figure 5.4).

The 19 kbp *tra1* region contains nine genes essential for transfer, the *oriT* (origin of transfer) site and a further 4-5 CDSs of unknown function (692). *Tra1* was conserved as a single sequence block with >99.9% identity across pAKU_1, pHCM1 and R27, differing from each other by only 14-16 bp. Few of these changes occurred in coding regions: nearly all the encoded proteins were 100% identical at the amino acid

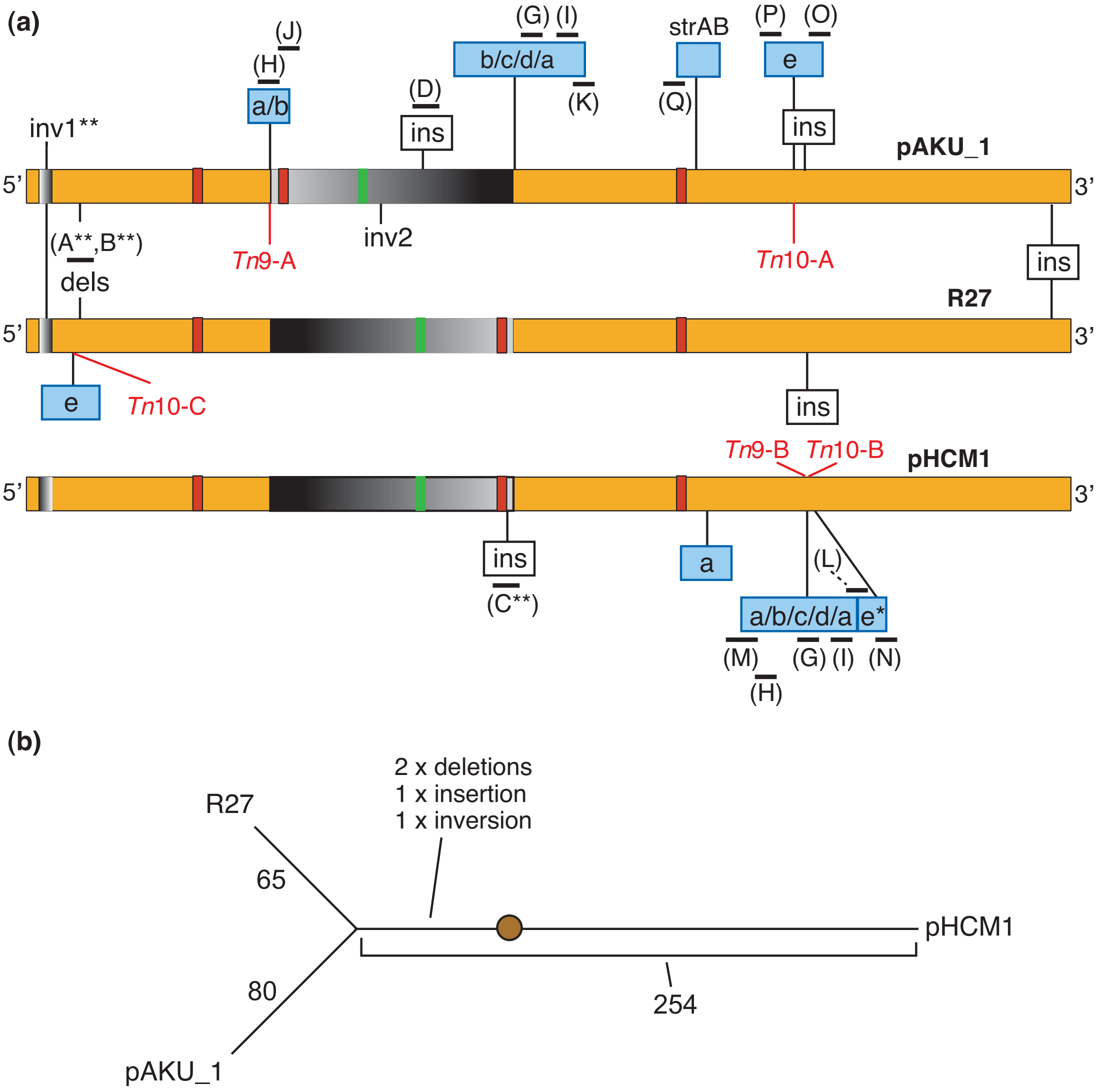


Figure 5.2: Comparison of three complete IncHI1 plasmid sequences from *Salmonella* - (a) Representative alignment of the 164kb IncHI1 backbone sequences of pAKU_1, R27 and pHCM1, with the sites of major insertions, deletions and inversions indicated. Note that the plasmids are actually circular and are shown as linear here merely for ease of comparison. Red boxes show the sites of IncHI1 replicons, green box represents the incompatibility region. Blue boxes represent resistance gene insertions, scaled to indicate relative size compared to backbone, and are labeled as in Figure 5.3 (a=*Tn*9, b=*Tn*21, c=Class I integron, d=*bla/sul/str*, e=*Tn*10, e*=truncated *Tn*10). Black bars indicate PCR target amplicons, labeled as in Table 5.2 and Figure 5.4. Transposon insertion sites are labelled in red, corresponding to labels given in the text and Figure 5.6. Other insertions are shown in white boxes (ins), note that the insertion targeted by PCR D is *Tn*6062. Inversions are shown as graded black/grey boxes, gradient indicates direction. (b) Tree showing the relationship between 164kb IncHI1 backbone sequences of the three plasmids, based on SNPs. Branch lengths are proportional to the number of SNPs, indicated next to branches. The position of four major differences (** in a) is indicated; the position of the root is imprecise due to lack of suitable plasmid sequences for use as outgroups.

level, with the exception of the pAKU_1-encoded *trhG* which differed from the R27 and pHCM1 orthologs at two amino acid residues (1253 and 1321) and from the pHCM1 ortholog at a third residue (670). The *oriT* site was completely conserved across all three plasmids (although not precisely defined, the site lies in the region between *traH* and *trhR*, which was 100% identical). The 23 kbp *tra2* region contains genes required for synthesis of the H-pilus, mating pair stabilisation and DNA transfer, genes required for plasmid partitioning and stability and the plasmid incompatibility region (*inc*). Apart from some insertions and rearrangements within *tra2* in pAKU_1 and pHCM1, the sequences first characterised in R27 (690, 691, 692) were highly conserved among pAKU_1, pHCM1 and R27 (>99.85% nucleotide identity). Most of the *tra2*-encoded proteins were 100% identical at the amino acid level, with single amino acid differences in four pHCM1 orthologs (*trhC*, *parA*, *orf16*, *trhP*), one R27 ortholog (*trhW*) and three residue changes in the pAKU_1 ortholog of *orf9*.

5.3.1.2 Comparison of drug resistance genes in pAKU_1 and pHCM1

The pAKU_1 plasmid sequence contained multiple antibiotic resistance gene elements inserted into the IncHI1 backbone. These insertions were highly clustered relative to the conserved IncHI1 backbone and were related to antibiotic resistance genes found on pHCM1 but not R27 as shown in Figure 5.2a. These resistance genes can be attributed to the insertion of previously described transposable elements (see Figure 5.3) into different positions in the plasmid backbones.

*Tn*10, carrying genes for tetracycline resistance (*tet*) (704, 705, 706) (Figure 5.3e), was present on pAKU_1 as well as R27 (706) and pHCM1 (46), although part of the transposon was missing from pHCM1. The insertion of *Tn*10 is mediated by flanking elements *IS*10-left and -right (704), and insertion of the transposon generates 9 bp direct repeats of the target sequence (target site duplications) (707). The site of *Tn*10 insertion into the backbone was different in each plasmid (see Figure 5.2a), generating distinct flanking repeats in each case and indicating that the transposon was independently acquired in each plasmid rather than by a common ancestor. No further resistance insertions were present on R27 (692).

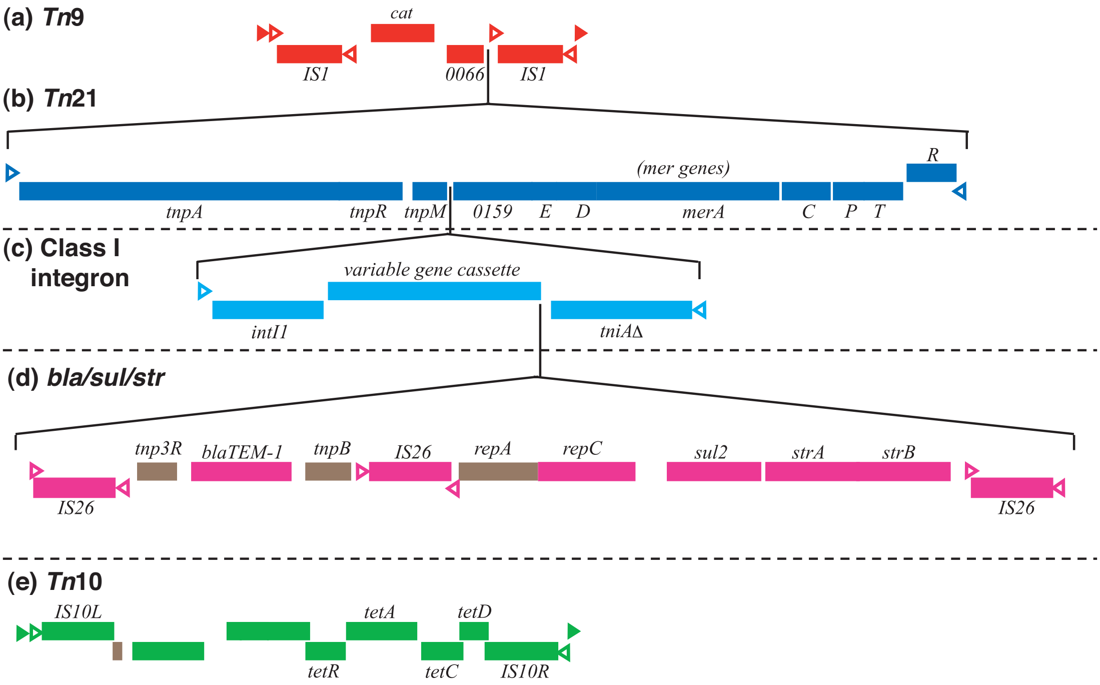


Figure 5.3: Transposons identified in pAKU_1 - The gene order shown here agrees with the consensus from other sequences; note that *Tn*9 and *Tn*21 have been disrupted by rearrangements in pAKU_1 and pHCM1 (see Figure 5.4a,d). Upper and lower bands represent forward and reverse strands; open triangles represent inverted repeats, filled triangles represent direct repeats (target site duplications). The insertion sites shown for (d) into (c), (c) into (b) and (b) into (a) are conserved in pAKU_1, pHCM1 and pRSB107.

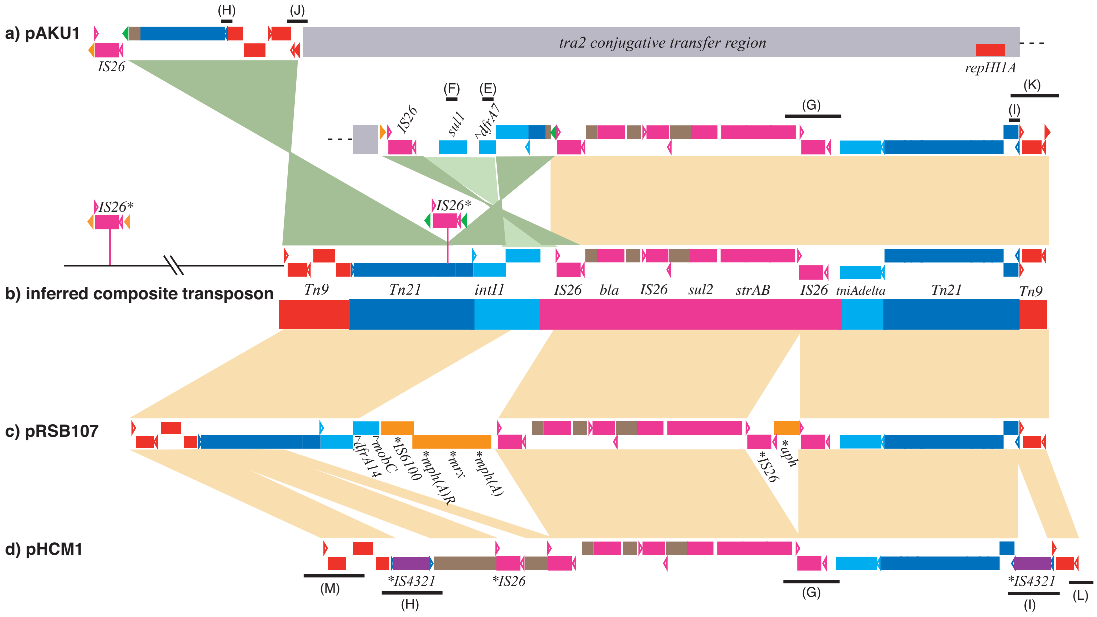


Figure 5.4: Rearrangements of composite transposons inserted in IncHI1 plasmids - The inferred ancestral version of the composite transposon (b) and its rearrangements in three plasmids. *=genes inserted relative to the composite transposon; ˆ=genes in the variable integron gene cassette. Colours and genes are the same as in Figure 5.3; upper and lower bands represent forward and reverse strands; open triangles represent inverted repeats. Filled triangles represent target site duplications: green="TTGCGCCG"; orange="AAAAAAAG". Regions of identical sequence are joined by coloured boxes (tan when in direct orientation, green when inverted). Black lines indicate PCR amplicons, labels correspond to those in Table 5.2 and Figure 5.2.

*Tn*9, with identical copies of the chloramphenicol resistance gene *cat* (368) (Figure 5.3a) was present on pHCM1 (46) and pAKU_1. The transposition of *Tn*9 is accompanied by 9 bp target site duplications (708). Distinct insertion sites, accompanied by distinct target site duplications, were evident in the two plasmids, suggesting that *Tn*9 was inserted independently into the backbones of pAKU_1 and pHCM1 (sites shown in Figure 5.2a).

*Tn*21, harbouring a class I integron and mercury resistance (*mer*) operon (709) (Figure 5.3b-c) was also identified on both pHCM1 (46) and pAKU_1, although in each case the primary transposable element has been disrupted by *IS*26 insertions and subsequent sequence rearrangements (Figure 5.4). *Tn*21 was inserted at the same site within *Tn*9 in pAKU_1 and pHCM1; pHCM1 also harbours a second, divergent copy of *Tn*21 elsewhere on the plasmid (a in Figure 5.2a). The resistance gene cassettes associated with the class I integrons (Figure 5.3c) differ in the two plasmids: *dhfR7* in pAKU_1 and *dhfR14* in pHCM1 (both encoding trimethoprim resistance (710, 711)). The *sul1* gene, encoding resistance to sulfonamides (712), was adjacent to *dhfr7* in pAKU_1. *Sul1* is frequently associated with class I integrons, however it is not believed to be part of the integron cassette (709).

An identical ∼9 kb sequence, incorporating bla_{TEM-1} (a beta-lactamase), *sul2* (sulfonamide resistance (713)) and *strAB* (streptomycin resistance (714)) genes flanked by *IS*26 elements (Figure 5.3d), was present on both pHCM1 (46) and pAKU_1. BLAST searching (most recently in June 2009) revealed that this is a promiscuous sequence, referred to hereafter as *bla/sul/str*, that is also present in the 120 kb IncF plasmid pRSB107 (unknown host, Germany, 2005) (715) and the F-like plasmid pU302L of Typhimurium strain G8430 (Centre for Disease Control, USA) (716). The sequence has also been identified in the chromosome of Typhimurium strain DT193 (Ireland, 1998) and (in part) in an IncI plasmid of Enteritidis (Italy, 1997) (717). As has been suggested previously (716), it is likely that this *bla/sul/str* sequence has moved as a single unit among enteric bacteria. Two *IS*26 in direct orientation can mediate tranposition of intervening sequences, generating 8 bp direct repeats on either side of the conjugate transposon (718). However no such repeats could be identified in the sequences under study. Thus the *bla/sul/str* sequence may have originally been transferred into

*Tn*21 via recombination between its flanking *IS*26 sequences and *IS*26 elements already inserted within *Tn*21.

5.3.1.3 A composite resistance transposon

Although the transposons of pAKU_1 and pHCM1 have been disrupted by several IS element-mediated insertions and sequence rearrangements, sequence identity at the boundaries of *Tn*21 and the *bla/sul/str* insertion suggests that *Tn*9, *Tn*21 and *bla/sul/str* may have been transferred as a single unit between plasmids. Specifically, it is hypothesised that some plasmid first acquired *Tn*9, followed by the transposition of *Tn*21 into *Tn*9, 3′ of the *cat* gene (see Figure 5.3a,b). A sequence described as *Tn*2670, identified in the *Shigella* IncFII plasmid NR1 (later called R100) (709), contains *Tn*21 inserted at the same locus in *Tn*9 and is therefore a likely precursor. At some point *bla/sul/str* was inserted into the integron in *Tn*21, adjacent to *tniA*Δ (Figure 5.3c,d). The resulting 24 kbp composite transposon has since been transferred between plasmids, at the very least between distinct IncHI1 backbones. The transposition mechanism is presumably by the *IS*1 ends of *Tn*9, as direct repeats are evident at opposite ends of the *IS*1 elements in pAKU_1. The same composite transposon is evident in plasmid pRSB107, sequenced from an unknown bacterial host from a waste water-treatment plant, albeit with additional resistance gene insertions (Figure 5.4c).

Once inserted into the ancestors of pAKU_1 and pHCM1, the composite transposon sequence has been disrupted by rearrangements mediated by IS elements (see Figure 5.4). In pAKU_1, two inversions in the 5′ end of the composite transposon appear to have been mediated by *IS*26 elements integrated into the plasmid in direct orientation (green regions in Figure 5.4a). A large inversion appears to have occurred between *IS*26 elements inserted in the backbone and into the *Tn*21 *tnpR* gene (marked **IS*26 in Figure 5.4), separating the 5′ ends of *Tn*9 (*IS*1, *cat*) and *Tn*21 (*tnpA*, *tnpR*–3′ fragment) from the rest of the composite transposon. This is supported by the present arrangement of 8 bp target site duplications adjacent to *IS*26 elements (green and orange arrows in Figure 5.4). A smaller inversion appears to have occurred between the *IS*26 inserted in *tnpR* and the 5′ *IS*26 of *bla/sul/str*, disrupting the class I integron (see Figure 5.4a). This is supported by analysis of the configuration of *IS*26 target site duplications, which were inverted along with the rest of the sequence between *IS*26 elements (green and orange arrows in Figure 5.4). These inversions have presumably deactivated the composite transposon in this plasmid, as the *IS*1 genes are now in opposite orientation and separated by 62 kb, thus disrupting *Tn*9. *Tn*21 and the integron are similarly disrupted, although *bla/sul/str* may still be capable of transfer via *IS*26-mediated transposition or homologous recombination.

In pHCM1, recombination between an *IS*26 element inserted between *tnpA* and *tnpR* and the 5′ *IS*26 element of *bla/sul/str* resulted in deletion of *tnpR*, *tnpM*, *intI1* and the integron gene cassette (Figure 5.4d). *IS*4321 elements (purple in Figure 5.4d) were also inserted within the *Tn*21 inverted flanking repeats, demonstrated to be a preferred target site for this IS element (719). In pRSB107, there were two additional resistance gene insertions within the composite transposon (Figure 5.4c). The *Tn*4352B kanamycin/neomycin-resistance transposon was inserted at the 3′ end of *bla/sul/str*. This transposon comprised the *aph* gene (aminoglycosid 3'-phosphotransferase, conferring kanamycin resistance (720)) flanked by *IS*26 elements, so may have been inserted at this position via recombination with the 5′−end *IS*26 element of *bla/sul/str*. A macrolide resistance module was also inserted between the integron gene cassette and *bla/sul/str*.

5.3.1.4 Other insertions in pAKU_1

Plasmid pAKU_1 contained an additional transposon not present in pHCM1, designated *Tn*6062, inserted within *tra2* (D in Figure 5.2a). The transposon is made up of four genes, including *betU* and a conserved hypothetical protein (SPAP0105) flanked by *IS*1 elements in direct orientation. A 9 bp target site duplication was evident on either side of *Tn*6062 (as with *Tn*9, which is also composed of genes flanked by *IS*1 elements), confirming that the four genes were inserted as a single unit. *BetU* contains a betaine-choline-carnitine transporter family domain and encodes a betaine uptake system, capable of transporting glycine betaine and proline betaine (721). It was first described in *E. coli* strains causing polynephritis (ascending urinary tract infection) and is believed to be an osmoregulator, allowing *E. coli* to survive the high osmolality and urea content in urine (721). However the gene is distributed among *E. coli* with a range of pathogenic phenotypes, so its osmoprotectant properties may be useful in other environmental contexts (722). The pAKU_1 *betU* sequence shares 99% identity with the amino acid sequences found in *E. coli*, and 99% and 98% identity respectively with protein sequences found in *Shigella boydii* and *Klebsiella pneumoniae.* The conserved gene SPAP0105 contains a signal peptide sequence and four probable transmembrane helices, suggesting it may be an outer membrane protein, however it contains no protein domains of known function. BLASTN searching of the EMBL database revealed the pair of genes, SPAP0105 and *betU*, are present adjacent to each other in the chromosomes of several *E. coli* strains and the plasmid pKPN3 of *K. pneumoniae.* However, the precise structure of *Tn*6062, with two flanking *IS*1 elements, was not detected outside of pAKU_1 (BLASTN search of EMBL database, June 2009).

Plasmid pAKU_1 also contained a second region, encoding four genes SPAP0280-83, that was not present in pHCM1. The genes include citrate transporters *citA, citB*, a hypothetical protein (SPAP0281) and lysR-family transcriptional regulator *nac*, and were inserted 3′ of *Tn*10 in pAKU_1. The entire sequence was also present in R27, as well as the chromosomes of Typhi, Paratyphi A, Paratyphi C and Choleraesuis (23% divergence between chromosomal and plasmid sequences at the nucleotide level). Early studies of IncHI1 plasmids found that most IncHI1 plasmids, but not other plasmids, were able to confer citrate utilisation (Cit+) upon transfer to otherwise Cit- *E. coli* and *Shigella* strains (723, 724). However the Typhi strains from which the Cit+ IncHI1 plasmids were isolated did not appear able to utilise citrate in culture (723), despite the presence of the *citA* and *citB* transporters and *citAB* two-component system that has subsequently been found in all sequenced Typhi genomes to date (see 2.3.4.1).

5.3.2 Evolution of IncHI1 plasmids and MDR

5.3.2.1 Phylogenetic analysis of the IncHI1 plasmid backbone

A further two finished IncHI1 plasmid sequences have recently become available, plasmid pMAK1 from *S. enterica* serovar Choleraesuis (EMBL: AB366440) and p0111 from an enterohemorrhagic *E. coli* (EHEC) strain (sequence provided by Tetsuya Hayashi, University of Miyazaki, Japan; May 2008). Both plasmids were similar to pHCM1 in their conserved backbone sequence and resistance gene insertions, and were highly similar to each other. These plasmids, as well as pHCM1 and R27, were compared to pAKU_1 using MUMmer to identify SNPs within the conserved IncHI1 backbone (5.2.2). SNPs were also identified within three IncHI1 plasmids sequenced with Solexa from H58 Typhi isolates E03-9804, ISP-03-07467 and ISP-04-06979 (2.3.3.3, see methods in 5.2.2). These plasmids were similar to pAKU_1, with 100% coverage of the pAKU_1 sequence, and were indistinguishable from each other.

A total of 345 SNPs were identified among the backbone sequences of all available IncHI1 plasmids, shown in Figure 5.5. Phylogenetic analysis of the SNP data confirmed that the plasmids from Typhi H58 isolates were much more closely related to the Paratyphi A plasmid pAKU_1 than the Typhi plasmid pHCM1 (Figure 5.6). It also confirmed that p0111 and pMAK1 from *E. coli* and Choleraesuis were closest to pHCM1 (Figure 5.6). An additional 16 SNPs were identified among plasmid sequences present in the Paratyphi A pools, as described in 3.3.3.6. These formed a tight cluster with pAKU_1, each differing from pAKU_1 and one another at <10 loci (see Figure 3.22). An MLST approach was recently developed by Minh Duy Phan at the Sanger Institute to study IncHI1 plasmids by comparing sequences from six gene fragments. The analysis of 40 plasmids (including pHCM1, pAKU_1, R27 and E03-9804) identified eight SNPs, defining eight sequence types (STs) (725). The relationships between STs were consistent with the phylogenetic relationships described here, with pAKU_1 and E03-9804 belonging to distinct STs separated by a single SNP (ST7 and ST6 respectively), and pHCM1 and R27 being much more distantly related (ST1 and ST5) (725).

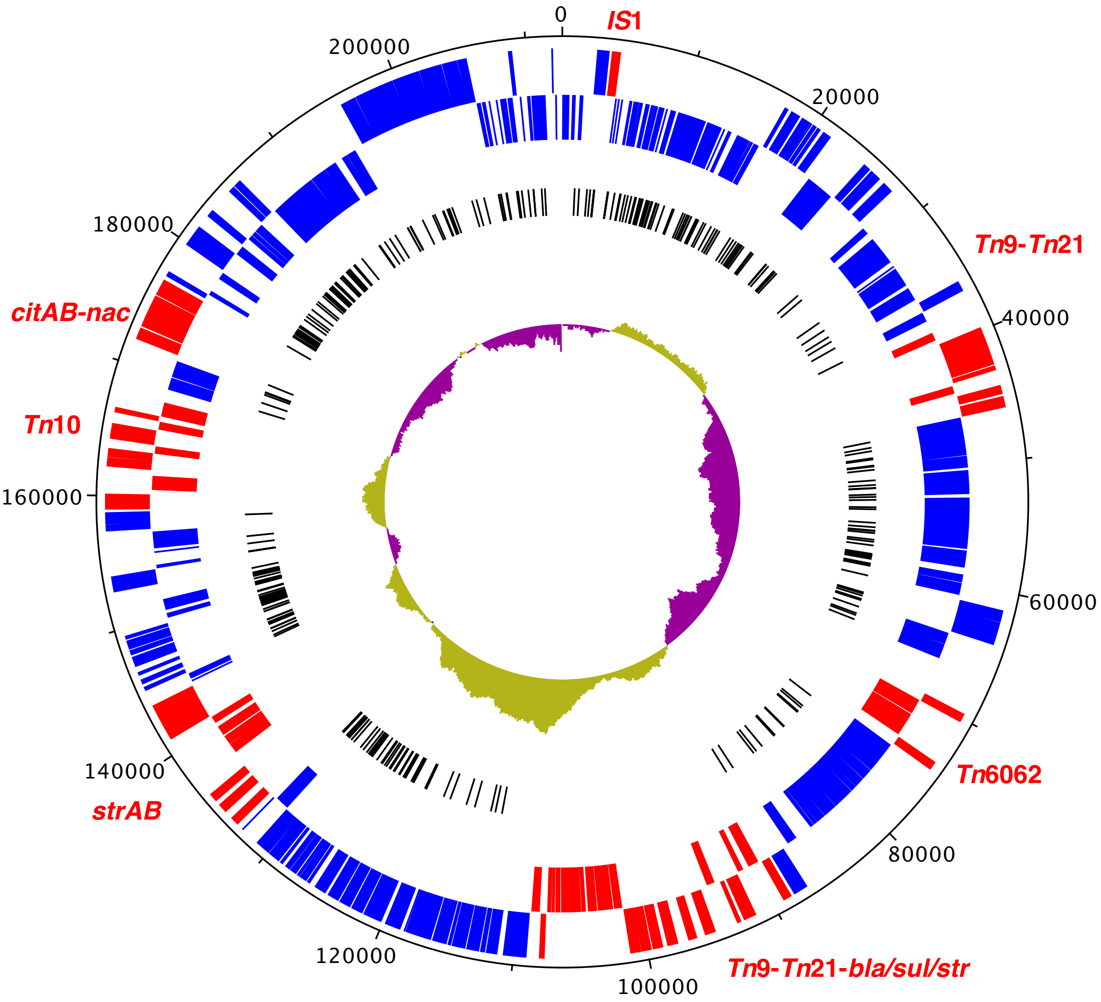


Figure 5.5: Distribution of SNPs in the pAKU_1 IncHI1 plasmid - Outer two rings = coding sequences on forward and reverse strands; blue = conserved backbone, red = insertions, labelled in red text. Third ring (black) = SNP loci included in phylogenetic analysis. Central plot shows GC deviation ((G-C)/(G+C), i.e. the difference in G content between the forward and reverse strands); inwards = negative deviation (low G), outwards = positive deviation (high G). Outer labels show pAKU_1 sequence coordinates (bp).

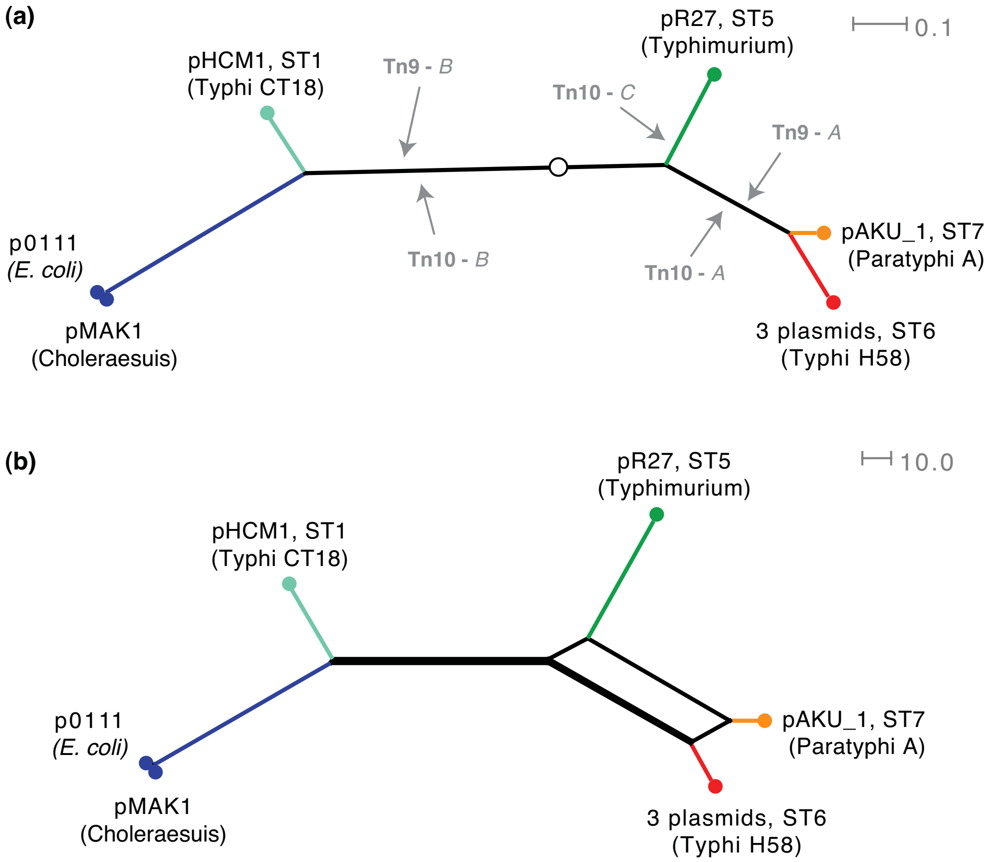


Figure 5.6: Phylogenetic trees of IncHI1 plasmids based on sequence data - (a) Maximum likelihood phylogenetic tree; each bipartition has 100% support from 1,000 bootstraps; scale bar is in substitutions per site estimated by maximum likelihood; position of the root (inferred from separate phylogenetic analysis of sequences shared between IncHI1 plasmids and IncHI2 plasmid R478) indicated with open circle. Inferred independent acquisitions of transposons are shown in grey; *Tn*9 has distinct insertion sites in two lineages, *Tn*10 has different insertion sites in three lineages, labelled A-C as in the text; positions of these sites in the IncHI1 backbone are indicated in Figure 5.2a. (b) Split network; scale bar is number of SNPs.

5.3.2.2 Drug resistance insertions in IncHI1 plasmids

The H58 Typhi plasmids contained the *Tn*9-*Tn*21-*bla/sul/str* composite transposon inserted within the *tra* conjugal transfer region, the same insertion site as in pAKU_1 (labelled *Tn*9-A in Figure 5.6a, site indicated in Figure 5.2a). They also contained *Tn*10, inserted in the same site as in pAKU_1 (*Tn*10-A in Figures 5.2a and 5.6a). This was confirmed by the presence of reads mapping across each insertion boundary in each isolate, and the successful amplification of PCR products across these boundaries (see 5.2.4 and Table 5.3). PCR results agreed with insertion sites determined from sequence data wherever both data types were available (see Table 5.3). Comparison of the finished sequences showed that plasmids p0111 and pMAK1 also contained the composite transposon, inserted within *Tn*10 at the same insertion site as in pHCM1 (*Tn*9-B in Figures 5.2a and 5.6a). The insertion site for *Tn*10 itself also matched that of pHCM1 (*Tn*10-B in Figures 5.2a and 5.6a).

A study of MDR IncHI1 plasmids collected in Vietnam between 1993 and 1996 found that pHCM1-like plasmids, common until 1996, were replaced by a novel IncHI1 plasmid with a distinct RFLP pattern (275). A plasmid representative of the novel type, pSTY7, was found by plasmid MLST to be of the ST6 type, confirming its distinction from pHCM1 (ST1) (725). PCR analysis of resistance genes in the Vietnam study showed that pSTY7 contained *sul1*, *dfrA7* and the full *tet* operon *tetRACDD*, similar to pAKU_1 but not pHCM1 (275). To check how closely related pSTY7 (ST6) was to pAKU_1 (ST7) and the Typhi H58 plasmids (ST6), PCR was performed as outlined above (5.2.4). The results indicated that pSTY7 was highly similar to the ST6 plasmids from Typhi, with the same insertion sites for all resistance genes tested (see Table 5.3). All ST6 plasmids contained the same resistance insertions as pAKU_1 (ST7), with the exception of an additional copy of *strAB* that was detected only in pAKU_1. These results are consistent with one acquisition each of *Tn*10 and the composite transposon in a common ancestor of all ST6 and ST7 plasmids, as shown in Figure 5.6a.

Insertion and data source	**ST1 HCM1**	**ST1 pMAK1**	**ST1 p0111**	**ST5 R27**	**ST6 6979**	**ST6 pSTY7**	**ST7 pAKU_1**
***Tn*10 insertion**	**B**	**B**	**B**	**C**	**A**	**A**	**A**
Sequence data	B	B	B	C	A	-	A
PCR N (*Tn*10-HCM1.247)	yes	-	-	no	no	no	no
PCR O (*tetD* /SPAP0276)	no	-	-	no	yes	yes	yes
PCR P (SPAP0261/*Tn*10)	no	-	-	no	yes	yes	yes
***Tn*9 insertion**	**B**	**B**	**B**	**no**	**A**	**A**	**A**
Sequence data	B	B	B	no	A	-	A
PCR J (*cat-trhN*)	no	-	-	no	yes	yes	yes
PCR K (*mer-trhI*)	no	-	-	no	yes	yes	yes
PCR M (*cat*-HCM1.203)	yes	-	-	no	no	no	no
PCR L (*insA*-*tetA*)	yes	-	-	no	no	no	no
***Tn*21 into *Tn*9**	**yes**	**yes**	**yes**	**no**	**yes**	**yes**	**yes**
Sequence data	yes	yes	yes	no	yes	-	yes
PCR H (*tnpA*-*Tn9*)	yes*	-	-	no	yes	yes	yes
PCR I (*merR*-*Tn*9)	yes*	-	-	no	yes	yes	yes
***bla/sul/str* into *Tn*21**	**yes**	**yes**	**yes**	**no**	**yes**	**yes**	**yes**
Sequence data	yes	yes	yes	no	yes	-	yes
PCR G (*strB*-*tniA*Δ)	yes	yes	yes	no	yes	-	yes
***strAB* 2nd copy**	**no**	**no**	**no**	**no**	**no**	**no**	**yes**
Sequence data	no	no	no	no	no	-	yes
PCR Q (*strB*-SPAP0228)	no	-	-	no	no	no	yes

Table 5.3: Resistance gene insertions in IncHI1 plasmids determined from sequence data and PCR - PCR products are labelled as in Table 5.2 (which gives primer sequences) and Figures 5.2a and 5.4 (which illustrate the positions of target amplicons); yes=successful amplification, no=no amplification with these primers; *=amplicon larger than that from pAKU_1. Insertion sites for *Tn*9 and *Tn*10 are labelled as in Figure 5.2a.

5.4 Discussion

5.4.1 IncHI1 plasmids in Paratyphi A and Typhi

The IncHI1 plasmid pAKU_1 was the first MDR plasmid from Paratyphi A to be sequenced and analysed in detail. Plasmids of this type were responsible for MDR in the majority of clinical isolates analysed from Pakistan in 2004 (283) and plasmids of a similar size have also been associated with Paratyphi A in Bangladesh and China (284, 285). Like IncHI1 plasmids isolated from MDR Typhi, the plasmid DNA sequence was composed of an IncHI1 backbone with numerous insertions of mobile elements, encoding resistance to chloramphenicol, streptomycin, beta-lactams, trimethoprim, sulfonamides and tetracycline. The comparative analysis presented here showed that pAKU_1 shares an IncHI1 backbone with plasmids that have been sequenced from other organisms: R27 from Typhimurium, pHCM1 from Typhi CT18, three plasmids from H58 Typhi, pMAK1 from Choleraesuis and p0111 from enterohemorrhagic *E. coli.* This shared backbone must have been inherited vertically from a common ancestral plasmid and was therefore analysed separately from the mobile elements contained in the plasmid sequences, which encode drug resistance genes and can be readily transferred horizontally into distinct DNA backbones.

Phylogenetic analysis of the IncHI1 backbone sequences confirmed the presence of distinct lineages of IncHI1 MDR plasmids within distinct lineages of Typhi. Specifically, pHCM1 (plasmid type ST1) was found within Typhi CT18 (Typhi haplotype H1), while plasmids of the ST6 type were found within Typhi isolates of the H58 haplotype (Figure 5.6a). The presence of distinct plasmid types in Typhi was detected previously by MLST (725), but this study links plasmid type to Typhi strain type for the first time, providing direct evidence for independent acquisitions of distinct MDR IncHI1 plasmids by distinct Typhi lineages. The spread of MDR Typhi may therefore be attributed not just to the spread of a particular plasmid within the Typhi population, or the expansion of a particular Typhi clone following the acquisition of resistance, but to the spread of multiple MDR plasmids within the Typhi population. A corollary of this is that distinct plasmid types may compete for maintenance in the Typhi population, driving selection for plasmid maintenance unrelated to drug resistance, and thereby contributing to competition between Typhi lineages. For example, the replacement of

ST1 plasmids (pSTY1-3,5) with ST6 plasmids (pSTY6-7) observed in Vietnam in the mid-1990s (275) may be related to the success of the H58 Typhi lineage (see 2.4.2.1, (2, 570)), as pSTY1 was present in H1 Typhi (CT18) and ST6 plasmids have been confirmed only in H58 Typhi isolates (see 5.3.2.1, Figure 5.6a). Unfortunately the haplotypes of the Typhi strains hosting pSTY6-7 cannot be determined as the strains themselves have not been maintained (the plasmid was transferred to *E. coli* for study in the laboratory), so this question can never be settled directly.

Phylogenetic analysis of the IncHI1 backbone sequences showed that the Paratyphi A plasmid pAKU_1 was distinct from plasmids analysed so far from Typhi and other organisms, although closely related to the plasmids found in H58 Typhi. So far there is no evidence of multiple independent acquisitions of MDR IncHI1 plasmids by Paratyphi A. Several Paratyphi A plasmids have been analysed by MLST and found to cluster into two closely related sequence types (ST7 and ST8), distinguished by a single deletion (725). In Chapter 3, just 16 SNPs were identified between pAKU_1 and sequencing reads from Paratyphi A pools. This small degree of variation could be due to mutations arising within the plasmid following a single acquisition event, or even sequencing errors in the case of SNP analysis. The question of multiple acquisitions may be resolved in the future by typing of individual plasmids and their host strains, which was not possible in this study since isolates were sequenced in pools (see 3.3.3.6).

5.4.2 Acquisiton of MDR by IncHI1 plasmids

Plasmids pAKU_1 and pHCM1 share very similar resistance gene complements, while R27 has only one resistance gene element (*Tn*10, encoding tetracycline resistance). However comparative sequence analysis found that while pAKU_1 and pHCM1 shared near-identical mobile elements encoding drug resistance, they were inserted into different loci within the IncHI1 backbones, indicating that they were acquired via independent insertion events. The accumulation of resistance in IncHI1 since the 1960s when R27 was first isolated, is therefore the result of independent acquisition of resistance genes by distinct IncHI1 plasmid lineages, rather than accumulation of resistance genes in the R27 lineage as first supposed (46). For example, R27, pAKU_1 and pHCM1 each encode a tetracycline resistance transposon *Tn*10, but the insertion site is at different positions in the IncHI1 backbone (*Tn*10-A,B,C in Figure 5.2a), suggesting that it has been independently acquired by each plasmid since their divergence. The insertion sites in pMAK1 and p0111 match the insertion site in pHCM1, consistent with a single acquisition of *Tn*10 in a common ancestor of these three plasmids (*Tn*10-B in Figure 5.2a). Similarly, the insertion sites in the Typhi H58 plasmids match that of pAKU_1, consistent with a single acquisition of *Tn*10 in a common ancestor of these plasmids (*Tn*10-A in Figure 5.2a).

In addition to *Tn*10, pHCM1 and pAKU_1 carry a highly similar set of mobile elements encoding resistance to chloramphenicol, streptomycin, beta-lactams, trimethoprim and sulfonamides. However this is not due to common ancestry of the plasmids, but to the independent acquisition of a single composite transposon by both plasmids (Figures 5.3, 5.4), which has since been subject to different rearrangements in each (Figure 5.4). The proposed composite transposon includes *Tn*9, *Tn*21 and a stretch of sequence including the bla_{TEM-1}, *sul2* and *strAB* resistance genes that may itself be mobile (*bla/sul/str*, Figure 5.3d). The insertion site of the composite transposon is different in pAKU_1 and pHCM1, supporting the hypothesis that the plasmids acquired their similar resistance genes independently by horizontal transfer rather than vertical inheritance from a common ancestral plasmid. BLAST searching the proposed composite transposon sequence (Figure 5.4b) against the EMBL database revealed its presence, without rearrangements, in a plasmid from an unknown source, pRSB107 (Figure 5.4c). This plasmid has a distinct IncF backbone, thus the composite transposon appears capable of insertion into a variety of genetic contexts. The strongest evidence for the transfer of the composite transposon as a single unit is the 100% sequence identity in pAKU_1, pHCM1 and pRSB107 across the boundaries of insertion of (a) *Tn*21 into *Tn*9, and (b) *bla/sul/str* into *Tn*21. If *Tn*9, *Tn*21 and *bla/sul/str* were acquired independently in each plasmid, it is unlikely that the insertion sites of *Tn*21 and *bla/sul/str* would be identical at the nucleotide level as they are in these three sequences. It is possible that multiple independent acquisitions of *bla/sul/str* via homologous recombination could result in identical sequences, however this would require identical recombinations between *IS*26 elements to occur at least three times. Thus the most parsimonious explanation is that the insertions occurred once to form a composite transposon, which was then able to move between distinct plasmid backbones as a single unit using the *IS*1 ends of *Tn*9.

As with *Tn*10, the insertion sites of the composite transposon in pMAK1 and p0111 match pHCM1, while those in the Typhi H58 plasmids and pSTY7 match pAKU_1 (Table 5.3). This suggests that the plasmids sequenced so far belong to just three lineages, represented by pHCM1, pAKU_1 and R27 (see Figure 5.6). This is compatible with the three major 'groups' proposed by MLST analysis of a global collection of IncHI1 plasmids: group 1 (including pHCM1), group 2 (including pAKU_1) and group 3 (R27) (725). The sequence analysis presented here suggests that group 1 and group 2 lineages each acquired, independently, *Tn*10 and the composite transposon before diversifying into the subtypes that are evident today; group 1 diversity including ST1 (pHCM1, pMAK1, p0111), ST2, ST3, ST4 and group 2 diversity including ST6 (pSTY6-7, Typhi H58 plasmids), ST7 (pAKU_1) and ST8 (725). Since the earliest reports of tetracycline resistant pathogens (*Tn*10) date back only 50 years (726) and multidrug resistant pathogens (*Tn*9, *Tn*21, *bla/sul/str*) less than 40 years (268, 684, 727), it follows that this diversification must represent recent evolution.

5.4.3 The spread of MDR via IncHI1 plasmids

The close relationships between IncHI1 backbones and resistance insertions presented above demonstrates that p0111, pMAK1, pHCM1, pAKU_1 and R27 share a recent common ancestry, indicative of spread between bacterial populations. The IncHI1 plasmid is self-transmissible, due to a conjugal transfer system which enables the plasmid to construct a pilus and transfer directly between bacterial cells (692). Each of the IncHI1 plasmids in this study (which includes all of those sequenced to date) was discovered in a different human enteric pathogen - enterohemorrhagic *E. coli*, and *S. enterica* serovars Choleraesuis, Typhi, Paratyphi A and Typhimurium, suggesting that the plasmids are able to spread between distinct pathogens occupying the human enteric niche. However it is unlikely that conjugal transfer occurs during coinfection of humans or any other mammalian host, since transfer is temperature-sensitive in IncHI1 plasmids, occurring at much higher efficiency at ambient temperatures (14-27°C) than *in vivo* temperature (37°) (689, 692). This is consistent with selection for plasmid exchange outside the animal host, perhaps in water which is a common transmission route for enteric and other bacteria.

Whatever conditions are required for conjugal transfer to take place, it is clear that MDR can be transferred between enteric pathogens via IncHI1 plasmids. There is also evidence that MDR can be transferred via phage (728) and integrated into the chromosome of host bacteria (e.g. *Tn*9-*Tn*21 composite transposon in the Typhimurium DT193 chromosome (717)). This mobility suggests that selection for resistance in one pathogen (via antibiotic treatment) can impact the development of resistance in another. It also suggests that selection for resistance to one antibiotic may lead to the proliferation of resistance to many, as complex composite transposons encoding resistance to multiple drugs can move together as a single unit. This is an important consideration in a clinical environment as it highlights that treatment choices for infection with one pathogen can impact the development of resistance in other pathogens, leading to a narrowing of options for the treatment of other, perhaps more serious infections. The phenomenon of MDR has already driven a switch to fluoroquinolone-based drugs for the treatment of many bacterial diseases including enteric fever (3, 297, 685), see Figure 1.7. Resistance to fluoroquinolones is on the rise in Typhi (16, 401), Paratyphi A (279) and other pathogens (729), however the most common mechanism of resistance to fluoroquinolones is mutations in the topisomerase targets *gyrA, gyrB, parE* and *parC* (396, 397, 398), which are encoded on the chromosome and therefore unlikely to spread between bacteria via plasmids or bacteriophage. However genes encoding resistance to fluoroqinolones (*qnr* genes) have been discovered in MDR plasmids present in many pathogens (730, 731, 732) including Typhimurium (407, 408, 409) and other *S. enterica* serovars (410, 411, 412, 413, 414, 415), including plasmids of different incompatibility groups (409). Thus it may simply be a matter of time before plasmids encoding MDR and fluoroquinolone resistance appear in Typhi and Paratyphi A, leaving very limited options for the treatment of enteric fever.

Chapter 6

Investigating Typhi populations using high throughput SNP typing

6.1 Introduction

Molecular typing of Typhi isolates is important in a variety of research and surveillance contexts. The aim is generally not just to distinguish genetically distinct groups of isolates, but to determine the phylogenetic relationships between those groups. It is therefore important that the genetic differences (mutations) uncovered by molecular typing are phylogenetically conserved, that is inherited directly during bacterial replication and not horizontally transferred or likely to revert to wildtype (e.g. temperate phage, genomic rearrangements). They should also be interpretable in a phylogenetic context, and reproducible in different sample populations and in different laboratories. Techniques currently in common use for typing of Typhi isolates (PFGE, ribotyping, phage typing) do not meet these criteria (see 1.3.1.2). Sequence-based analysis, however, provides the ultimate typing tool. Multi-locus sequence typing (MLST) has been introduced with great success in a wide variety of bacteria (458, 460). However Typhi is a relatively young serovar with very little sequence variation and current MLST schemes have virtually no power to discriminate within the Typhi population (1, 464).

Having compared whole-genome sequences of 19 Typhi isolates, we now have a set of nearly 2,000 SNPs with which to type large populations of Typhi isolates. By definition, typing these SNPs in Typhi populations will not lead to the identification of novel SNP loci, as MLST can within populations of sufficient nucleotide variation. However the comparative genome analysis revealed no regions of the Typhi genome that were sufficiently variable to construct a new Typhi-specific MLST scheme with sufficient resolution. A SNP typing scheme, based on 88 SNP loci identified in 66 Typhi gene fragments (2), has been used successfully to type Typhi isolates using Sequenom assays (256). However the scheme covers just a tiny fraction of the genome and provides limited resolution, particularly among the more common strain types, for instance H58 in South East Asia (2) or H59 in Indonesia (256). Furthermore, the Sequenom assay is difficult to scale up to large numbers of SNP loci, which will be important for large population-based studies of Typhi. To allow higher resolution SNP typing for Typhi, a genome-wide SNP typing scheme was developed using the GoldenGate platform (Illumina) (733, 734). This platform allows simultaneous typing of up to 1,536 SNP loci in 96 samples, and multiple sets of 96 samples can be assayed together with very small increases in handling time. In this chapter, the design and validation of GoldenGate arrays for SNP typing in Typhi is described, including not only chromosomal SNPs but also loci on the IncHI1 multidrug resistance plasmid, z66-encoding linear plasmid and cryptic plasmid pHCM2. This is followed by studies of a global collection of Typhi isolates, including many with MDR IncHI1 plasmids, and Typhi isolates from four endemic areas, demonstrating how high throughput SNP typing can contribute to studies of evolution and disease epidemiology in Typhi and potentially other bacterial pathogens.

Previous studies of Typhi populations have relied upon PFGE, phage typing or ribotyping to identify subpopulations of clones and study changes over time or space. For example, a study of 142 Typhi isolates from Northern, Central and Southern Vietnam from 1995-2002 using PFGE, Vi phage typing and ribotyping found that three quarters of the isolates from this period were attributable to one or two clones (the authors were unsure whether or not two distinct phage types represented two distinct clones) (735). Another study reported similar findings for Vietnamese Typhi isolates, but much more variability for isolates from Hong Kong (736). A study of isolates from four typhoid outbreaks in Southern Vietnam in 1993-1997 used PFGE and phage typing to show that each outbreak was caused by a single clone, although different outbreaks were caused by different clones (380). Several other studies of typhoid outbreaks have attributed the outbreaks to a clonal source (357, 359, 474, 737, 738, 739), although there are also studies reporting diverse clones during outbreaks (740, 741, 742, 743, 744). These studies appear to reflect two different kinds of outbreaks: those associated with infection from a single source of Typhi, often an asymptomatic carrier (359) or a contaminated water source in a non-endemic area (357, 737); and those associated with a generalised increase in exposure to Typhi bacteria through a contaminated water supply in an endemic area (740, 742, 743).

The highest incidence of typhoid fever, among both inhabitants and travellers, occurs in Asia and in particular southern Asia (see Figure 6.1) (10, 11). In these high incidence regions, typhoid affects children more than adults, with the vast majority of patients aged under 20 (10). This chapter involves analysis of Typhi isolates collected from three sites where typhoid fever is particularly common - the Mekong Delta region of Vietnam, the city of Kathmandu in Nepal and an urban slum in Kolkata (Calcutta) in the east of India (see Figure 6.1). The isolates were collected during distinct studies in each site, the details of which are given in Table 6.1. The Mekong Delta study was a two-year hospital-based treatment study involving children and adults with typhoid fever (745), led by Christiane Dolecek of the Oxford University Clinical Research Unit (OUCRU) at the Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam. The Kathmandu study was a two-year hospital-based study of the burden of vaccine-preventable bacterial disease in children under 13 (746), led by Andrew Pollard of the Department of Pediatrics, University of Oxford, UK. The Kolkata study was a population-based cohort study designed to assess the burden of typhoid fever and the efficacy of a novel Vi conjugate vaccine (343, 421, 747), led by Shanta Dutta at the National Institute for Cholera and Enteric Disease (NICED) in Kolkata, India and John Clemens of the International Vaccine Institute (IVI), Korea. Typhi isolates from a fourth site, Nairobi in Kenya, were also analysed. Kenya, like most of Africa, is also considered an endemic area for typhoid fever but has a medium level of incidence (10-100/100,000 annually (10), see Figure 6.1). DNA from 96 Typhi isolates, collected in Nairobi as part of surveillance studies over a 21 year period, was provided by Sam Kariuki of the Kenya

Medical Research Institute (KEMRI), Nairobi, Kenya.

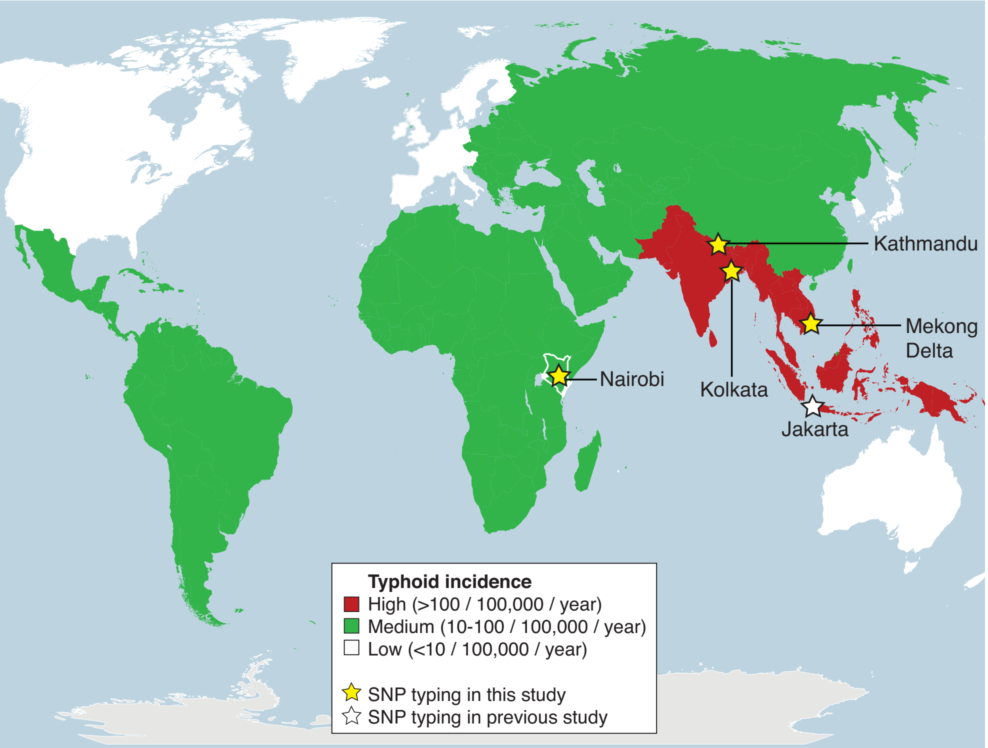

Kathmandu
Mekong Delta
Nairobi
Kolkata
Jakarta
Typhoid incidence
🟥 High (>100 / 100,000 / year)
🟩 Medium (10-100 / 100,000 / year)
□ Low (<10 / 100,000 / year)
★ SNP typing in this study
☆ SNP typing in previous study

Figure 6.1: Typhoid incidence around the world and SNP typing study sites - Incidence levels are taken from a meta-analysis of typhoid incidence data and regional extrapolatations by Crump *et al.*, published in the Bulletin of the WHO, 2004 (10). A previous SNP typing study of Typhi in Jakarta was published by Baker *et al.* in 2008 (256).

	Mekong Delta, Vietnam	**Kathmandu, Nepal**	**Kolkata, India**
Site	Rural	Urban	Urban
Wet season	Apr-Nov	Jun-Aug	Jun-Sep
Typhoid incidence	78-198/100,000 (328)		214/100,000 (327)
Study design	Hospital-based	Hospital-based	Population-based
Typhoid patients	Adults and children	Children <13 years	Adults and children
Time period	Jun 2004 - Feb 2006	Apr 2005 - Dec 2006	May 2003 - Jan 2007
Collaborator	Christiane Dolecek	Andrew Pollard	Shanta Dutta
	OUCRU	Oxford	NICED & IVI
Reference	(745)	(746)	(343, 421, 747)
Typhi isolates	358	46	188

Table 6.1: Study sites for SNP typing of localised Typhi populations - Details of studies during which Typhi isolates were collected and later made available for SNP typing with the GoldenGate assay. Note that 96 isolates collected during surveillance studies in Nairobi, Kenya between 1988-2008 were also SNP typed.

6.1.1 Aims

The aim of the work presented in this chapter was to design and validate a novel high throughput SNP typing assay for Typhi and to apply the assay to study the population structure of Typhi from both global and regional perspectives. Specific aims were to:

- design a GoldenGate assay to type Typhi chromosomal and plasmid SNPs identified in Chapters 2 and 5;
- develop analysis methods and validate the assay using data from control isolates with known alleles;
- compare the assay to previously published SNP typing data for a global collection of isolates;
- study the movement of IncHI1 MDR plasmids in the Typhi population by analysing both IncHI1 plasmid types and Typhi chromosomal haplotypes of their host strains; and
- investigate the structures of Typhi populations circulating in endemic areas, including looking for correlations between clinical or epidemiological features of typhoid and particular Typhi haplotypes.

6.2 Methods

6.2.1 DNA preparation and quantitation

DNA was provided for SNP typing analysis by the collaborators listed in Appendix E. DNA was extracted from the 19 sequenced isolates by Stephen Baker, Robert Kingsley and Derek Pickard at the Sanger Institute.

Accurate analysis of the data generated by the GoldenGate assay depends on DNA from each sample being present in equal amounts, thus DNA quantitation is an important step in the preparation of DNA for SNP typing. Quantitation was performed using the Quant-iT PicoGreen dsDNA reagent (Invitrogen) which fluoresces when bound to double stranded DNA (dsDNA) but not when bound to nucleotides, single stranded DNA, RNA or contaminants. Quantitation was performed in black flat-bottomed 96-well plates (the PicoGreen reagent is light sensitive). To the first two columns were added 90μL of buffer (provided in the Quant-iT kit) and 10μL of DNA standards at concentrations of 0, 5, 10, 20, 40, 60, 80 or 100 ng/μL (Invitrogen). One μL of sample DNA and 99μL buffer was added to each of the remaining 80 wells. A 1:100 mix of Quant-iT PicoGreen dsDNA reagent was mixed with buffer and 100μL added to each well, to reach a total volume of 200μL per well. Plates were covered for three minutes, then the fluorescence read using a FluoSTAR (Omega) fluorescence microplate reader and standard fluorescein wavelengths (excitation 480 nm, emission 520 nm). A standard curve was calculated from fluorescence of the DNA standards and used to determine the concentration of each sample (using Omega STAR software). DNA samples were arrayed in 96-well plates for genotyping, with each plate including one blank well (water) and duplicate wells of at least one sequenced isolate as a control. DNA concentrations were adjusted so that each well containined 30 μL of DNA solution at a concentration of 10 ng/μL. DNA quantitation and preparation was performed by myself (sequenced isolates, Pasteur isolates, Kathmandu isolates), Christiane Dolecek (Mekong Delta isolates), Minh Duy Phan (IncHI1 plasmid control isolates) and Derek Pickard (Kolkata and Kenya isolates).

6.2.2 Illumina GoldenGate assay

The GoldenGate assay (Illumina) allows simultaneous genotyping of up to 1,536 loci in a single DNA sample. For each SNP locus, three oligonucleotides are designed: a locus-specific primer joined to an addressing sequence and a universal PCR primer sequence; and two allele-specific primers, each joined to a different universal PCR primer sequence. Up to 1,536 of these oligonucleotide sets are combined to form an oligonucleotide pool. The oligonucleotide pool is hybridised to the DNA sample, followed by extension and ligation to generate a single allele-specific product for each SNP locus, which incorporates both allele-specific and locus-specific sequences. These products are amplified using universal PCR primers; those that bind the allele-specific PCR primer sequences are Cy3- and Cy5-labelled. The resulting amplicons are bound to their complement bead types via the locus-specific addressing sequences and the presence of specific alleles on each bead type is measured via detection of Cy3 or Cy5 fluorescent signal. All steps in the GoldenGate protocol were performed by members of the dedicated genotyping lab at the Sanger Institute, under the guidance of Ranganath Bangalore Venkatesh, Rhian Gwilliam and Panos Deloukas.

SNP alleles and 100 bp sequences flanking each SNP locus were submitted to Illumina for design of oligonucleotides for use in the GoldenGate assay. The design process takes into account interference between oligonucleotides targeted to different loci in the same pool and the potential for non-specific binding in the genome. In particular, two SNPs that are less than 60 bp apart can not be targeted in the same oligonucleotide pool, as it is not possible to design non-overlapping oligonucleotides on either side of both SNPs. In this study two oligonucleotide pools were used, each containing 1,536 SNPs, so SNPs less than 60 bp apart could be targeted by assigning each SNP to a different pool. However if more than two SNPs were pesent within 60 bp, they could not all be typed using just two GoldenGate arrays. Furthermore, if two SNPs occur within 10 bp of each other, the variation can interfere with the binding of allele-specific oligonucleotides and so such SNPs should not be typed using GoldenGate. During the design of Typhi oligonucleotides, 1.8% of known SNP loci could not be targeted due to these issues.

6.2.3 Genotype calling from raw data

The raw data provided by the GoldenGate assay is, for each SNP in each sample, a fluorescence signal corresponding to each allele-specific fluorescent-labelled probe. Raw data was normalised using the proprietary Illumina BeadStudio software, but the mean normalised signal intensities and signal:noise ratios vary among SNPs. Thus turning allele-specific signals into genotype calls requires each SNP to be analysed individually, across a range of samples. This process is known as genotype clustering, and is essentially a two-dimensional clustering problem, where each allele-specific probe contributes one dimension. The problem is best illustrated by two-dimensional cluster plots, like those in Figure 6.2. DNA samples in which one allele is present will cluster at a point along the axis corresponding to the fluorescent marker attached to amplicons containing that allele. For example in Figure 6.2 samples lying in the x-axis cluster contain allele A, while samples lying in the y-axis cluster contain allele B. In a haploid bacterial sample this would imply the x-axis cluster allele or haplotype is A, and the y-axis cluster allele or haplotype is B (Figure 6.2a); in a diploid human sample this would imply the x-axis cluster genotype is AA and the y-axis cluster genotype is BB (Figure 6.2b). Among diploid samples like human DNA, heterozygotes are common, whereby both alleles are present in a single sample. For example in Figure 6.2b a number of samples cluster around a third point which has positive signal for both alleles, this cluster indicates the heterozygous genotype AB. In haploid bacterial samples we would not expect to see heterozygous clusters, unless the SNP locus is present in multiple copies within the bacterial genome.

In the present study care was taken not to include such SNP loci (see 2.3.1.5), so heterozygous clusters are not expected. However several loci were included that are expected to be absent from some samples, for example plasmid SNPs. These SNPs should generate a cluster around the origin of the graph, as in Figure 6.2c-d, corresponding to a lack of signal for either allele-specific probe at the SNP locus. The same lack of signal would be observed if the allele-specific probes failed to bind for some other reason, e.g. (a) a secondary SNP or indel mutation was present within one of the primer-binding sites or (b) a third allele was present at the SNP locus which would fail to bind to either of the two allele-specific primers. Thus while these ‘no signal’ clusters

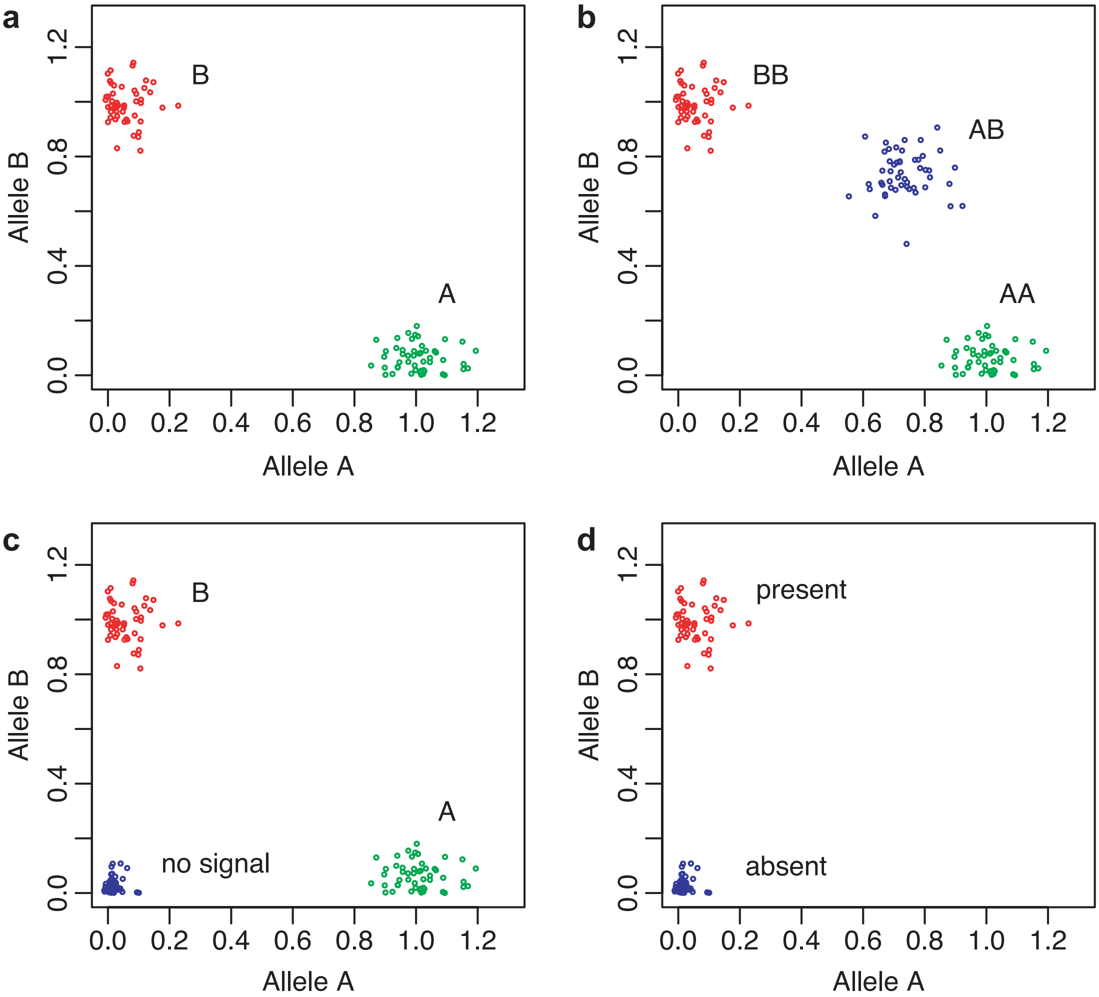


Figure 6.2: Example cluster plots for genotyping assays - Genotyping with the Illumina GoldenGate and similar array-based assays generate allele-specific signals for each SNP in each sample. By plotting these values for all samples at a given SNP, clusters can be identified which represent different genotypes at that SNP locus. (a) Haploid clusters at a SNP locus. (b) Diploid clusters at a SNP locus, with heterozygotes in blue. (c) Haploid clusters at a SNP locus with some samples displaying no signal for either allele; e.g. for plasmid SNPs, this would represent strains with either allele (A, B) or no plasmid present (no signal). (d) Haploid clusters at a nonpolymorphic locus with some samples displaying no signal; e.g. for a nonpolymorphic plasmid locus that is absent from some strains.

most likely indicate absence of known plasmid or chromosomal deletion loci, more care must be taken in the interpretation of 'no signal' clusters for chromosomal SNP loci.

6.2.3.1 Genotype calling with Illuminus-P

Since each SNP needs to be clustered individually, manual genotype clustering is extremely time consuming and can introduce bias into genotyping results. An automated clustering algorithm is implemented in the Illumina software (BeadStudio), but has been optimised for clustering genotypes in diploid samples, i.e. where three clusters are expected for each locus, corresponding to two homozygous and one heterozygous genotype (AA, BB or AB, as in Figure 6.2b). However for the present study involving haploid bacterial genotyping, we expect no heterozygous genotype clusters but some legitimate 'no signal' genotype clusters (Figure 6.2c-d). Illumina BeadStudio is proprietary software and unable to be modified, but third-party opensource genotype clustering algorithms are available. Among the best performing for the Illumina genotyping platform is Illuminus (748). On request the software's author, Yik Y Teo (Wellcome Trust Centre for Human Genetics, University of Oxford, UK), modified Illuminus to fit a third 'no signal' cluster centred at the origin rather than a heterozygous cluster. This version, referred to hereafter as Illuminus-P, was applied to genotype clustering of all GoldenGate data presented in this chapter.

6.2.3.2 Heuristic to identify 'no signal' cluster

Despite the modification, Illuminus-P occasionally failed to differentiate the cluster around the origin from the other clusters. To remedy this, a heuristic was applied following genotype clustering with Illuminus-P, such that for each SNP locus, samples with both allele signals below 15% of the maximum observed signal at that locus were assigned to the 'no signal' cluster. This heuristic improved genotype clustering accuracy for plasmid but not chromosomal loci and was applied to all plasmid loci in the analysis presented in this chapter.

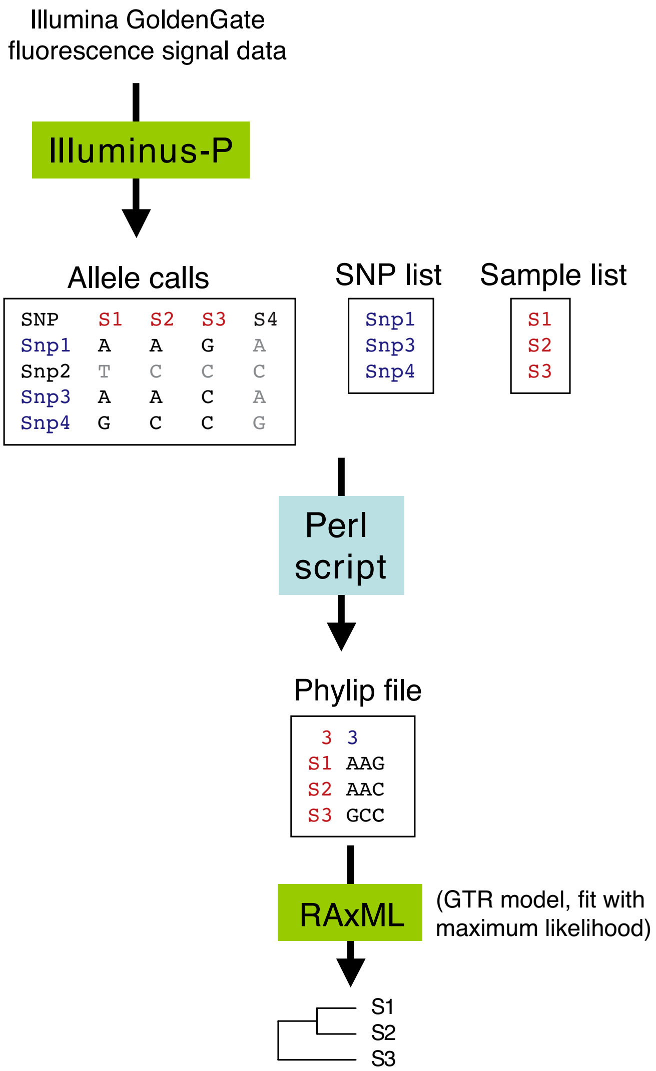
6.2.3.3 Clustering across plates

Since clustering requires each SNP to be considered individually across a range of samples, the accuracy of genotype clustering depends on the number and uniformity of the samples. Theoretically, increasing the number of data points increases the accuracy of cluster definition, so the more samples the better the performance. However clustering can be affected by the presence of outlier samples or subsets of samples. DNA samples are prepared for genotyping in 96-well plates, with all steps in the GoldenGate assay (PCR, array hybridisation and scanning) performed in 96-well format. We therefore expect to see little variation within 96-sample sets due to technical factors, but variation between 96-sample sets is sometimes observed. For example, one 96-sample set may display lower mean signal intensity for some SNPs compared to other sample sets, presumably due to run-specific technical factors such as reagent concentration or hybridisation conditions. In such cases including the plate with lower mean signal will diminish the clustering performance across all samples at loci that are most sensitive to technical variation.

For consistency in the present study, all sample sets were clustered together except where there was evidence of poor performance. Specifically, each 96-well plate was clustered individually in addition to clustering across all available sample sets. Each 96-well plate contained at least one control sample (a Typhi isolate whose genome has been sequenced and therefore all alleles were known), usually in duplicate. The accuracy of genotype calling in these control samples using individual-plate clustering was compared to that for all-plate clustering. Only four plates, all from the second oligonucleotide pool, gave better results on individual clustering. Each of these four plates displayed low mean signal intensity compared to the other plates and gave better results when clustered together as a set (compared to individual clustering or clustering with all other plates).

6.2.4 Phylogenetic analysis of genotyping data

Alleles determined by genotype clustering with Illuminus-P were analysed as follows. A Perl script was written to extract allele data from subsets of high-quality SNPs (determined by comparison with alleles for control isolates, see below 6.3.1.1) and samples from the clustered data set, see Figure 6.3. For example, chromosomal SNP loci were extracted for analysis separately from plasmid SNPs; samples from different data sets were analysed separately (e.g. global collection, individual local studies, etc). The


Illumina GoldenGate fluorescence signal data
Illuminus-P
Allele calls
SNP list
Sample list
Perl script
Phylip file
RAxML
(GTR model, fit with maximum likelihood)
S1
S2
S3

SNP	S1	S2	S3	S4
Snp1	A	A	G	A
Snp2	T	C	C	C
Snp3	A	A	C	A
Snp4	G	C	C	G

```
Snp1
Snp3
Snp4
```

```
S1
S2
S3
```

```
 3 3
S1 AAG
S2 AAC
S3 GCC
```

Figure 6.3: Phylogenetic analysis workflow for SNP typing studies - Illuminus-P was used to cluster fluorescence data across samples (including control and experimental samples from multiple studies) and call SNP alleles. A Perl script was written to extract allele data for subsets of SNPs (in this case Snp1, Snp3, Snp4) and subsets of samples (in this case S1, S2, S3) and convert it into a form suitable for phylogenetic analysis by RAxML.

script outputs alignments of concatenated strings of SNP alleles in phylip format, suitable for analysis using a range of phylogenetics software. Chromosomal alleles from a global collection of 180 isolates plus 19 sequenced isolates was analysed in ModelTest (645) (implemented in FindModel (646)), which suggested a general time reversible (GTR) model was most appropriate for analysis of this data. All phylogenetic analysis presented in this chapter was performed using maximum likelihood approaches to fit a GTR model, implemented in the software program RAxML (644). SNP typing with the GoldenGate assay only provides genetic information at the specific assayed loci; in the case of chromosomal SNPs, these are mostly loci determined by whole genome comparison of 19 Typhi strains (Chapter 2). Thus the SNP typing analysis essentially places each Typhi isolate at the appropriate position along branches defining the phylogenetic tree of these 19 strains (Figure 2.6) and branch lengths reflect genetic divergence only at the assayed loci. Short branches separating very closely related clusters (e.g. within the H58 group) were verified by manually inspecting cluster plots for SNPs that differentiated within those clusters. Similarly, plasmid SNPs were originally determined from comparison of a few plasmid sequences (Chapter 5) and thus genotyping at these loci essentially assigns each plasmid to a position in the phylogenetic tree defined by these SNPs (Figure 5.6).

6.2.5 Visualisation of temporal and spatial data

Data analysis was performed using the open-source statistical programming package R (749). For the Kathmandu data set, typhoid cases were so infrequent as to be easily visualisable using monthly counts; for the Kenya dataset years but not precise dates of isolation were available, so annual counts were used. For the Mekong Delta and Kolkata data sets typhoid cases were more frequent, so a smoothing function was used to visualise the incidence rate using the recorded day of isolation. Temporal distributions of Typhi isolates were plotted using R's `density` function for kernel density estimation, using days since study began to indicate the time of each isolation and bandwidth of 10 days.

Spatial clustering within the Kolkata data set was performed using Openshaw's Geographical Analysis Machine (GAM) (750), implemented in the R package `DCluster` (751). Note this is intended to highlight clusters of unusually high density but does not provide a formal statistical test of spatial clustering, hence additional data and statistics is provided for each cluster discussed in the text. The centre of each geographical cluster (i.e. the clusters defined by the study team for distribution of vaccine) was determined using a spatial grid in which the smallest unit corresponded to the size of the smallest cluster (cluster 44). The observed and expected number of cases for each cluster i were defined as follows:

$$\begin{aligned} observed_i &= n_i, \quad (6.1) \\ expected_i &= \frac{p_i}{P} * N, \quad (6.2) \end{aligned}$$

where n_i = number of cases observed in cluster i, N = $\sum_{i=1}^{k} n_i$, p_i = population of cluster i, P = $\sum_{i=1}^{k} p_i$. The `opgam` method of `DCluster` (751) was used to assess whether or not each point in the spatial grid represented a cluster of cases, based on a random Poisson distribution of cases among the study population; statistically significant clusters (p<0.005) were plotted.

6.2.6 Simpson's diversity index

Simpson's diversity index, 1-D was calculated as follows:

$$1 - D = 1 - \sum_{i=1}^{k} \frac{n_i * (n_i - 1)}{N * (N - 1)}, \quad (6.3)$$

where n_i number of Typhi isolates of haplotype i, N = total number of Typhi isolates.

Diversity indices were compared for two sets of Typhi isolates using Student's T-test, as follows:

$$\begin{aligned} s_x^2 &= \frac{4}{N} * (\sum_{i=1}^{k_x} p_{i,x}^3 - \sum_{i=1}^{k_x} p_{i,x}^2)^2, \quad (6.4) \\ T &= \frac{D_2 - D_1}{\sqrt{s_1^2 + s_2^2}}, \quad (6.5) \\ df &= k_1 + k_2 - 2, \quad (6.6) \end{aligned}$$

where $p_{i,x} = \frac{n_i}{N}$ for sample x, k_x = number of taxa in sample x.

For the Kolkata Typhi data, s^2 prior to December 2004 was 0.0005, s^2 from January 2005 was 0.0015, thus T = 82.85 with 14+10-2 = 22 degrees of freedom (df), which has a p-value of <1x10^{-6}.

6.3 Results

6.3.1 Validation of GoldenGate assay for target loci in Typhi

The accuracy of the GoldenGate assay and genotype clustering for the two Typhi oligonucleotide pools was assessed using data from the first ten 96-well plates run on the arrays. These samples include all sequenced isolates, 180 isolates previously genotyped at 88 loci (isolates from the Pasteur Institute, see Appendix E (2)) and several hundred isolates from diverse regions across Asia and some from Africa (including Mekong Delta and Kathmandu collections, part of Kenya collection). Each of the 19 sequenced isolates was assigned alleles at each of these SNP loci based on sequencing data (Chapter 2). Each plate contained at least one of the sequenced isolates, often in duplicate, and comparison of the expected alleles with those assigned following genotype clustering of GoldenGate data was used to assess the accuracy of the GoldenGate SNP typing results.

6.3.1.1 Chromosomal loci

The GoldenGate assay utilises mega-plex PCR (up to 1,536 oligonucleotide nucleotides per pool) followed by hybridisation to custom bead arrays (see 6.2.2). Due to the PCR step, it is not possible to uniquely target two SNP loci separated by less than 60 bp in a single oligonucleotide pool. It is also not possible to assay any SNP locus that lies within 10 bp of another SNP, insertion or deletion, as these variants can interfere with primer binding. For these reasons, 35 (1.8%) of the 1,964 SNPs identified in Chapter 2 were not suitable for SNP typing with GoldenGate. Oligonucleotides were designed to target the remaining 1,929 SNPs, as well as 72 additional SNP loci ("BiPs") identified by Roumagnac *et al* (2). As explained above, the GoldenGate assay generates fluorescence signals for each SNP locus in each sample, which must be converted into genotype calls using a two-dimensional clustering approach (see 6.2.2). Illuminus (748), originally designed for clustering of diploid human genotypes, was used to assign alleles to the Typhi data. This resulted in perfect allele calls in control samples for just 1,104 SNPs (57%). A modified version of Illuminus, Illuminus-P (see 6.2.3.1) gave better results, with perfect allele calls in control samples for 1,436 SNPs (74%). Thus Illuminus-P was used for the remainder of this study.

Analysis of GoldenGate data for sequenced isolates was used to determine whether each SNP was (a) assayed successfully, (b) clustered accurately and (c) truly polymorphic as expected from the sequence data. Each SNP assay was considered successful if it generated signals of reasonable strength that were able to be clustered. For 1,402 SNPs, alleles assigned by Illuminus-P clustering of GoldenGate data agreed with all those expected from sequence data and these were considered high quality SNP assays for downstream analysis. For 19 SNPs ($\sim$1%), GoldenGate analysis detected no evidence of the derived allele (confirmed by manual inspection of the signal plots in addition to Illuminus-P clustering). These SNPs were originally detected in sequence data from just one isolate each and are likely to be genuinely nonpolymorphic sites. These SNPs were concentrated in isolates E01-6750 (5 SNPs), E03-9804 (3 SNPs), J185SM (3 SNPs) and M223 (3 SNPs), and were not included in downstream analysis. Note that none of these were nonsense SNPs, so would not affect analysis of pseudogenes or selection in Chapter 2. For a further 12 SNPs, GoldenGate analysis found evidence of the derived allele, but allele assignments did not match exactly those from sequence data. Signal plots for these SNPs were manually inspected and assessed to be good quality signals and accurately clustered. Therefore these loci are truly polymorphic, but the earlier sequence-based allele assignments likely contained some errors which can now be corrected by the GoldenGate analysis. These errors were concentrated in J185SM (4 SNPs), M223 (3 SNPs) and E01-6750 (3 SNPs), and resolve two apparent homoplasies (reported in Table 2.9) in genes STY0347 (*tsaC*) and STY1689 (*ydhD*).

Thus 1,448 SNP loci (75% of those targeted) were considered successful GoldenGate assays of polymorphic loci identified previously from sequence analysis. While the high rate (25%) of failed assays means a reduction in resolution, the SNP loci concerned were distributed evenly within the phylogenetic tree defined by the complete SNP set (Figure 2.6). Figure 6.4 shows the length of each branch as defined by the complete SNP set (x-axis) vs that defined by the successfully assayed SNPs (y-axis), demonstrating that these lengths are highly correlated (Pearson $R^2 = 0.994$, $p<2x10^{-16}$). Thus the loss of resolution does not lead to significant change in relative branch lengths of the resulting phylogenetic tree.

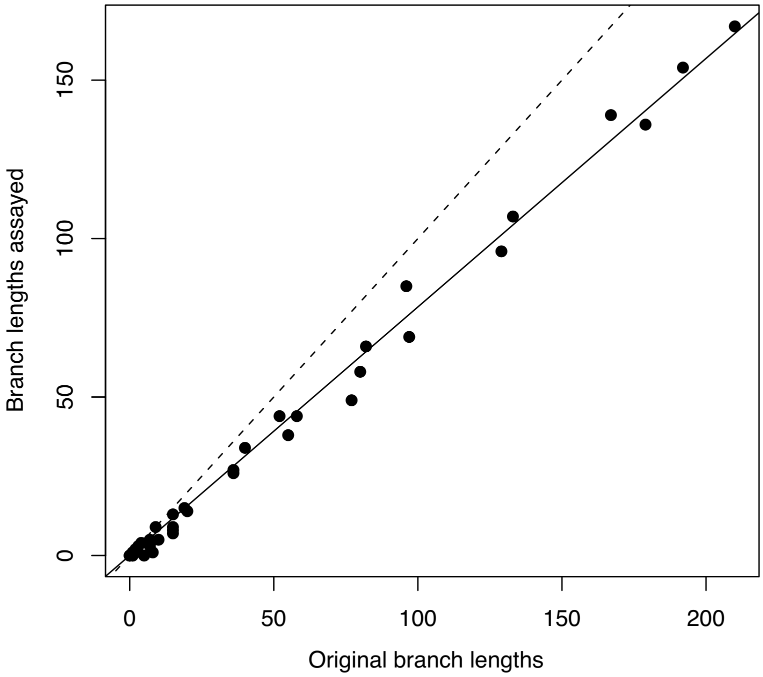


Figure 6.4: Effect of assay failure on relative branch lengths for Typhi chromosomal SNPs - Solid line indicates linear model fit (Pearson R^2 = 0.994), dashed line indicates y=x.

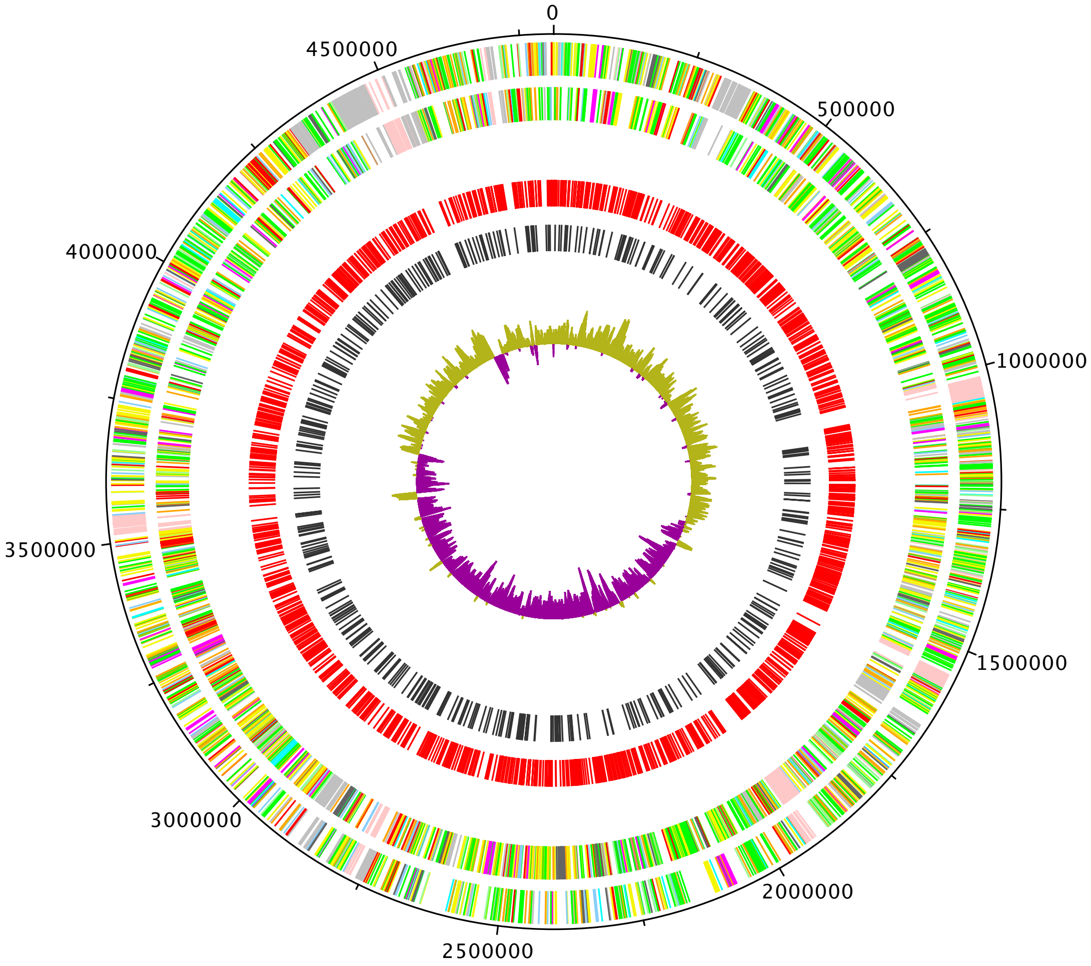


Figure 6.5: Distribution of assayed SNPs in the Typhi CT18 chromosome - Coordinates are bp in the Typhi CT18 genome. Genes annotated in the CT18 genome are shown in the two outer rings: outer-most ring = forward strand, second outer-most ring = reverse strand. Genes are coloured according to broad functional groups, note phage genes are coloured pale pink. Red and black rings show the location of all SNPs detected so far in the Typhi population (red) and those successfully assayed with GoldenGate (black). Inner-most rings indicate GC deviation in the CT18 genome ((G-C)/(G+C), i.e. the difference in G content between the forward and reverse strands).

Signal plots were manually inspected to assess the GoldenGate assays of 72 SNPs defined by Roumagnac *et al* (2), of which 60 (83%) were of high quality (signals of reasonable strength, clustered accurately). Thus the phylogenetic analysis of experimental Typhi isolates presented in this study is based on 1,508 chromosomal SNP loci. These loci are distributed randomly in the Typhi chromosome (Figure 6.5), indicating that failures in the design, signal generation or clustering do not bias the distribution of SNPs that were assayed successfully with GoldenGate.

6.3.1.2 IncHI1 plasmid loci

A total of 345 SNPs were identified in the IncHI1 plasmid backbone in Chapter 5, 294 (85.2%) of these were included in the GoldenGate assay, and their genotypes called using Illuminus-P followed by the heuristic. IncHI1 plasmid SNPs were validated by comparing alleles from SNP typing and sequence data for the plasmids previously sequenced in Typhi isolates CT18, E03-9804, ISP-03-07467 and ISP-04-06979, the Paratyphi A plasmid pAKU_1 and the Typhimurium plasmid R27, as well as plasmids used to develop the pMLST scheme (Chapter 5, (725)). Alleles from genotyping and sequencing matched perfectly for 200 IncHI1 plasmid SNPs (68.0%) and these loci were used for the remainder of the study. The distribution of these SNPs is shown in Figure 6.6. Positive allele calls for IncHI1 plasmid SNPs (i.e. fluorescence signals clustered outside the 'no signal' cluster) correlated strongly with presence of the plasmid. The four control isolates known to contain IncHI1 MDR plasmids each gave positive allele calls for 193-197 of the 200 IncHI1 plasmid SNP loci, whereas the IncHI1 plasmid-free control isolates each gave positive signals for 0-5 plasmid SNPs. To assess whether excluding the poorly-assayed SNPs would bias phylogenetic analysis of the IncHI1 plasmid, the number of SNPs lying on each branch of the phylogenetic tree was compared for (a) the complete set of 345 SNPs identified in Chapter 5 and (b) 200 SNPs with 100% genotype calling accuracy. The branch lengths represented by the two SNP sets were highly correlated (R^2=0.964) (see Figure 6.7). This demonstrates that the exclusion of low quality SNPs does not significantly alter the relative branch lengths of phylogenetic trees determined from analysis of the SNP typing data and therefore will not bias estimates of relative genetic distance between samples. Note however that since 32% of SNP loci are being excluded, genetic distance will be underestimated by phylogenetic analysis of the SNP data.

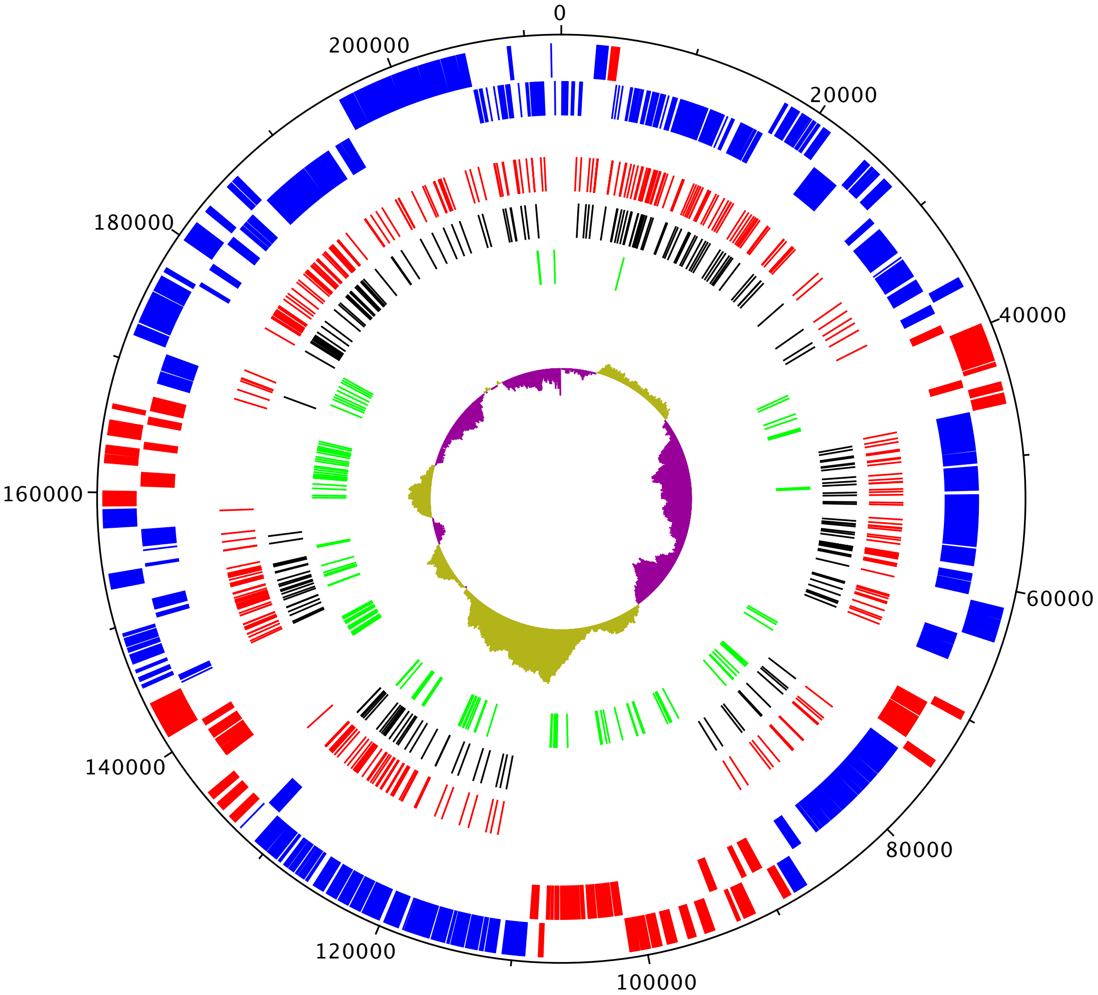


Figure 6.6: Distribution of assayed SNPs in the IncHI1 plasmid - Coordinates are bp in the IncHI1 plasmid pAKU_1. Genes annotated in pAKU_1 are shown in the two outer rings: outer-most ring = forward strand, second outer-most ring = reverse strand. Genes are coloured to indicate the IncHI1 conserved background (blue) or insertion sequences (red). Rings 3 and 4 show the location of all SNPs detected so far in the IncHI1 plasmid population (red) and those successfully assayed with GoldenGate (black). Ring 5 shows loci targeting resistance genes and deletions in the plasmid backbone (green). Innermost rings indicate GC deviation in pAKU_1 ((G-C)/(G+C), i.e. the difference in G content between the forward and reverse strands).

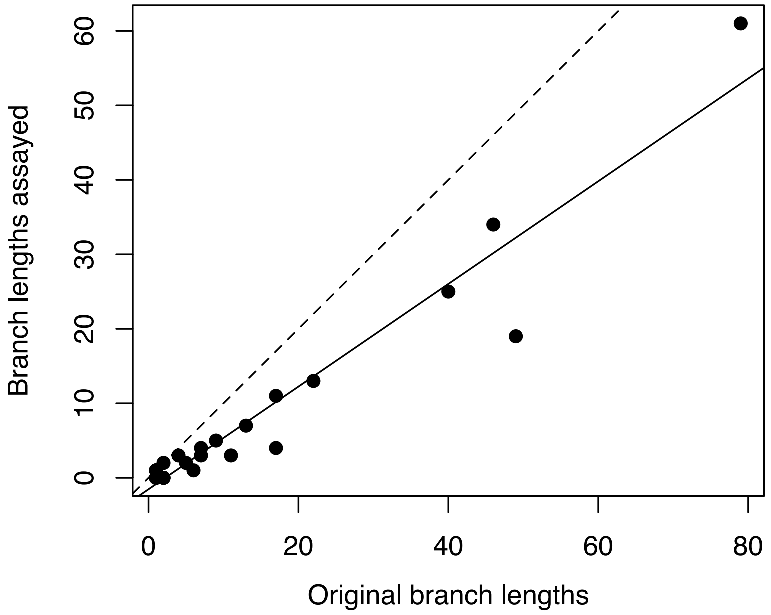


Figure 6.7: Effect of assay failure on relative branch lengths for IncHI1 plasmid SNPs - Solid line indicates linear model fit (Pearson $R^2 = 0.964$), dashed line indicates y=x.

A total of 218 SNPs designed to assess the presence or absence of resistance genes and specific IncHI1 sequences were included on the arrays, and their genotypes called using Illuminus-P and the heuristic. Here the 'no signal' cluster implies absence of the target sequence, which may be due to absence of the entire plasmid (for IncHI1-specific sequences, if no other IncHI1 targets are detected) or absence of the specific locus (if most other IncHI1 targets are detected). Note that resistance genes may be present on plasmids of a different type, or potentially integrated into the chromosome, and so are not always associated with the presence of IncHI1 sequences. Perfect matches were obtained between sequence and genotyping data for 119 of these loci (54.6%). This provides reasonable coverage of resistance genes and insertion sequences, as well as several deletions characterised earlier by comparative analysis of the three finished plasmid sequences pHCM1, pAKU_1 and R27 (Chapter 5), shown in Table 6.2.

Gene(s)	Sequence type	Number of targets
*cat**	*Tn*9 - chloramphenicol res	3
*tetACDR**	*Tn*10 - tetracycline res	3, 3, 2, 2
*merAPRT**	*Tn*20 - mercury res	2, 1, 2, 2
*dfhR7**	integron cassette - trimethoprim res	3
*dhfR14**	integron cassette - trimethoprim res	3
*sul1**	integron cassette - sulfonamide res	1
bla_{TEM-1}*	bla/sul/str - beta-lactam res	2
*sul2**	bla/sul/str - sulfonamide res	1
*strAB**	bla/sul/str - streptomycin res	2, 2
IntI1	integrase of class I integron	1
tnpAB	*Tn*21 transposase	2, 2
*IS*10-L,-R	*Tn*10 transposase	2, 2
*IS*26	transposase	2
*IS*30	transposase	2
*IS*6100	transposase	2
repC	replication initiation gene	1
betU	*Tn*6062 - betaine transporter	1
SPAP0105	*Tn*6062 - hypothetical protein	2
citAB	citrate transporters	3, 2
nac	transcriptional regulator	1
SPAP0281	hypothetical protein	3
R0140	backbone in/del	1
SPAP0266-9	backbone in/del	2, 3, 3, 2
SPAP0329	backbone in/del	2
HCM25-6	backbone in/del	1
HCM102-106	backbone in/del	3, 2, 3, 0, 1
HCM160	backbone in/del	2
HCM163	backbone in/del	2
HCM170	backbone in/del	2
HCM174	backbone in/del	3
HCM177-8	backbone in/del	1, 2
HCM189-193	backbone in/del	3, 3, 2, 1, 3
HCM195-199	backbone in/del	2, 3, 1, 0 1
HCM203	backbone in/del	2
HCM278	backbone in/del	1

Table 6.2: SNPs for detection of resistance genes and IncHI1 plasmid deletions - *Antimicrobial resistance genes, res resistance, *Tn* transposon, *IS* insertion sequence.

6.3.1.3 Other target loci

Two loci provided accurate detection of the cryptic plasmid pHCM2, which was successfully detected in the two sequenced isolates CT18 and E02-2759. A further two SNPs provided accurate detection of plasmid pBSSB1 which carries the z66 flagella antigen, present in the two sequenced H59 isolates 404ty and E03-4983. Replication (*rep*) genes from another 17 plasmids, detailed in (752), were also included on the array. Control plasmids were not available to test these assays, but allele signals were only detected in two isolates, both multidrug resistant Typhi containing plasmids of different incompatibility groups (see below 6.3.2.2).

Two SNPs specific to Paratyphi A were included in the assay to facilitate detection of erroneously serotyped isolates. The SNPs were validated by typing five Paratyphi A control isolates, which gave distinct allele signals from the Typhi control isolates at the two Paratyphi A-specific loci. Alleles were determined for 89% of Typhi chromosomal SNPs in the Paratyphi A strains, resulting in these strains clustering at the root of the Typhi phylogenetic tree (see Figure 6.8b).

6.3.2 Validation of GoldenGate SNP typing in a global collection of Typhi isolates, previously typed at 88 loci

DNA from a global collection of 180 Typhi isolates was provided by Francois-Xavier Weill of the Pasteur Insitute, Paris, France. The isolates, listed in Appendix E, were collected between 1958 and 2004 from travellers returning to France with typhoid fever. Their geographical origin is considered to be the country in which the patient initially became ill, and includes Africa (89 isolates), Asia (77 isolates) and South America (10 isolates). This collection was previously used to identify SNPs in fragments of the Typhi genome (2), generating the tree reproduced in Figure 6.8a.

6.3.2.1 Phylogenetic analysis of chromosomal SNPs

As expected, phylogenetic analysis of chromosomal SNPs in this collection placed each isolate at different points on the phylogenetic tree defined by the 19 sequenced isolates (Figure 6.8b). The location of each isolate according to GoldenGate SNP typing was consistent with their previously defined haplotypes based on 88 SNPs (2) (Figure 6.8a),

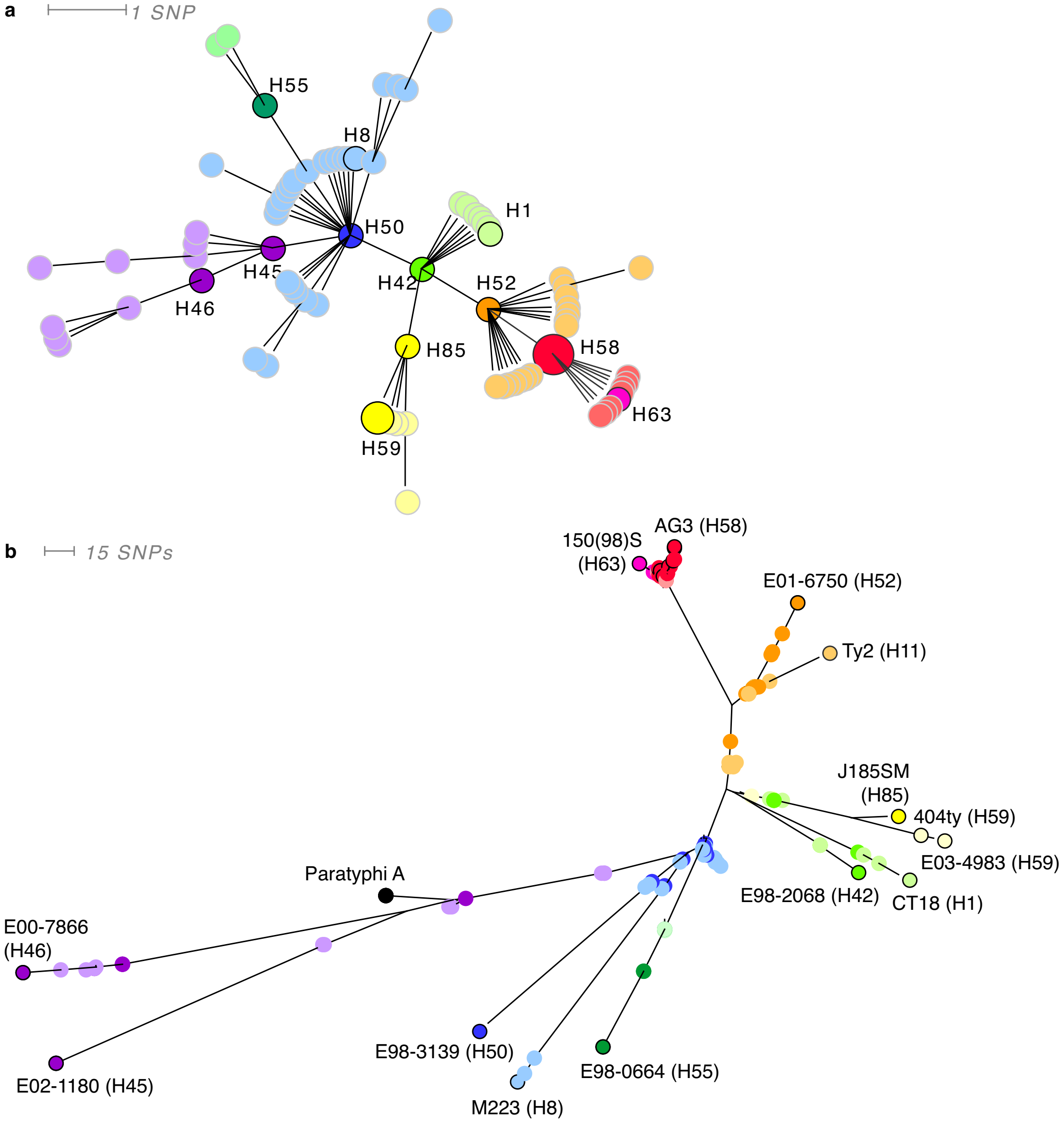


Figure 6.8: Phylogenetic trees for a global collection of 180 Typhi isolates (1958-2005) - (a) Minimum spanning tree based on 88 SNPs, reproduced from (2). Each circle (node) represents a distinct haplotype group. Haplotype groups from which isolates were sequenced are shown in bold colours with black outline, haplotypes not represented in sequencing are show in paler colours. Non-sequenced haplotypes are considered the result of clonal expansion of the sequenced haplotypes, and are coloured in a paler shade of the same colour as the haplotype from which they are assumed to have descended. (b) Maximum likelihood tree based on 1,508 chromosomal SNPs typed with the GoldenGate assay. Sequenced strains have black outlines; black circle is Paratyphi A strains, indicating the position of the root of the Typhi tree. All genotyped isolates lay along branches defined by the phylogenetic tree of sequenced strains. Colours indicate the previously determined haplotype of each isolate as given in (a); the same colour scheme is used in Figure 6.9.

demonstrating the ability of the Typhi GoldenGate assay to correctly cluster Typhi isolates into known haplotype groups.

In addition, the assay was able to differentiate within some haplotypes, even when only one member of the haplotype group had been included in SNP detection via sequencing. For example, E98-0664 was sequenced as a representative isolate of the H55 haplotype. As Figure 6.9 shows, the SNPs identified in this isolate were able to differentiate the four SNP typed H55 strains (77-303, 77-302, 75-2507 and E98-0664, coloured dark green) into three clusters lying along the branch that ends at E98-0664. Similarly, SNPs identified in M223 (H8) differentiate all three SNP typed H8 isolates (M223, E00-4626, E01-5612, coloured pale blue) and cluster H77 isolates into two groups (1458, 81424, 81918, pale blue). Differentiation between H50 isolates is also evident in Figure 6.9.

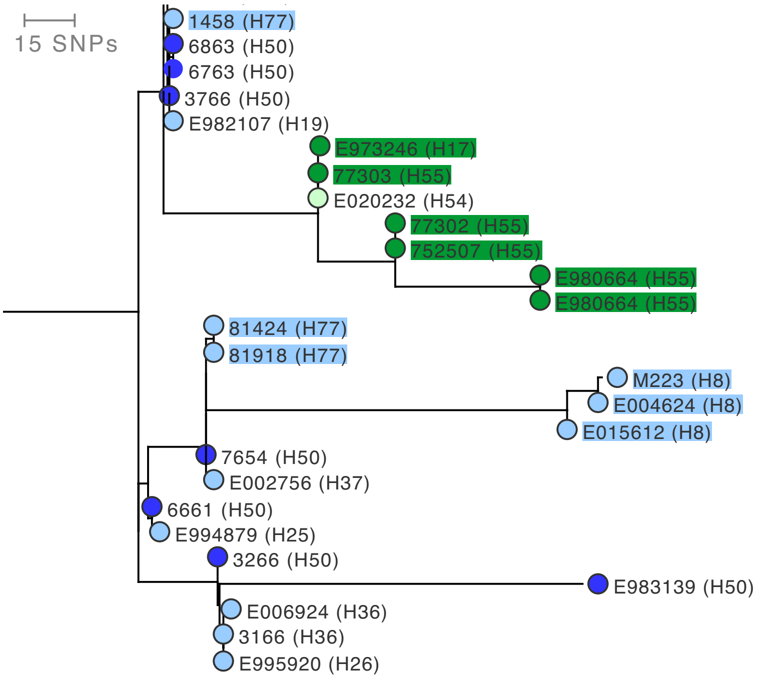


Figure 6.9: Discrimination within known Typhi haplotypes - Part of the phylogenetic tree based on chromosomal SNPs in the global collection of Typhi isolates. Isolates are coloured according to haplotype groups to which they were previously assigned (2) as shown in Figure 6.8a.

6.3.2.2 IncHI1 plasmids and multidrug resistance

A total of 40 isolates were recorded in the Pasteur laboratory as being multidrug resistant (defined as resistant to ampicillin, chloramphenicol and co-trimoxazole). The distribution of the number of IncHI1 plasmid SNPs in these and drug sensitive isolates is given in Figure 6.10. There was a clear distinction between isolates that gave positive signals for IncHI1 loci (>180 out of 200 SNPs) and those that did not (≤20 out of 200 SNPs, see Figure 6.10).

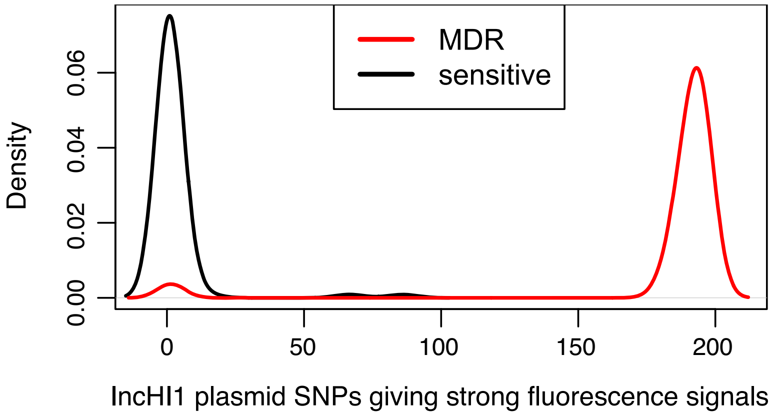


Figure 6.10: Distribution of IncHI1 plasmid SNPs detected in MDR and drug sensitive isolates - 'Strong fluorescence signal' is defined as clustering within an allele cluster as opposed to the 'no signal' cluster for a given SNP, this implies a fluorescence signal >15% that of the maximum signal detected for either allele of that SNP across all Typhi samples.

Thirty-eight of the isolates recorded as MDR had allele signals for >180 IncHI1 plasmid SNPs, consistent with the presence of an IncHI1 MDR plasmid. The remaining two MDR isolates had positive fluorescence signals for *rep* genes of other plasmid types. Isolate 76-1292 had positive signals for both IncI target loci, one of two IncK loci and one of three IncC loci. Plasmid conjugation and incompatibility group typing experiments performed by Francois-Xavier Weill (Pasteur Institute) confirmed that the isolate contained a 100 kbp plasmid of the IncI1 type, which transferred multidrug resistance to *E. coli*. Isolate 80-2002 had positive signals for two of three IncA/C target loci, one of two IncN loci and one of four IncW loci. Experiments by Francois-Xavier Weill confirmed that the isolate contained a 130 kbp plasmid of the IncA/C type, which transferred multidrug resistance to *E. coli.* No evidence of IncHI1 plasmids was found in DNA from isolates recorded as drug sensitive. Phylogenetic analysis of the IncHI1 plasmid SNPs clustered the isolates into eight groups, shown in Figure 6.11.

The majority of plasmids (23) clustered into a single group of the ST6 IncHI1 plasmid type, defined by SNPs found in the sequenced Typhi H58 strains E98-0664, ISP-03-07467 and ISP-04-06979. All of these plasmids were found in H58 or H58-derived Typhi strains (see Figure 6.11), consistent with a single acquisition of the ST6 IncHI1 plasmid by a common ancestor of the H58 lineage. (Note that it was not possible to confirm this for the ST6 control plasmid pSTY7, which was transferred into *E. coli* for laboratory storage and experiments and the original host strain is not available.) Nearly all resistance gene target loci gave the same signals for all ST6 plasmids, matching the GoldenGate assay profile of pAKU_1. A notable exception was the isolate 38(98)S which appeared to be missing the composite transposon insertion (lacking *IS26*, *bla*, *sul2*, *strAB* (*bla/sul/str*); *sul1*, *dhfR7*, *mer* genes, *tnpA* (*Tn*21); and *cat* (*Tn*9); but carrying *tet* genes (*Tn*10)). The isolate, from which DNA had been prepared for SNP typing in July 2008, was re-tested in December 2008 by Francois-Xavier Weill at the Pasteur Institute, who confirmed that it was sensitive to all drugs at that time, including tetracycline. It is likely therefore that this isolate harboured the MDR IncHI1 plasmid when it was initially isolated in 1998, but the resistance genes have been lost from the plasmid during its 10 years in the laboratory.

Three of the Typhi IncHI1 plasmids clustered with ST8, originally defined by SNPs found in plasmid pAKU_1 sequenced from Paratyphi A (orange in Figure 6.11). The Typhi plasmids differed from the ST8 Paratyphi A plasmids (isolated from Karachi in 2003-2004 (725)) at just one SNP locus and were all isolated in Peru in 1981. The Typhi strains themselves were of haplotypes H77 (two isolates) and H50. This is the first clear evidence that very closely related IncHI1 plasmids have successfully transferred into both the Paratyphi A and Typhi populations. However, the resistance gene profiles of these three Typhi plasmids differed from that of pAKU_1. They gave positive signals for the *cat*, *strAB*, *sul1* and *tet* genes, consistent with their resistance phenotypes (chloramphenicol, streptomycin, sulfonamide and tetracycline resistant). However there was no evidence of *bla*, *dhfR7*, *dhfR14*, *sul2* or *IS*26. The two H77 strains were found to

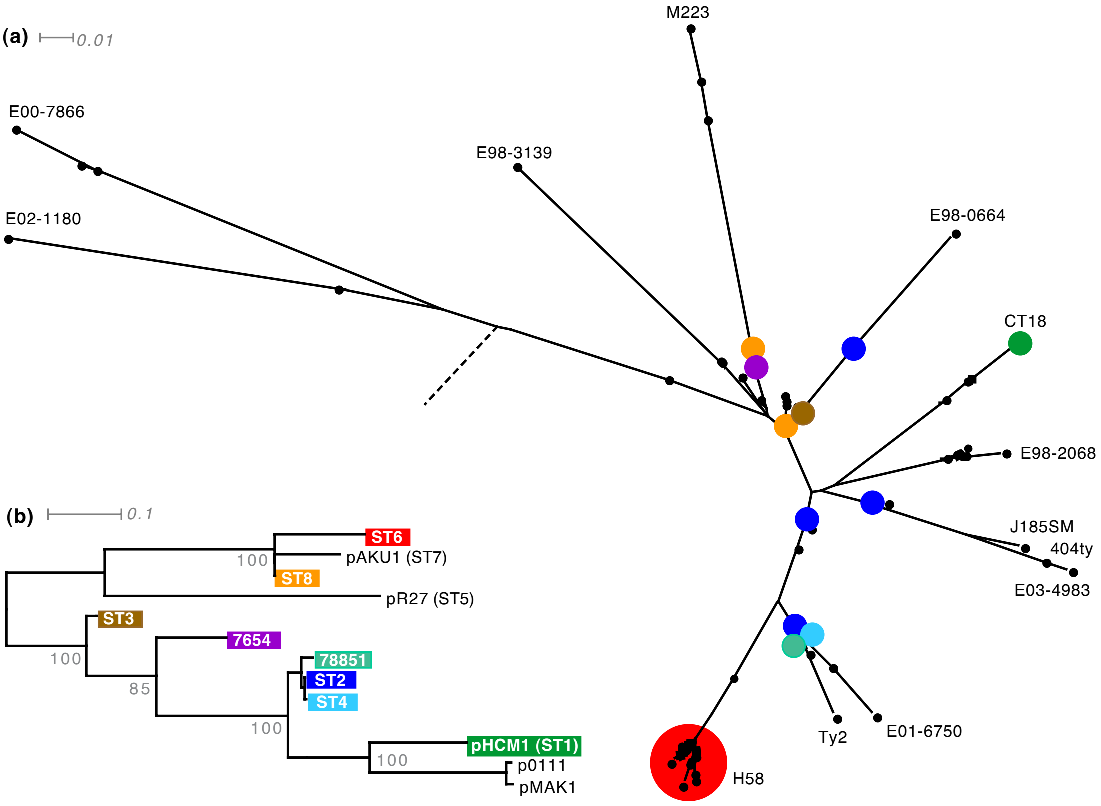


Figure 6.11: Phylogenetic trees of Typhi chromosomes and IncHI1 plasmids in a global collection of Typhi isolates - Maximum likelihood phylogenetic trees, scale bar is in divergence per assayed SNP (1,508 SNPs for a, 200 SNPs for b). (a) Tree of Typhi based on typing of chromosomal SNPs in a global collection of 180 isolates plus 19 control isolates (labelled). Nodes including isolates containing IncHI1 plasmids are highlighted with circles, coloured according to IncHI1 plasmid subgroup as shown in (b). (b) Tree of IncHI1 plasmids based on typing of SNPs in the conserved IncHI1 plasmid backbone in a global collection of 180 isolates, including 38 containing MDR IncHI1 plasmids, and control plasmids. Nodes are labelled by the plasmid ST type (as defined by plasmid MLST in (725)) and/or by the name of a representative plasmid. Each node including an IncHI1 plasmid that has been found in Typhi is assigned a unique colour. Uncoloured nodes represent plasmids that have so far only been found in other bacteria: pAKU_1 in Paratyphi A, R27 in Typhimurium, pMAK1 in Choleraesuis, p0111 in enterohemorrhagic *E. coli.* Bipartitions are labelled with bootstrap values from 1,000 bootstraps.

be trimethoprim resistant, so likely carry a different trimethoprim resistance gene from those sequenced to date from Typhi IncHI1 plasmids (*dhfR7*, *dhfR14*). The lack of resistance to ampicillin is consistent with the lack of signal detected for the beta-lactamase gene (*bla*). The SNP typing data for the Peruvian ST8 plasmids is consistent with the presence of *Tn*9, *Tn*21, *Tn*10 and *strAB* insertions in the same plasmid backbone as pAKU_1; the same elements are present in pAKU_1, with *Tn*21 inserted within *Tn*9.

To test whether the insertion sites were the same as in pAKU_1, PCR was performed using the same primer sets used to assay IncHI1 plasmids in earlier studies (see 5.2.4 and Table 5.3). Primers were designed by myself and PCR performed by Minh Duy Phan at the Sanger Institute. For all three plasmids, amplicons were generated across the left and right boundaries of the insertion site of *Tn*10 into pAKU_1 (PCR loci O, P in 5.2.4), and the insertion of the second copy of *strAB* into the pAKU_1 backbone (PCR locus Q in 5.2.4). For one plasmid, no further PCR amplicons could be generated. For the other two plasmids, 81863 and 81424, amplicons were generated across the insertion boundaries of *Tn*21 into *Tn*9 (PCR loci G, H in 5.2.4), and the insertion site of *Tn*9 into the pAKU_1 backbone (PCR loci J, K in 5.2.4). No amplicons were generated for the insertion sites of *Tn*9 or *Tn*10 observed in pHCM1 (PCR loci L, M, N in 5.2.4), or the insertion of *strAB* into *Tn*21. These results suggest that the Peruvian ST8 plasmids carry resistance insertions very similar to those in pAKU_1, with the exception of the *bla/sul/str* element. The failure of PCR to confirm the expected insertion sites of *Tn*9 and *Tn*21 in isolate 81918 suggests there may have been rearrangements within this particular plasmid, or even transposition of the *Tn*9-*Tn*21 composite transposon to a novel location. The Typhi strains carrying the ST8 plasmid type were all isolated in Peru in 1981, yet fall into distinct haplotypes H77 (isolates 81424 and 81918) and H50 (isolate 81863), separated by 28 target SNPs (Figure 6.9). This could be explained by a single acquisition of the plasmid by a common ancestor, followed by diversification of the strain types. However given the short time frame and genetic distance it is more likely that the ST8 plasmid was independently acquired by multiple Typhi strains circulating in Peru in 1981.

6.3.2.3 Distribution of other plasmids

Positive signals were detected in four experimental samples for a SNP in plasmid pHCM2 (coordinate 38,509). The SNP was originally detected between copies of the pHCM2 plasmid sequenced from genetically distant strains CT18 (H1) and E02-2759 (H58). The four novel strains harboured the E02-2759 allele at this SNP locus. Two of the strains were closely related (haplotype H39), but the other strains were genetically distant (H52 and H58-derived H34). To confirm the presence of pHCM2 in these four strains, PCR was performed as described in 2.2.2 using primers previously described in (278) (three targeting sites in pHCM2, one targeting chromosomal gene *aroC*). All three pHCM2 target amplicons, as well as the chromosomal target, were detected in the four novel isolates. In each case, amplicon sizes matched those amplified from control isolates CT18 and E02-2759. Thus, although rare in the Typhi population (278), pHCM2 can be found in distinct Typhi lineages, confirming multiple acquisition events as opposed to clonal spread within a single a lineage. Furthermore, we can expect that the array targets provide accurate detection of the pHCM2 plasmid.

The SNP typing assay included two targets for sequences within the linear plasmid pBSSB1, harbouring the z66 flagella antigen and previously found to be present only in H59 isolates (256). These targets gave positive signals for control isolates 404ty and E03-4983 (both H59) but not for any other isolates of the global collection (which did not include additional H59 isolates). This is consistent with PCR assays reported in (2), which failed to detect z66 sequences among these isolates.

6.3.3 Endemic typhoid in the Mekong Delta, Vietnam

Typhoid fever is endemic to the Mekong River Delta region in the south of Vietnam, shown in pink in Figure 6.12. With a mean of ~80 cases per 100,000 people annually (328, 349, 422), the region accounts for over 75% of typhoid in Vietnam (354). The first MDR typhoid outbreak in Vietnam occurred in Kien Giang (Figure 6.12) in 1993 (380). MDR typhoid has continued to be a problem in the region, with rates peaking at 100% in 1996-1998 and declining to ~50% in 2002-2004 (2, 16, 380). In Vietnam, MDR typhoid is usually treated with fluoroquinolones, however most Typhi isolated from the Mekong Delta (>95%) are now resistant to nalidixic acid (Nal) and show reduced

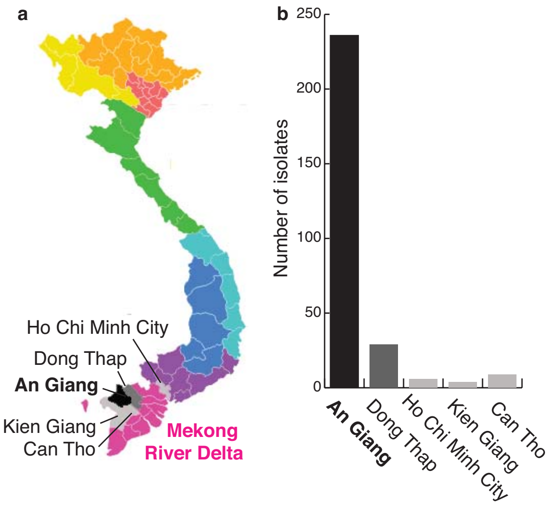


Figure 6.12: Geographical sources of isolates from the Mekong Delta - (a) Map of Vietnam, broken down into regions (shown in different colours) and provinces. Provinces from which Typhi was isolates for this study are highlighted. (b) Distribution of Typhi isolates from each province included in the study.

susceptibility to fluoroquinolones (2, 16). In order to compare alternative treatments for MDR Nal-resistant typhoid, a treatment study was conducted in 2004-2005 in the Mekong River Delta region, led by Christiane Dolecek at OUCRU, Ho Chi Minh City (745). Typhoid patients (adults and children) were recruited into the study from three hospitals: the Hospital for Tropical Diseases in Ho Chi Minh City, the Dong Thap Provincial hospital in Cao Lanh, Dong Thap province and the An Giang Provincial hospital in Long Xuyen, An Giang province (Figure 6.12). Patients were treated with gatifloxacin (a second generation fluoroquinolone) or azithromycin (745). A total of 282 patients had typhoid fever confirmed by blood culture during the two-year study period (745). The majority of these patients (82.2%) lived in An Giang province, with 10.1% of patients from Dong Thap, 1.7% from Ho Chi Minh City and the rest from neighbouring provinces (see Figure 6.12). In total, 96% of the Typhi isolates were Nal-resistant and 58% were MDR. Treatment failure with gatifloxacin or azithromycin occurred in 9% of patients, all of whom made a full recovery after treatment with ceftriaxone (745). A total of 264 of the 282 Typhi isolates were available for analysis and these were all SNP typed with the GoldenGate assay. DNA was prepared and quantified by Christiane Dolecek; data analysis was performed by myself as described above.

6.3.3.1 Phylogenetic analysis

The 264 SNP typed Typhi isolates are listed in Appendix E. A total of 258 isolates (97.7%) were of the H58 haplotype, the remaining isolates were of haplotypes H1 (N=3), H45, H50 and H52, see Figure 6.13a. H58 isolates displayed variation at 10 SNP loci, which differentiated seven distinct haplotypes, shown in Figure 6.13b. However 239 (92.6%) of these isolates belonged to just three closely related haplotypes, labelled C, E1 and E2 in Figure 6.13b. The sequenced isolate AG3, isolated in An Giang province during the study (March 2004), belongs to the H58-E2 haplotype. The SNPs separating E1 and E2 from C were originally identified from the AG3 sequence, i.e. the ability to differentiate within this cluster of 239 isolates is due to the inclusion of one of these isolates in the initial SNP detection study. Four isolates clustered with H58 based on chromosomal SNP alleles, but could not be fitted into the H58 phylogenetic tree due to conflicting SNP alleles. These isolates displayed unusual heterozygous signals for several chromosomal SNPs. They were also positive for IncHI1 plasmid loci previously identified in enterohemorrhagic *E. coli* (see 5.3.2.1), *qnr* (quinolone-resistance) gene loci which have not been reported in Typhi, and several additional plasmid *rep* genes. These samples are therefore likely to be contaminated with other bacterial DNA, possibly *E. coli.* There was no association between Typhi haplotype and patient age, fever clearance time or relapse. Diarrhea was more common in patients infected with H58-C compared to H58-E2 (74% vs 55%, p=0.006 with Pearson χ^2 test; 64% among other haplotypes), whereas constipation was more common in patients infected with H58-E2 (14% H58-E2, 6% H58-C, p=0.076 with Pearson χ^2 test; 6% among other haplotypes). Headache was also more common among patients infected with H58-C compared to H58-E2 (70% vs 55%, p=0.036 with Pearson χ^2 test; 64% among other haplotypes) and was associated with diarrhea ($p<10 \times 10^{-5}$, Pearson χ^2 test). Only one instance of Typhi carriage was observed during 3 month follow-up, this was an MDR H58-C isolate.

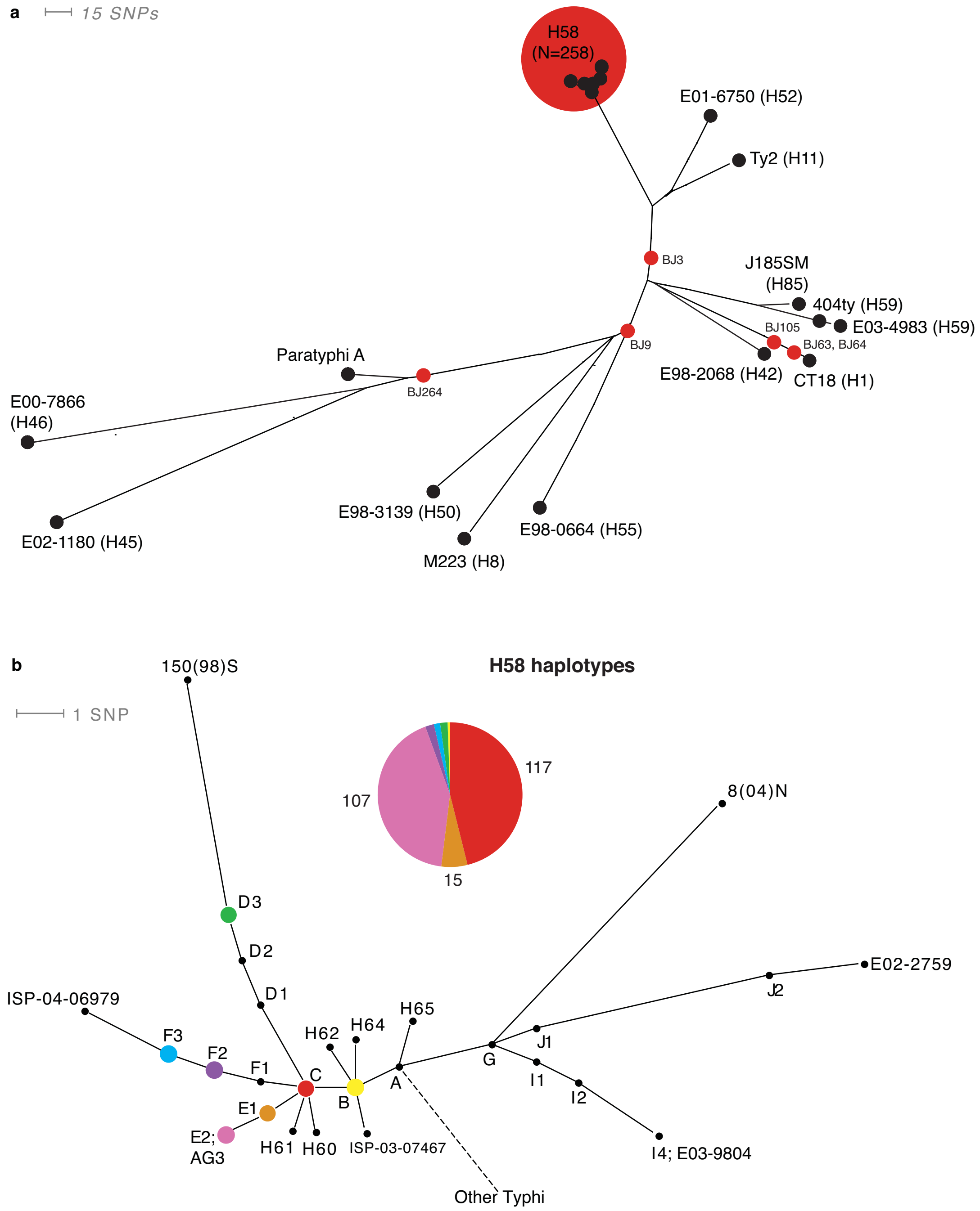


Figure 6.13: Phylogenetic distribution of Typhi isolates from the Mekong Delta - (a) Typhi phylogenetic tree. Black nodes indicate control isolates, red nodes show the position of 264 isolates from the Mekong Delta. Most isolates (258) were of the H58 haplotype, non-H58 isolates are named individually. (b) Zoom in on Phylogenetic tree of the H58 cluster. Coloured nodes indicate haplotypes that were detected among Typhi isolates from the Mekong Delta, inset pie chart indicates the frequency of each of these nodes.

6.3.3.2 Plasmids and drug resistance

The presence of fluorescence signals for IncHI1 SNP loci indicated that a total of 137 samples contained IncHI1 plasmids. All plasmids were of the ST6 type, and all host isolates were of the H58 haplotype. Of these isolates, 135 were classified as MDR by resistance testing performed by Christiane Dolecek at OUCRU (defined as MIC $\geq$ 32 μg/mL for chloramphenicol and ampicillin, and MIC $\geq$ 8/152 μg/mL for co-trimoxazole) (745). The other two isolates tested positive by GoldenGate SNP typing for *sul1*, *sul2*, *dfrA7*, *tetACDR*, *strAB*, *bla* and*cat*, just like the MDR isolates, despite very low recorded MICs for chloramphenicol, ampicillin and co-trimoxazole. An additional 18 isolates were recorded as MDR but did not test positive for IncHI1 plasmid SNPs. One of these, BJ5, did give positive signals for resistance genes *strA*, *bla*, *sul1*, *sul2*, *dfrA7*; the *repC* replication initiation gene of IncHI1; transposases and *merAPTR* from *Tn21*; and *IS*26 and *IS*10 transposases. The MDR IncHI1 plasmid was much more common among C and E1 isolates than E2 isolates (80% vs 15%, see Figure 6.14).

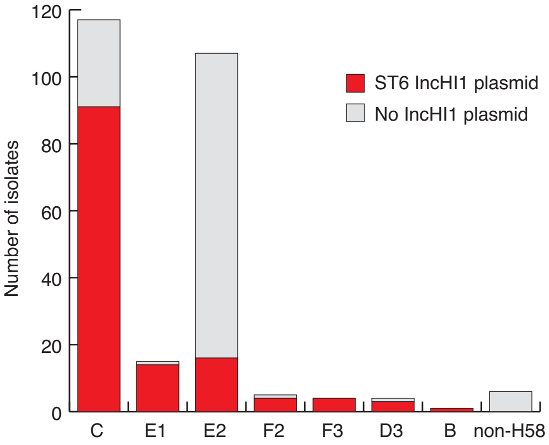


Figure 6.14: Distribution of IncHI1 plasmids among Typhi isolates from the Mekong Delta - B-F H58 haplotypes.

Nal resistance was tested by Christiane Dolecek at OUCRU. A total of 254 isolates were Nal resistant, defined as MIC $\geq$ 32 μg/mL (745). All of these were H58 isolates and all were susceptible to the fluoroquinolone drugs gatifloxacin, ciprofloxacin and ofloxacin. The only Nal susceptible H58 isolates were the singleton H58-B isolate (see Figure 6.13b), one H58-C isolate and the two isolates suspected of contamination.

6.3.3.3 Spatial and temporal distribution

Figure 6.15 shows the distribution of Typhi haplotypes among Mekong Delta provinces. All of the non-H58 isolates were from patients living outside An Giang province. Two H1 isolates BJ63 and BJ64 were identical at all assayed SNP loci and were taken from patients in Dong Thap on consecutive days. The third H1 isolate (BJ105 in Figure 6.13a) differed from these at 16 SNP loci and was collected in Dong Thap 14 months after BJ63 and BJ64. The H45 isolate BJ264 originated in Can Tho province, the H50 isolate BJ9 originated in Ho Chi Minh City and the H52 isolate BJ3 originated in Dong Nai province. The H58 C/E1/E2 cluster was dominant among isolates from each province (except Dong Nai from which just one isolate originated).

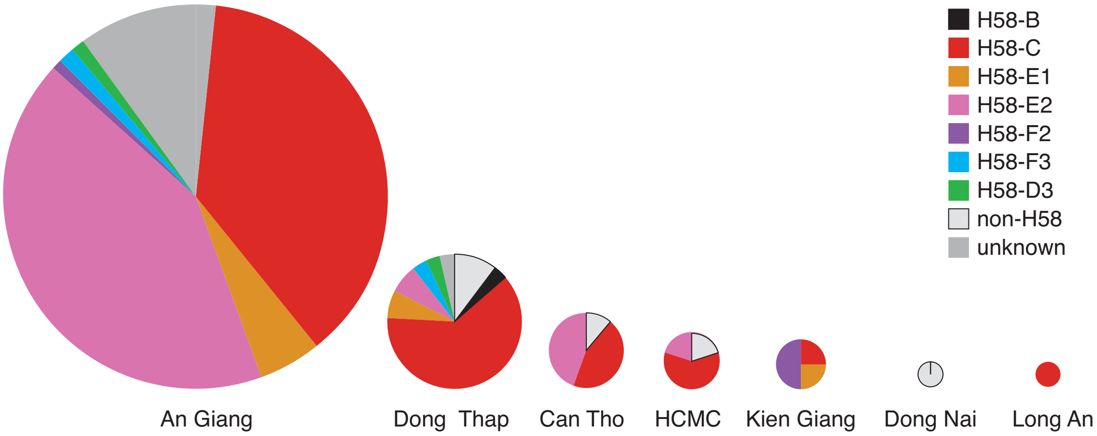


Figure 6.15: Distribution of haplotypes among provinces in the Mekong Delta - Pie charts are scaled to represent the total number of isolates from each province.

The distribution of Typhi haplotypes over time is shown in Figure 6.16a. The majority of typhoid cases occurred in the wet season, between July and December each year. In the season of 2004 H58-E2 and H58-C were both prevalent, whereas very few isolates of H58-E2 Typhi were collected during the 2005 wet season. The decline of H58-E2 may be associated with the IncHI1 MDR plasmid (see Figure 6.14), which was much more common in H58-C. As Figure 6.16b highlights, the majority of isolates collected during the second season were MDR and carried the IncHI1 plasmid ST6.

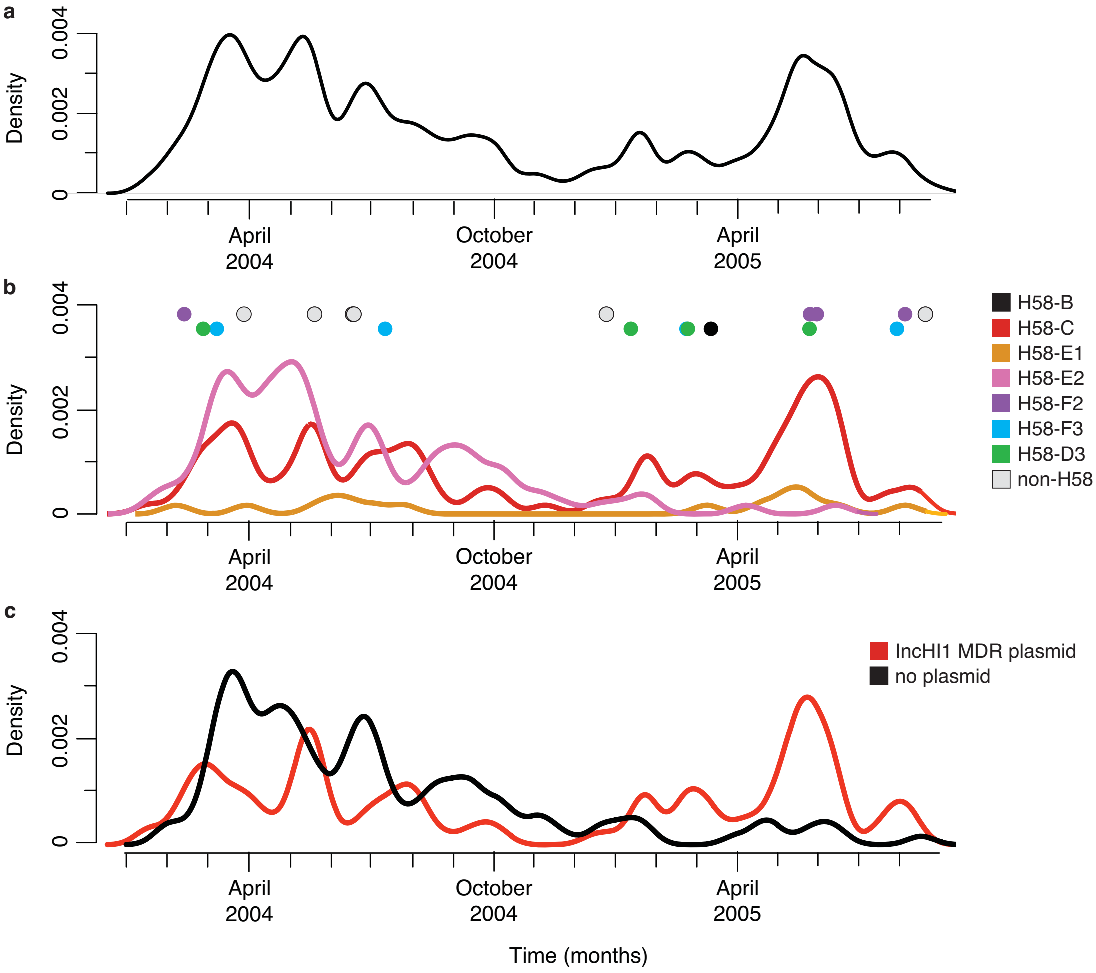


Figure 6.16: Distribution of typhoid fever cases over two years in the Mekong Delta - (a) All typhoid fever cases in the study. (b) Typhoid fever cases split by haplotype. Density distributions are shown for the three dominant H58 haplotypes C, E1 and E2. Other haplotypes are shown as circles. All lines and circles are coloured by haplotype according to the legend provided. (c) Typhoid fever cases split by presence of the MDR IncHI1-ST6 plasmid.

6.3.4 Pediatric typhoid in Kathmandu, Nepal

Although precise incidence data is not available, the prevalence of typhoid fever in Kathmandu, the capital city of Nepal, has been observed to be very high (753). A ten-year retrospective study of typhoid in 1993-2003 found the number of enteric fever cases (including both Typhi and Paratyphi A) more than doubled in 2001-2003 compared with the previous three years (336). MDR was not a significant problem during the study period, although there were increasing levels of reduced susceptibility to fluoroquinolones (336). In 2005-2006, researchers from Oxford University and Patan Hospital in Kathmandu conducted a study of the burden of disease caused by encapsulated bacteria among children in Patan, a central district of Kathmandu (746). Children under 13 years of age admitted to Patan Hospital with suspected bacteraemia, meningitis or pneumonia were recruited into the study (N=2,039), and blood cultures performed (N=141 positive cultures). Typhi (N=53 isolates) and Paratyphi A (N=6 isolates) were responsible for 49% of all bacteraemias and were the most frequently cultured pathogens in children older than 12 months. DNA samples from 46 of the 53 Typhi isolates was provided by Andrew Pollard of the Department of Paediatrics, University of Oxford, UK and were SNP typed at the Sanger Institute using the GoldenGate assay (following DNA quantitation by myself).

6.3.4.1 Phylogenetic analysis

The SNP typed Typhi isolates are listed in Appendix E. The majority of isolates were H58 (32, or 70%), with most of these belonging to the specific haplotype H58-G (30 isolates), see Figure 6.17. There was also a significant phylogenetic cluster of nine isolates (20%) of a subgroup of H42 (H42-A, blue in Figure 6.17), demonstrating that the dominance of H58 was not as complete in Kathmandu as it appeared to be in the Mekong Delta. There was also a cluster of three H50 isolates and two isolates from different subgroups of H42 (yellow and white in Figure 6.17). There was no association between Typhi haplotype and patient age or sex, see Table 6.3. The slightly higher mean age of children infected with Typhi of the H58-G haplotype (4.6 years vs 2.9 years for other haplotypes) is most likely due to the higher frequency of H58-G, as infections in older children were generally rarer and thus the chance to observe older children infected with other haplotypes was reduced compared to the more common H58-G haplotype. In support of this, direct comparison of the age distributions of children infected with H58-G versus other haplotypes (see Figure 6.18) showed no evidence that the underlying age distributions differ between haplotypes (two-sample Kolmogorov-Smirnov test, p=0.45).

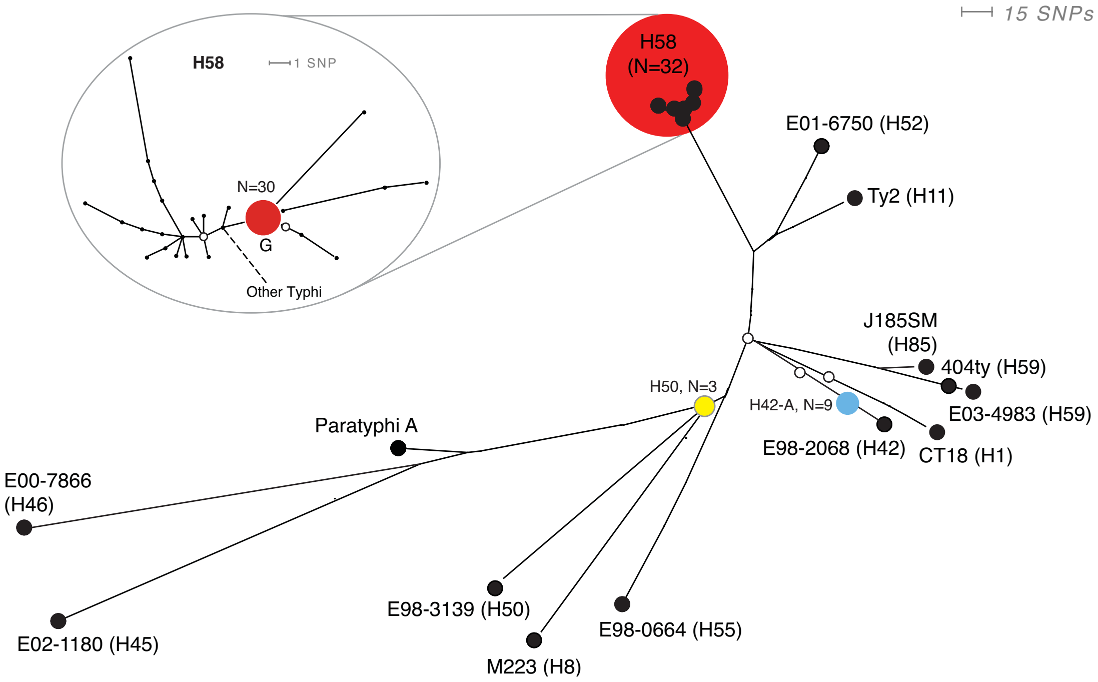


Figure 6.17: Phylogenetic distribution of Typhi isolates from Kathmandu - Black nodes indicate control isolates, coloured nodes show the position of 46 isolates from pediatric typhoid cases in Kathmandu. Most isolates were of H58 haplotypes, inset shows phylogenetic structure within the H58 group.

Haplotype	No. isolates	Mean age	Female	Nal-R
H58-G	30	4.6	47%	28
H42-A	9	3.8	44%	0
Other	7	1.9	57%	1
All	46	4.0	48%	29

Table 6.3: Typhoid case parameters by Typhi haplotype in Kathmandu - Nal-R indicates resistance to nalidixic acid.

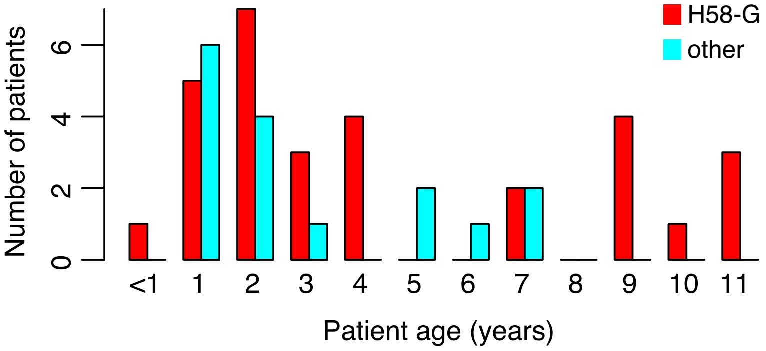


Figure 6.18: Distribution of patient ages for H58-G vs other haplotypes detected in Kathmandu - Cases are split according to Typhi haplotype, coloured as shown.

6.3.4.2 Drug resistance

Typhi isolates were tested for resistance to ampicillin, chloramphenicol, co-trimoxazole, gentamicin, ciprofloxacin, ceftriaxone and nalidixic acid (Nal). (Resistance testing was performed by David Murdoch at the University of Otago, Christchurch, New Zealand.) All isolates were susceptible to the former six drugs, while 29 isolates were resistant to Nal. MDR Typhi has been reported previously in Nepal, usually associated with the presence of plasmids (16, 598, 687, 754). However the GoldenGate assay detected no signals for resistance genes, IncHI1 plasmid SNPs or other plasmid loci in these isolates, consistent with the absence of resistance phenotypes. Nal resistance was restricted to H58 isolates, with 28 of 30 H58-G isolates and the closely related isolate 872 exhibiting MICs greater than 256 μg/mL. Quinolones target bacterial topoisomerase genes, in particular DNA gyrase (GyrA), and mutations at positions 83 and 87 of the GyrA protein have been shown to confer Nal resistance in Typhi (396). These SNPs could not be typed using the GoldenGate assay, which cannot target SNP loci that lie within 10 bp of each other. Instead, all H58 isolates were tested by Yajun Song at the Environmental Research Institute, Cork, Ireland for six known GyrA mutations (Pro83, Phe83, Tyr83, Asn87, Tyr87 and Gly87 (2)) using a Luminex 200 assay (755). All H58 isolates harboured the Phe83 mutation, with the exception of the H58-B isolate 959 which was Nal sensitive and carried wildtype alleles at GyrA positions 83 and 87.

6.3.4.3 Temporal distribution of haplotypes

The distribution of typhoid cases across the course of the study is shown in Figure 6.19. Surprisingly, the incidence of pediatric typhoid cases requiring hospitalisation was fairly constant throughout the study period (mean 2.5 cases per month), with no increase associated with the wet season (see Figure 6.19). Patterns of infection differed among Typhi haplotypes (Figure 6.19b-c). Hospitalisation of patients infected with H58-G Typhi occurred at a constant rate during the course of the study (mean 1.4 cases per month, Figure 6.19b). However H42-A Typhi was not seen until the second half of the study (Figure 6.19c, blue), after which point this haplotype was detected at a mean rate of 0.8 cases per month. Hospitalisations due to infection with other Typhi haplotypes occurred throughout the study, at a mean rate of 0.33 cases per month.

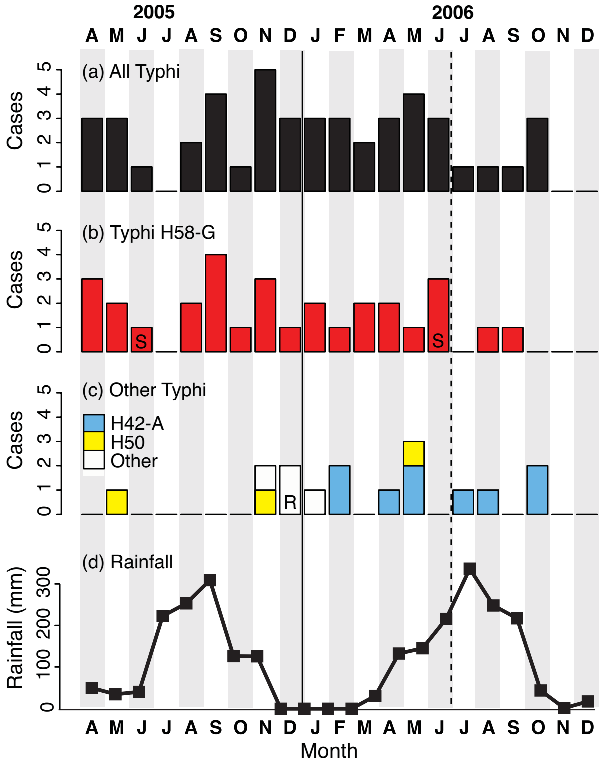


Figure 6.19: Distribution of typhoid cases in Kathmandu by month - (a) All cases. (b) Infections with haplotype H58-G, note all of these cases were nalidixic acid resistant except two marked 'S'. (c) Infections with Typhi of other haplotypes, colours indicate haplotype according to legend provided. All of these cases were nalidixic acid sensitive except one marked 'R'. (d) Total monthly rainfall at Kathmandu airport.

6.3.5 Endemic typhoid in an urban slum in Kolkata, India

The annual incidence of typhoid fever in India has been estimated at 662/100,000 among inhabitants and 42/100,000 among travellers, the highest rates in the world (10, 11). In Kolkata in the east of India, typhoid is most common in urban slum areas ("bustees"). The first MDR typhoid outbreak in Kolkata occurred in 1989-1990 (756) and MDR typhoid has persisted, declining from 100% in 1991-1992 to below 15% in 2003-2004 (417, 418, 757). In 2004, the first fluoroquinolone-resistant Typhi isolates were observed in Kolkata, with MICs >16 μg/mL to ciprofloxacin and ofloxacin (401). In 2003, a four year study of typhoid fever was begun in an urban slum in East Kolkata (Wards 29 and 30), led by researchers at NICED in Kolkata and IVI in Seoul, Korea. The 1 km^2 study site was home to nearly 60,000 people, who were encouraged to report all episodes of fever lasting at least three days to study centers, where they were offered free diagnosis and reimbursed transport costs as well as treatment costs for enteric fever and malaria (343, 421, 747). Surveillance began on May 1 2003 and continued until 31 January 2007, during which time 378 cases of typhoid fever were confirmed by blood culture (2.8% of fever cases reported; 50% of fever cases reported in children aged 5-15 (327)). During December 2004, residents were vaccinated with either a Vi conjugate vaccine (designed to protect against Typhi), a Hepatitis A vaccine, or no vaccine (421). Residents were assigned to one of 80 geographical clusters based on the location of their dwelling, geographic clusters were assigned to either the Vi conjugate or Hepatitis A vaccine (40 clusters each, see Figure 6.22 below), and 25-85% of individuals within each cluster received vaccine. DNA was available for 188 (50%) of the 378 cases of typhoid fever confirmed during the study (see Appendix E); these samples were SNP typed with the GoldenGate assay. DNA was prepared in Kolkata by Shanta Dutta at NICED and quantified by Derek Pickard at the Sanger Institute. Data analysis was performed by myself.

6.3.5.1 Phylogenetic analysis

The majority of isolates were H58 (139, or 74%), see Figure 6.20a. There was also a significant phylogenetic cluster of 28 isolates (20%) of H42-A (blue in Figure 6.20a), as well as nine H50 isolates and 12 other isolates scattered around the phylogenetic tree (white in Figure 6.20a). Although the basic composition of the population was similar to that observed among isolates from Kathmandu, Nepal (70% H58, 20% H42-A) the H58 subgroups were different. Among the Kolkata H58 isolates, the majority were of subgroups B or G (77 and 43 isolates respectively), which represent distinct sublineages of H58 (see Figure 6.20b-c). There were also small clusters of subgroups A and H64 (nine and eight isolates respectively), which are each separated from subgroup B by a single target SNP (see Figure 6.20b-c). The remaining two H58 isolates were of two different, more distantly related subgroups, separated from G by at least two target SNPs (see Figure 6.20b). There was no association between Typhi haplotype and patient age or sex, see Table 6.4. IncHI1 plasmids were detected in just six isolates, all of them H58 (see Table 6.4).

Haplotype	No. isolates	Mean age	Female	IncHI1
H58-B	77	10	43%	0
H58-G	43	12	42%	4
Other H58	19	10	47%	2
H42-A	28	10	43%	0
Other	21	10	44%	0
All	188	10	45%	6

Table 6.4: Typhoid case parameters by Typhi haplotype in Kolkata - IncHI1 indicates the presence of the IncHI1 plasmid as detected by SNP typing with GoldenGate.

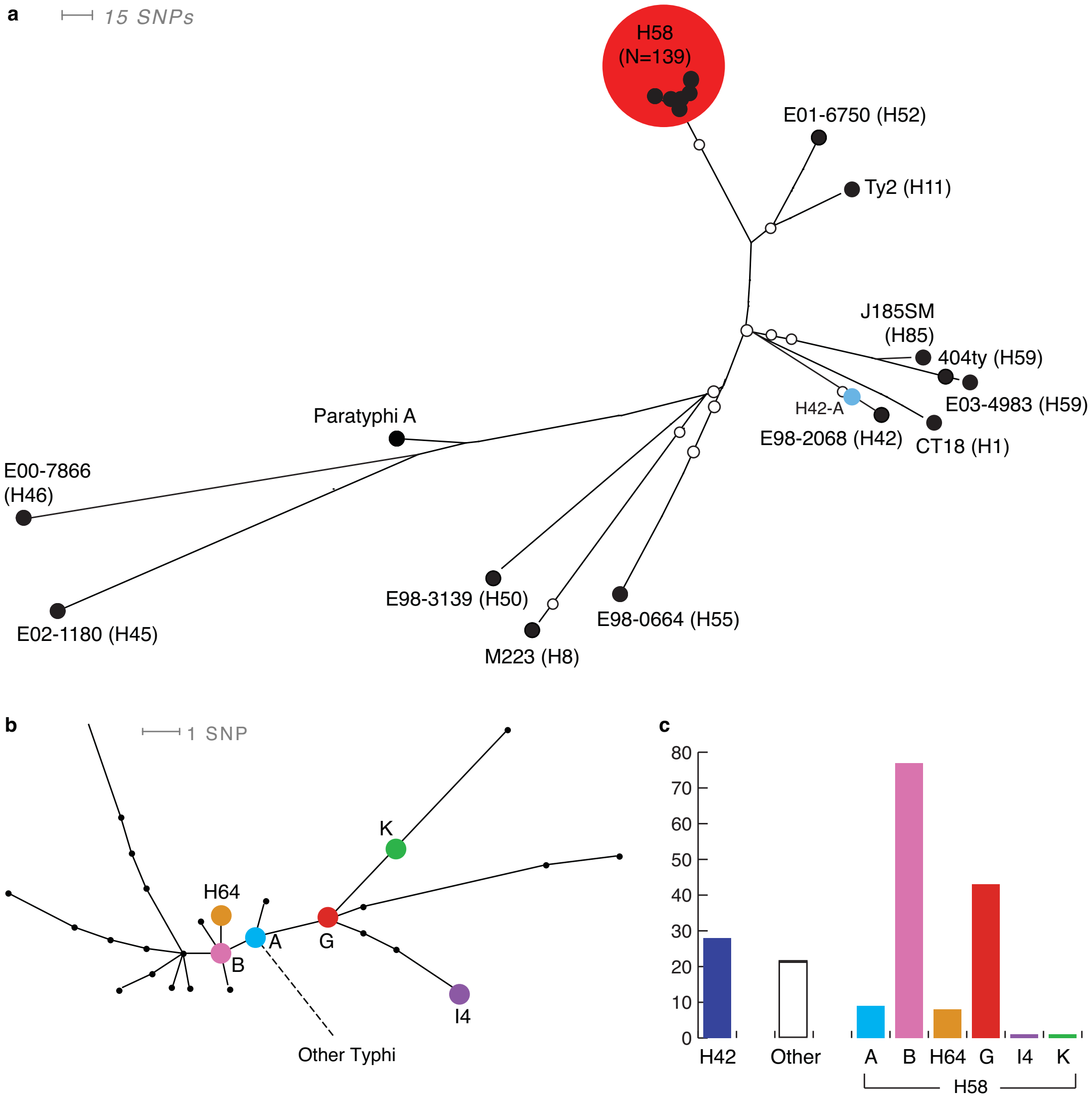


Figure 6.20: Phylogenetic distribution of Typhi isolates from Kolkata - (a) Typhi phylogenetic tree. Black nodes indicate control isolates, other nodes show the position of 188 isolates from Kolkata. Most isolates (139) were H58, although there were also 28 H42 isolates (pale blue). (b) Zoom in on phylogenetic tree of the H58 cluster. Coloured nodes indicate haplotypes that were detected among the Kolkata isolates. (c) Frequency of each haplotype observed. Nodes are coloured as in a-b.

6.3.5.2 Spatial and temporal distribution of haplotypes

Confirmed typhoid cases occurred at a rate of 0.16 per day (approximately one per week) during the study period, both before and after the introduction of the vaccine in December 2004. The distribution of typhoid cases across the course of the study is shown in Figure 6.21a, which shows a number of peaks in typhoid incidence. The distribution of haplotypes, ascertained for 188 (50%) of the Typhi isolates, is shown in Figure 6.21b. From this plot it is clear that peaks in typhoid cases in 2004 resulted from infections with a diverse range of Typhi haplotypes (Figure 6.21b, peaks 1 and 2). However later peaks in typhoid incidence were due almost entirely to infection with Typhi haplotype H58-B (peaks 4 and 5) or H58-G (peak 6, see Figure 6.21b).

The spatial distribution of typhoid cases was non-random, with a higher incidence of cases in certain geographical clusters including high population density clusters in the south-west of Ward 29 (Figure 6.22). There were also spatial areas in which specific haplotypes were overrepresented (see 6.2.5), shown in Figure 6.22. For example, 27 out of the 77 infections with H58-B occurred in a spatial area containing seven of the geographical clusters (red in Figure 6.22), at a rate of 546/100,000 people compared to 91/100,000 across the rest of the study site and making up 79% of isolates collected in these clusters compared to 40% of all SNP typed isolates. Seventeen of these patients lived in just six dwellings in the same street, with 2-4 cases per household. In geographical cluster 17 H58-G was overrepresented, with incidence 456/100,000 compared to 66/100,000 in other clusters, and 50% of Typhi infections in this cluster due to H58-G compared to 23% of all SNP typed Typhi isolates (blue in Figure 6.22). H42 isolates were overrepresented in geographical clusters 32 and 17 (incidence 230/100,000 vs 40/100,000; 50% in these clusters compared to 15% overall, green in Figure 6.22) and H64 isolates were overrepresented in geographical cluster 3 (incidence 412/100,000 vs 8/100,000; 60% compared to 4% overall, pink in Figure 6.22). Two of the five typhoid cases (40%) in neighbouring geographical clusters 9 and 10 were of haplotype H50 (the two patients lived in the same street), compared to 5% of all cases (incidence 134/100,000 vs 12/100,000, pale blue in Figure 6.22). The haplotype-specific spatial clusters did not show strong correlations with peaks in typhoid incidence highlighted in Figure 6.21, however most peaks had identifiable spatial foci, shown in Figure 6.23.

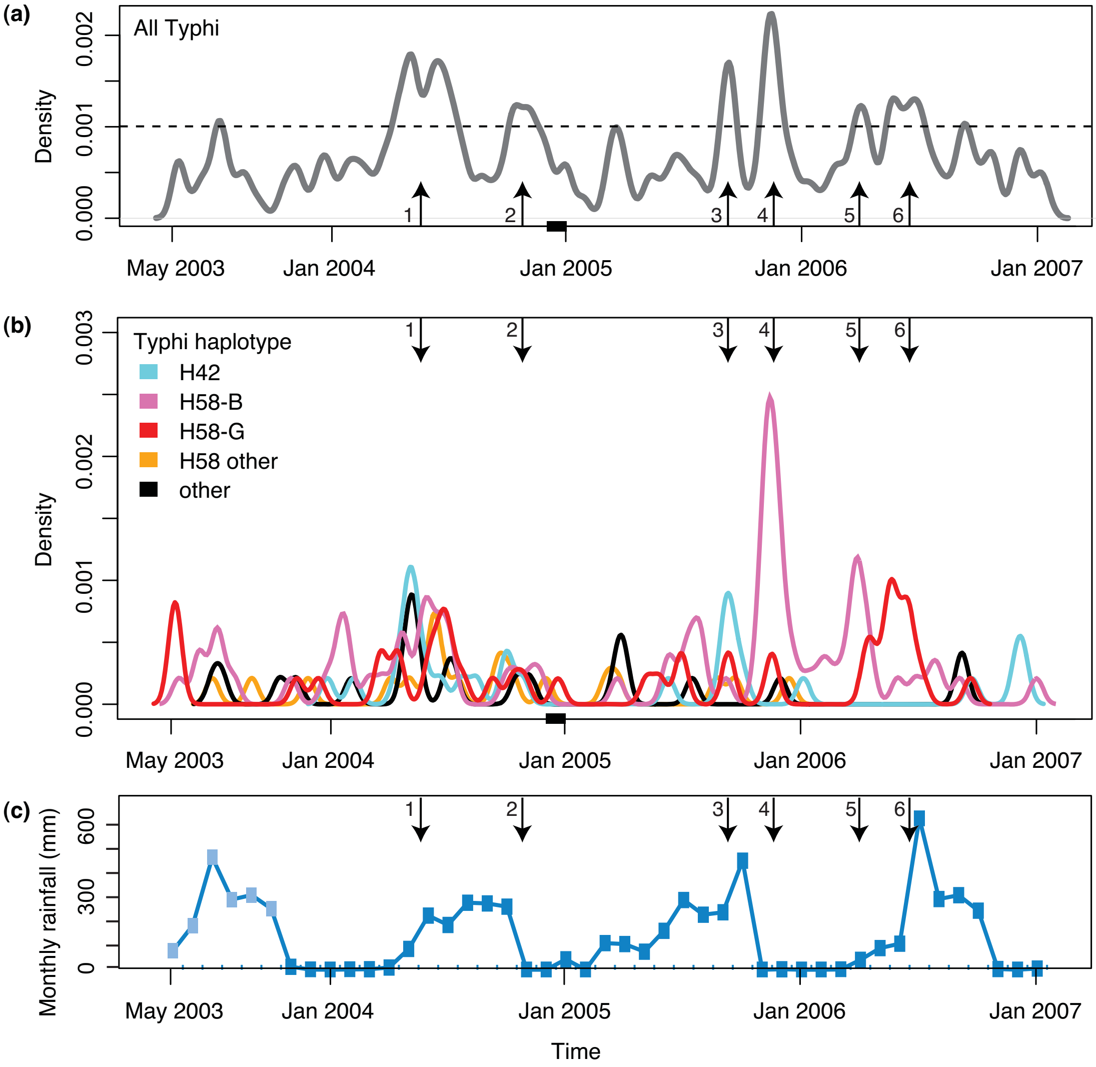


Figure 6.21: Distribution of typhoid cases during a four year study in Kolkata - Black rectangle indicates December 2004, during which time vaccinations were given. Outbreaks labelled 1-6 are highlighted with arrows. Arrows indicate outbreaks discussed in the text. (a) All typhoid cases (N=378). (b) SNP typed typhoid cases (N=188) broken down by Typhi haplotype, coloured as shown. (c) Monthly rainfall in the Howrah district where the study site is located. Data was sourced from the India Meteorological Department (http://www.imd.gov.in). Data was not available for 2003, so the mean values across 2004-2006 were plotted in its place (pale blue).

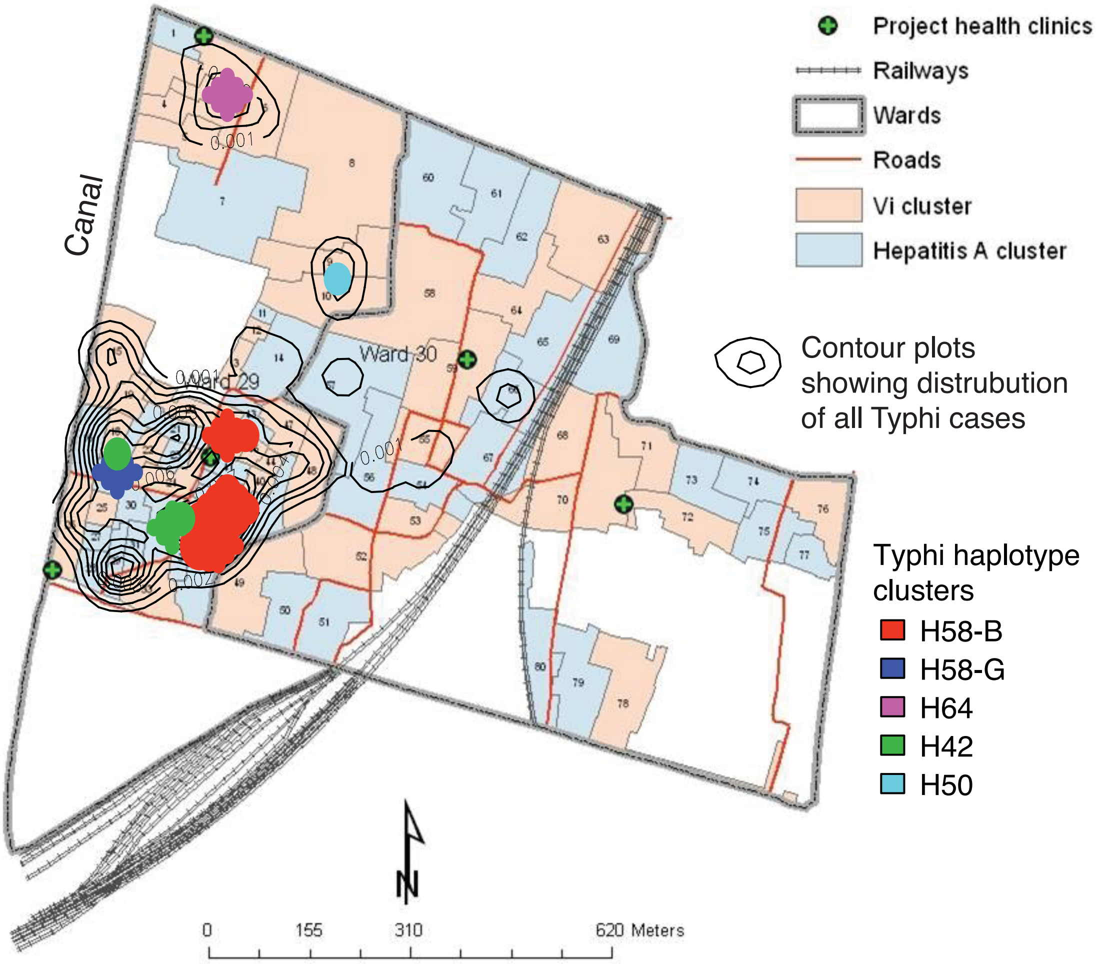


Figure 6.22: Spatial clustering of typhoid cases in Kolkata - Geographic clusters 1-80 are shown as orange and blue blocks. Countour plots (black lines) show the distribution of typhoid fever cases during the study period. Geographic clusters with an overrepresentation of typhoid cases of particular haplotypes (using Openshaw's Geographical Analysis Machine to test for spatial clustering within each haplotype) are coloured according to the legend.

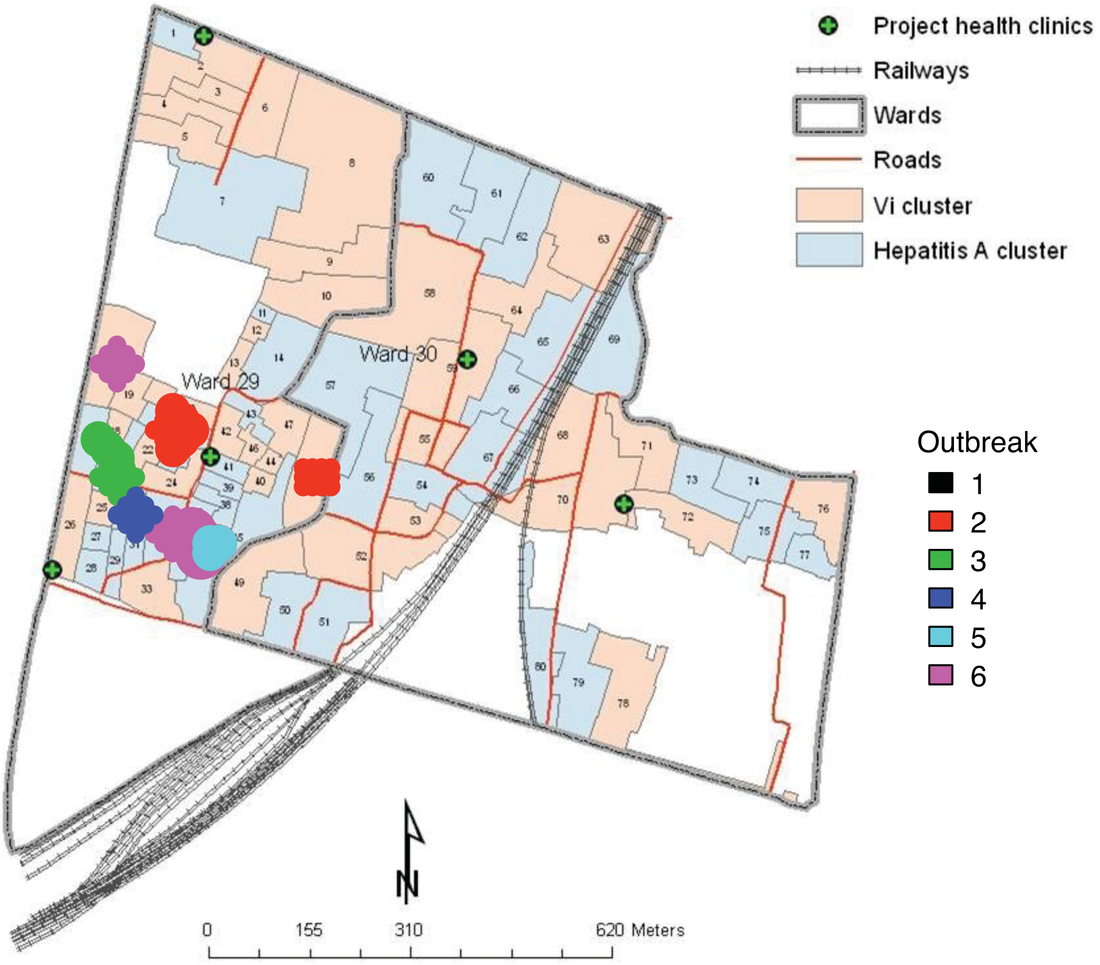


Figure 6.23: Spatial clustering during typhoid peak-incidence periods in Kolkata - Geographic clusters 1-80 are shown as orange and blue blocks. Geographic clusters with an overrepresentation of typhoid cases during six periods of high incidence (labelled 1-6 as in Figure 6.21) (using Openshaw's Geographical Analysis Machine to test for spatial clustering within each period) are coloured according to the legend.

There was evidence of typhoid fever cases clustering within househoulds, with 88 (47%) of the SNP-typed Typhi isolated from patients living in 34 (0.3%) of the 10,954 households in the study site (2-6 per household, median 2, see Figure 6.24a). Nineteen of these households had cases of infection with multiple Typhi haplotypes, see Figure 6.24a. The median time between cases in the same household with the same haplotype was 93 days (3 months; range 1-722 days), while the median time between any cases within a household was 154 days (5 months; range 1-1078 days). The downwards skew of time between infections in the same household (see Figure 6.24b) suggests these infections are not independent. Infections within a household may be caused by transmission between individuals (i.e. same haplotype), or multiple people sharing infections from a common external source (which could be of the same or different haplotypes). The distribution of times between infections with the same haplotype was not significantly different from those between infections with different haplotypes (p=0.3, one-sided Kolmogorov-Smirnov test). The high number of households reporting infections with multiple Typhi haplotypes (Figure 6.24) suggests that shared exposure to external sources of contaminated food or water may make an equal if not greater contribution to shared typhoid fever illness than direct transmission of Typhi between household members, at least in high-risk endemic areas.

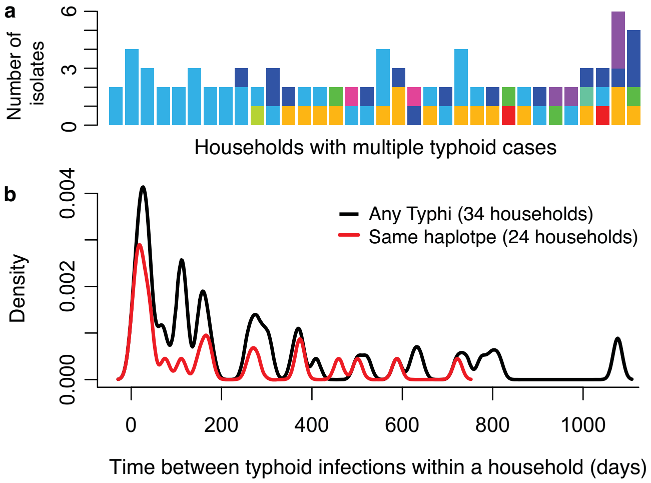


Figure 6.24: Distribution of Typhi cases among households in Kolkata - (a) Typhi diversity among households with multiple SNP-typed Typhi infections. Colours represent distinct Typhi haplotypes. (b) Distribution of time between Typhi cases within a household.

6.3.5.3 Association with the vaccination programme

The Vi conjugate vaccine was effective at reducing the incidence of typhoid fever among vaccinees, with overall effectiveness reported at 61% (421). Among the post-January 2005 cases for which isolates were SNP typed, the odds ratio (OR) for typhoid among Vi vaccinees compared to unvaccinated inhabitants of clusters assigned the Vi vaccine was 0.55 (95% confidence interval 0.28-1.08, p=0.04 (758, 759)). This is equivalent to vaccine effectiveness of (1-0.55) x 100% = 45%. The diversity of Typhi haplotypes causing disease appeared to have been reduced after the introduction of the vaccine, with a Simpson's diversity index (1-D) of 0.81 prior to December 2004 and 0.68 after January 2005 (p-value<1x10^{-6}, see 6.2.6; see also Figures 6.21b and 6.25). The proportion of Typhi isolates of the H58-B haplotype increased after the vaccinations, from 33% to 50%. In addition, the vaccine appeared to be ineffective against H58-B, with an OR of 1.56 (95% CI 0.44-7.16, p=0.27) for H58-B infection among vaccinees compared to unvaccinated inhabitants, in contrast to OR 0.33 (95% CI 0.14-0.77, p=0.005) for infection with other Typhi haplotypes. All isolates expressed Vi (421) and all of the H58-B isolates gave positive signals for all SNP loci targeted in SPI7, including seven SNPs within the *tviD* and *tviE* genes involved in Vi biosynthesis. It is possible that the H58-B isolates may have a modified Vi structure or a modified pattern of Vi expression which facilitates their escape from immunity conferred by the Vi vaccine, however further experiments will be needed to test these hypotheses.

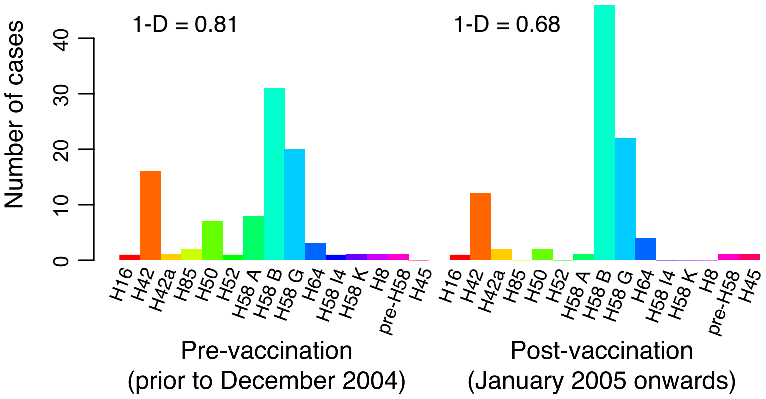


Figure 6.25: Frequency distribution of Typhi haplotypes before and after the introduction of a Vi conjugate vaccine in Kolkata - 1-D = Simpson's index.

6.3.6 The Typhi population in Nairobi, Kenya over a 21 year period

The incidence of typhoid fever in Kenya has not been studied, however the annual incidence of typhoid fever in Eastern Africa has been estimated at 39/100,000, and 50/100,000 for Africa as a whole (10) (see Figure 6.1). Diagnostic tests and surveillance systems for *Salmonella* are not as well developed in Kenya as they are in high-incidence endemic areas in South Asia (760, 761, 762) and as a result there are few studies addressing the population structure or drug resistance of Typhi. However available studies reveal a high rate of MDR typhoid, associated with the presence of a >100 kbp plasmid (383, 763). A collection of 96 Typhi isolates was assembled by Sam Kariuki at KEMRI from hospitals in Nairobi, Kenya (see Appendix E). The isolates were collected between 1988 and 2008 for a number of surveillance studies. The collection is biased towards more recent isolates, with 73 isolates (76%) collected between 2001-2008. DNA was provided by Sam Kariuki for SNP typing on the GoldenGate array. DNA quantitation was performed by Derek Pickard at the Sanger Institute, analysis was performed by myself.

Figure 6.26 shows the distribution of Typhi haplotypes over the 21 year period represented by the Nairobi collection. The majority of isolates (76%) were H58, with the remaining isolates distributed among seven distinct haplotypes (see Figure 6.26a). The H58 isolates increased in frequency over time, but at least one H58 strain was isolated in every year represented, including the earliest year 1988. The distribution of other haplotypes was more sporadic, with most detected in just one or two years (see Figure 6.26b). The distribution of H58 subgroups is shown in Figure 6.26c-d. The most common H58 subgroup was J1, represented by 47 isolates (64% of H58 isolates) spread across the entire period (see Figure 6.26d). Two isolates of H58-J2, derived from H58-J1 (see Figure 6.26c) were detected among those from 2001. H58-B was detected among isolates between 2004-2008. The H58-B subgroup is from a different sublineage to H58-J1 and H58-J2 (see Figure 6.26c), thus the appearance of H58-B in 2004 may represent the introduction and spread of a novel clone in this region.

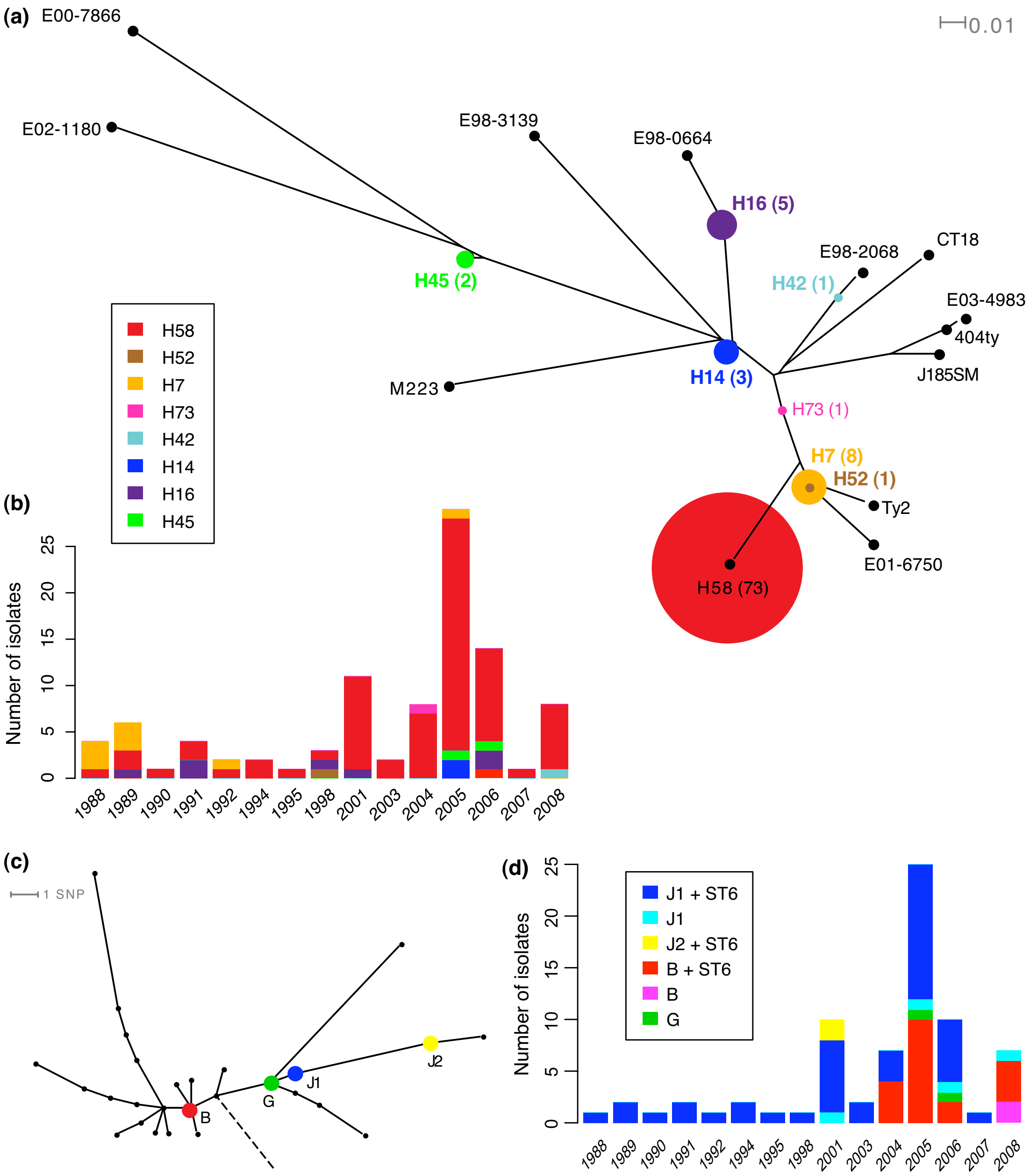


Figure 6.26: Distribution of Typhi haplotypes in Nairobi, Kenya - (a) Phylogenetic tree of Typhi, the position of Kenyan isolates is shown with coloured nodes, labelled by haplotype and the number of isolates (in brackets). The size of nodes are also scaled to represent the number of isolates. (b) Distribution of haplotypes over time, colours indicate haplotypes as shown in the legend and in (a). (c) Phylogenetic tree of Typhi H58, the position of Kenyan isolates is shown with coloured nodes, labelled by H58 subgroup. (d) Distribution of H58 subgroups over time, colours indicate subgroup and presence of the ST6 IncHI1 plasmid, as shown in the legend and in (c).

A total of 66 IncHI1 plasmids were detected, of which 65 were ST6 plasmids found in H58 strains. The exception was a single plasmid, similar to the ST6 plasmids but with different alleles at 18 SNP loci and no signal at 39 loci, found in an isolate of the H73 haplotype collected in 2004 (pink in Figure 6.26a). Among the H58 isolates, ST6 plasmids were found in the majority of J1 and B isolates (96% and 91%, respectively) and in both J2 isolates (see Figure 6.26d). The two H58-G isolates, collected in 2005 and 2006, were plasmid-free. The plasmid-free B and J1 isolates were observed in later years (see Figure 6.26d), consistent with a gradual loss of the MDR plasmid over time.

6.3.7 Typhi H58 and the IncHI1 ST6 plasmid

Figure 6.27 shows the distribution of Typhi H58 subtypes among the regional datasets analysed above. As those analyses showed, the distribution of H58 subtypes was different in the different regions studied. However Figure 6.27 highlights some broader trends in the geographic distribution of H58. In particular node C and derived lineages (D, E, F) was mainly restricted to Vietnam, while node G and derived lineages (I, J, K) were common among isolates from India but not Vietnam. The Kenyan isolates were mostly from the JI and B nodes, which were associated with Indian but not Vietnamese isolates (B was common in Kolkata; the J nodes are derived from the G node which was also common in Kolkata). This may reflect human patterns of travel and migration between India and Kenya, which has had an Indian community since the early 20th century, generating a constant flow of migrants and visitors between the two countries. The isolates found in Kathmandu were from the G node, which was also common in Kolkata, perhaps reflecting the close geographic proximity of India and Nepal.

The earliest H58 and derived haplotypes were first reported by Roumagnac *et al.* (2), including a 1958 H61 isolate from Morocco, a 1966 H62 isolate from Morocco and a 1968 H60 isolate from the Cote d'Ivoire (see Figure 6.27). The 1964 H58 isolate 43-64 from Chad was retyped in the present study and found to be H58-E2, indicating that diversification of the H58 B lineage had already occurred by the early 1960s. The earliest representative of the G lineage was among the earliest isolates from the Kenya collection (1988). Thus the earliest record we have of H58, including diverse sublineages, is from Africa. This is consistent with an African origin for H58, however there

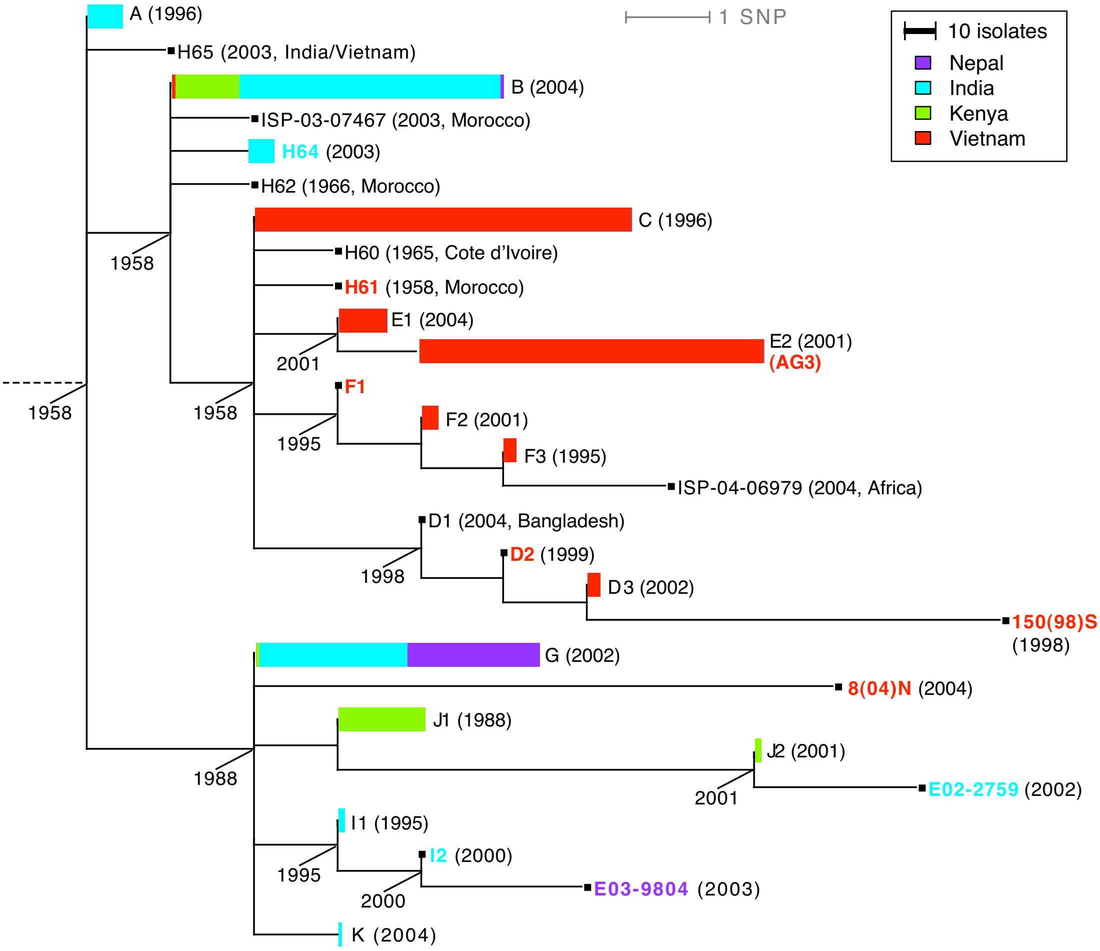


Figure 6.27: Distribution of H58 subtypes among Typhi isolates from four regional collections - The phylogenetic tree of H58, dashed line represents the root of this tree, i.e. where H58 joins the rest of the Typhi phylogeny. Nodes are labelled with their name as assigned in this study (A-K) or previously assigned H-group, or with the sequenced isolate which defines the node. Coloured bars indicate the number of isolates at that node from each of the four regional collections. Nodes that are not represented by isolates from these collections (i.e. defined by isolates from the Pasteur collection, (2) or sequenced isolates) are coloured to indicate their origin from Vietnam, India, Nepal or Kenya, or are labelled to indicate their origin elsewhere. Nodes are also labelled with the earliest year of isolation (among the isolates typed in this study or in (2)).

is not enough data available from early Asian isolates to confirm this. Only ten pre-1995 Asian isolates have been tested (sourced from the Pasteur collection and tested both here and in (2)), all of which were from Vietnam and belonged to the non-H58 associated haplotypes H1 (N=3), H50 (N=4), H87 (N=2) and H68 (N=1). The earliest appearance of H58 among Vietnamese isolates in the Pasteur collection was 1995, with an isolate from the H58 F lineage (reported in (2) and analysed in this study). The earliest isolate of the H58 D lineage was from 1999 and the earliest H58 E lineage isolate from Vietnam was from 2001. The earliest example of the J lineage was a 1988 isolate from the Kenyan collection, and the earliest example of the I lineage was an Indian 1995 isolate from the Pasteur collection. The non-random distribution of H58 subtypes shown in Figure 6.27 is in contrast to the global distribution of Typhi haplotypes first noted in (2). This is likely a function of the different time scales and levels of resolution involved - the diversification of H58 likely represents well under 100 years of evolution, with much of this occurring during an expansion over the last 10-20 years; the broader scale comparison of H-groups represents diversification over many thousands of years. Thus while it is broadly true that Typhi has a global distribution, high-resolution SNP typing highlights some geographic patterns. This suggests there are barriers to the spread and maintenance of novel Typhi clones, even between high-incidence endemic areas that are relatively close geographically.

In the data presented above, a clear trend emerged in which MDR IncHI1 plasmids detected in the last ten years have been almost exclusively of the ST6 type (see summary plot in Figure 6.28c) and have been found in H58 isolates. While the data is biased towards the four locations studied intensively (Mekong Delta, Kathmandu, Kolkata and Kenya), the Pasteur collection revealed Typhi H58/ST6 plasmid isolates originating from numerous locations in Asia, Africa and the Middle East (see Figure 6.28a), suggesting that this MDR clone has global reach. Figure 6.30 shows the distribution of IncHI1 plasmids within H58 subgroups, as well as the distribution of *IS*1. This IS element is encoded in the IncHI1 plasmid (as part of *Tn*9, *Tn*6062 and singleton insertions) and in the sequencing study presented in Chapter 2 was also found inserted in the chromosomes of IncHI1 plasmid-containing isolates (see 2.3.3). The association between plasmid and *IS*1 is further supported by the GoldenGate data, with *IS*1 detected far more frequently within haplotypes in which plasmids had also been detected. The

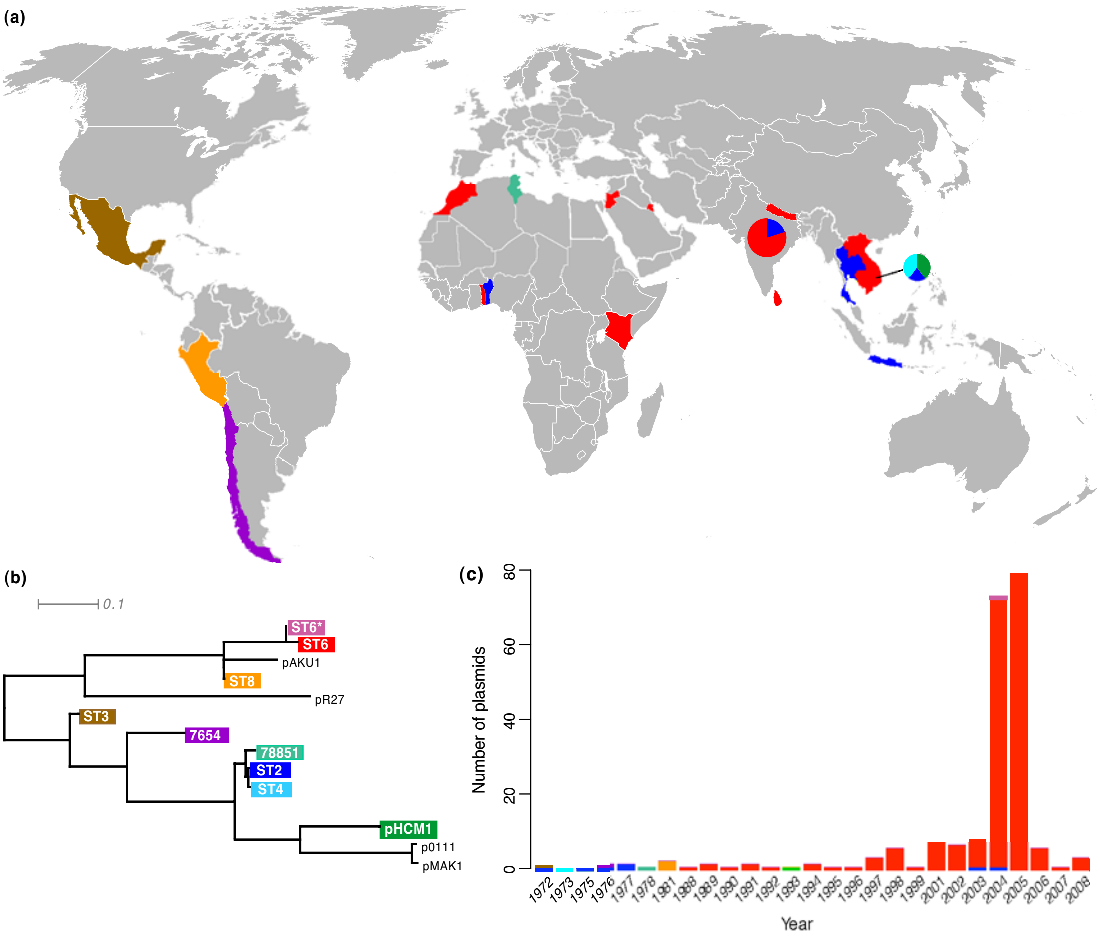


Figure 6.28: Distribution of IncHI1 plasmids in time and space - Each colour indicates a unique IncHI1 subtype, as labelled in (b). (a) Countries from which IncHI1 plasmids were analysed, coloured to indicate the IncHI1 subtype detected. (b) Phylogenetic tree of IncHI1 plasmids detected in this study, scale bar indicates divergence among 200 SNPs. (c) Distribution of IncHI1 plasmid types over time.

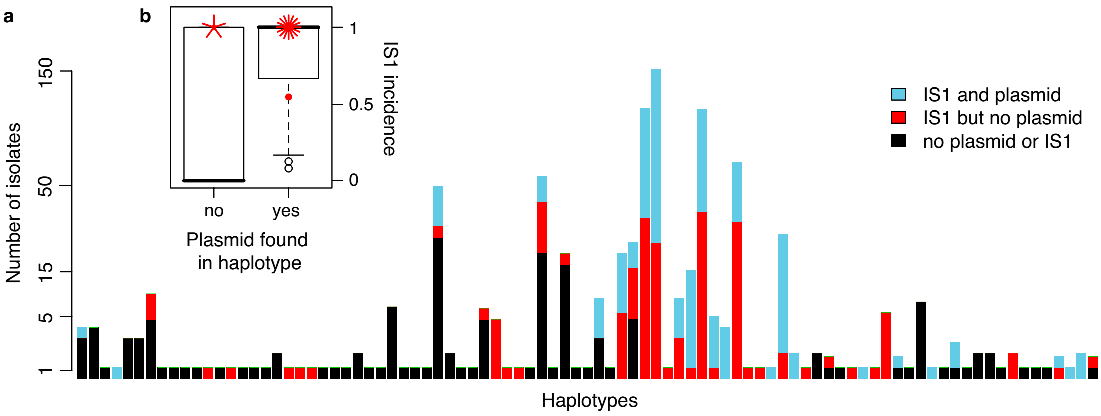


Figure 6.29: Distribution of IncHI1 plasmids and *IS*1 among Typhi haplotypes - (a) Presence of *IS*1 and IncHI1 plasmid among Typhi haplotypes. All plasmid-containing isolates also carried *IS*1. Note the y-axis is on the log scale. (b) Incidence rate of *IS*1 among haplotypes in which the presence of IncHI1 plasmids was demonstrated. Data from H58 haplotypes are highlighted using a sunflower plot, in which each radial line represents a different data point (i.e. haplotype) sharing the value represented by the central point.

*IS*1 probe gave positive signals for every isolate in which the IncHI1 plasmid was found (230 isolates), demonstrating its reliability in detecting the presence of *IS*1. It was detected in only 21 out of 61 haplotypes in which plasmids were not detected, which could be the result of plasmid loss after transposition of *IS*1 into the chromosome. Figure 6.29 shows the distribution of plasmids and *IS*1 elements in all haplotypes, as detected by GoldenGate assay and Figure 6.30 shows the distribution within the H58 phylogenetic tree. As the latter shows, the ST6 IncHI1 plasmid was detected in all lineages of the H58 phylogenetic tree, although many isolates did not have the plasmid. However, apart from those in the ancestral node A, all H58 isolates gave positive signals for two *IS*1 loci targeted in the GoldenGate assay. This is consistent with a single acquisition of the ST6 plasmid by the common ancestor of H58, perhaps belonging to node A itself, followed by transposition of *IS*1 from the plasmid into the chromosome and gradual loss of the plasmid from sublineages following the diversification of H58. However, given that H60 and H61 were present as early as 1965 and 1958, respectively (see Figure 6.27) this would imply that the ST6 plasmid had already been acquired by this time, which seems unlikely. It is not impossible, given that chloramphenicol resistant typhoid was described as early as 1950 (363), but would imply that further resistance genes accumulated in ST6 while it was resident in H58 Typhi.

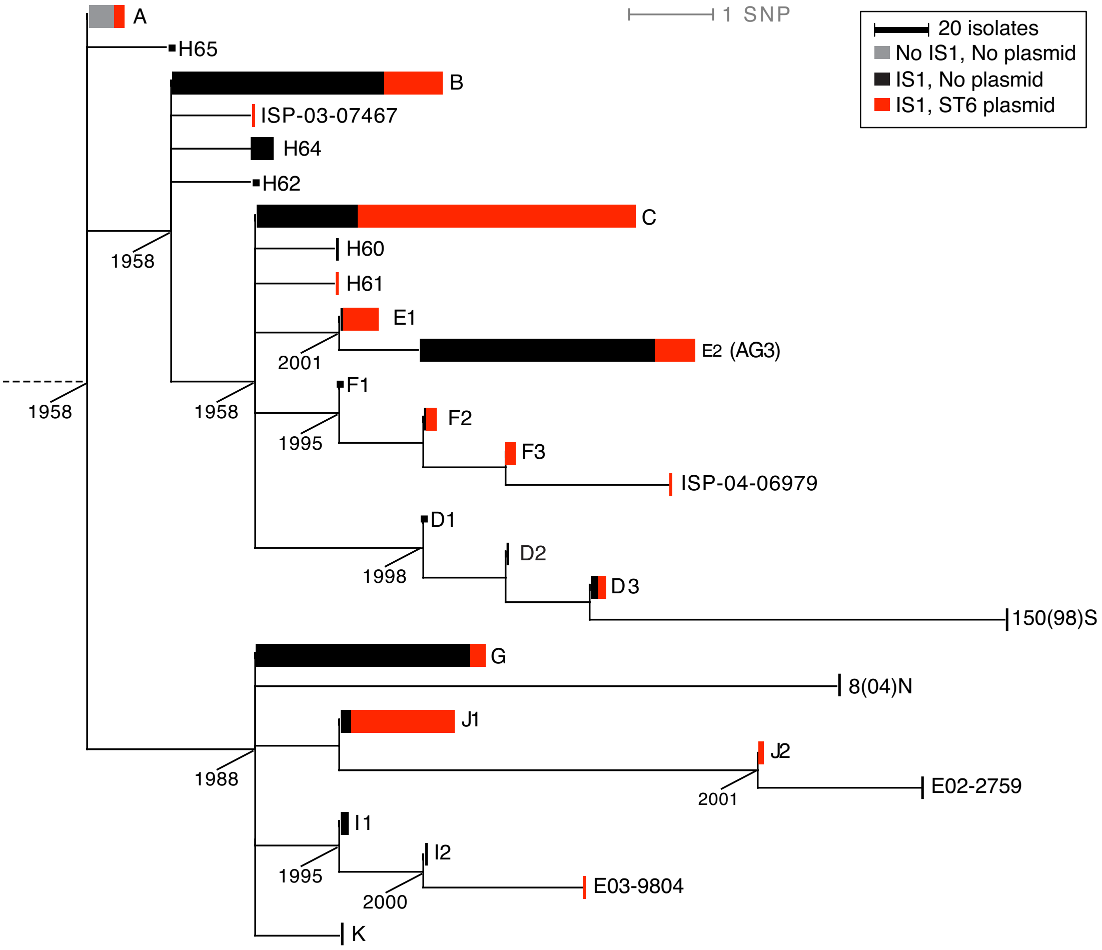


Figure 6.30: Distribution of IncHI1 plasmids and *IS*1 among Typhi H58 subtypes - The phylogenetic tree of H58, dashed line represents the root of this tree, i.e. where H58 joins the rest of the Typhi phylogeny. Nodes are labelled with their name as assigned in this study (A-K) or previously assigned H-group, or with the sequenced isolate which defines the node. Coloured bars indicate the number of isolates at that node (from the regional collections or the Pasteur collection) that contain *IS*1 and/or the IncHI1 plasmid ST6 as indicated in legend. Base nodes are labelled with the earliest year of isolation (among the isolates typed in this study or in (2)).

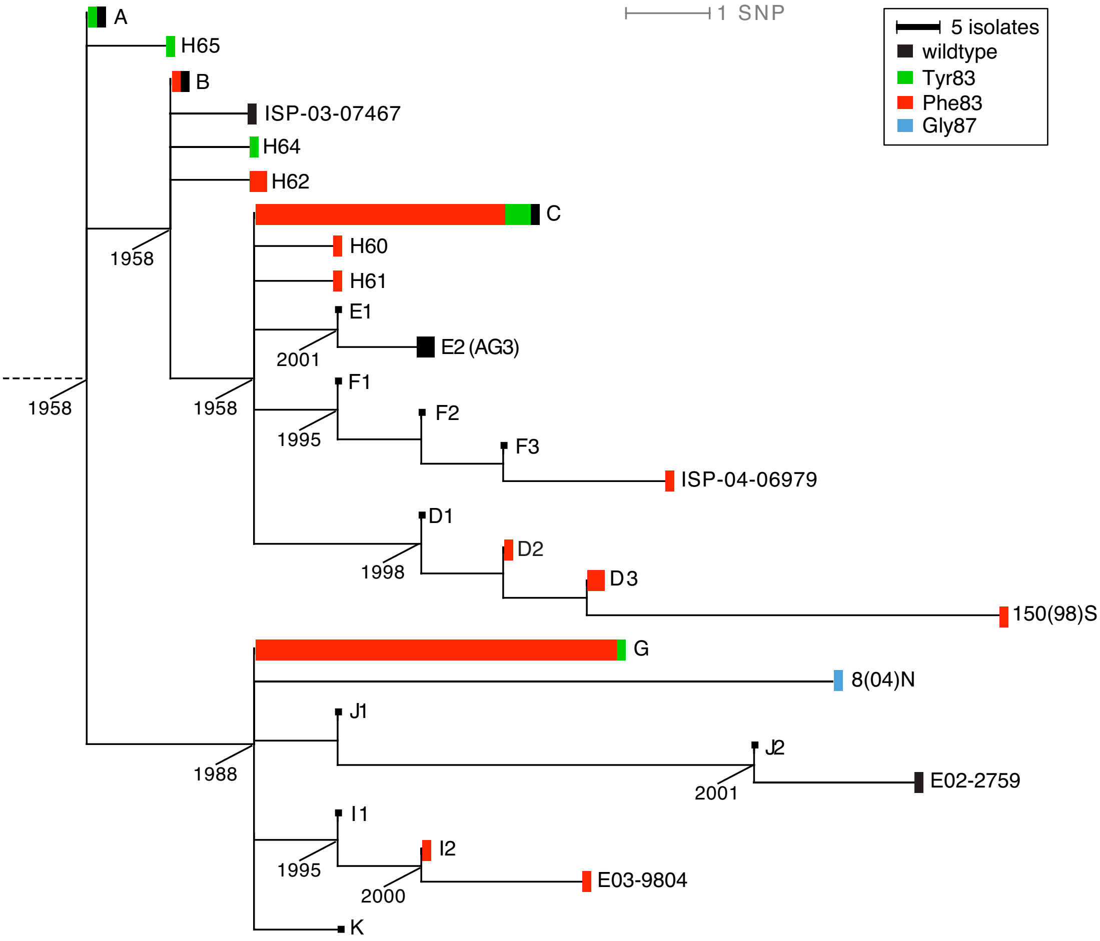


Figure 6.31: GyrA SNPs distributed among Typhi H58 subtypes - The phylogenetic tree of H58, dashed line represents the root of this tree, i.e. where H58 joins the rest of the Typhi phylogeny. Nodes are labelled with their name as assigned in this study (A-K) or previously assigned H-group, or with the sequenced isolate which defines the node. Coloured bars indicate the number of isolates at that node that have been SNP typed at GyrA positions 83 and 87 (total of 97 H58 isolates). Base nodes are labelled with the earliest year of isolation (among the isolates typed in this study or in (2)).

6.3.8 Typhi H58 and mutations in GyrA

A total of 246 isolates (from the Pasteur collection (2), the Kathmandu collection above and the sequenced isolates) have been typed for SNPs in *gyrA* that are known to be associated with fluoroquinolone resistance. Three isolates had a SNP at position 87: Asn87 in a H6 isolate from Morocco in 2005 (Pasteur collection), and Gly87 in sequenced isolates E02-1180 (H45, from India, 2002) and 8(04)N (H58, from Vietnam, 2004). Of the remaining isolates, 97 were from the H58 cluster, of which 83 carried the Phe83 SNP and five carried the Tyr83 SNP. The distribution of these SNPs among H58 subtypes is shown in Figure 6.31, and supports the notion that mutations arose at GyrA position 83 not once but multiple times within the H58 cluster (2). The sequenced H50 isolate E98-3139 also carried the Phe83 SNP.

6.4 Discussion

6.4.1 Strengths and limitations of the study

The major strengths of the GoldenGate high throughput SNP typing assay for the study of Typhi populations are (a) the reproducibility and phylogenetic informativeness of the sequence-based approach, (b) the increased resolution offered over other sequence-based approaches that have been attempted previously (MLST (1, 725) and SNP typing at <100 loci (2, 256)), (c) the ability to simultaneously target hundreds of loci on both chromosome and plasmid, and (d) the ability to screen hundreds of isolates in a single assay. The increased resolution, particularly within the H58 cluster, has revealed dynamics of the Typhi population at a scale that was largely inaccessible until now. For example, while it has been observed previously that H58 is common in South Asia (2, 570) and Typhi lineages are globally distributed (2), high resolution SNP typing in this study revealed fluctuations within the H58 populations of endemic regions over time (see Figures 6.16, 6.19 and 6.21) as well as geographic differences in population structure (see Figure 6.22). These observations were only possible using high resolution analysis of large numbers of isolates and sequence-based technologies that facilitated direct comparison between data sets. By linking Typhi strain types with IncHI1 plasmid types for the first time, this study provided direct evidence of multiple independent acquisitions of distinct IncHI1 plasmids by distinct strain types, spread of IncHI1 plasmids within specific Typhi strain types, and the presence of very closely related IncHI1 plasmids in Typhi and Paratyphi A (see Figure 6.28). It also highlighted the close association between H58 Typhi and ST6 plasmids in recent years (see Figures 6.11 and 6.30) and provided evidence for multiple independent occurrences of the same fluoroquinolone-resistance mutations within closely related H58 strains (see Figure 6.31).

The GoldenGate assay has some major limitations, including the inability to accurately type a quarter of known SNP loci, and in particular the inability to target SNPs that are close to other mutations, including the known fluoroquinolone resistance-associated SNPs in *gyrA* and other topoisomerase target genes. Thus to study the acquisition of Nal resistance within the Typhi population would require additional assays to screen for mutations in these genes, such as the Luminex assay used to type *gyrA* SNPs within the Nal-resistant isolates from Kathmandu (6.3.4.2) or the Sequenom SNP typing assay (not yet demonstrated for Nal-resistance Typhi SNPs). However, all SNP typing approaches suffer the much more general problem of being limited to known SNP loci. This limitation is acceptable for many purposes, for example the present study as well as other SNP-typing studies (2, 256) have yielded novel insights into the evolution and population structure of Typhi. However SNP typing by its nature cannot discover novel mutations, thus there will always be a limitation to the resolution possible with this technique. Since H58 was known to be the most prevalent haplotype (2), care was taken in this study to screen for SNPs that would discriminate within the H58 cluster (seven out of nineteen sequenced isolates were chosen from within this cluster, Chapter 2) and indeed 45 SNPs were successfully typed. Yet it is almost certain that there is much more variation within the larger nodes of the H58 tree than we were able to detect. This is most clearly illustrated in Figure 6.27, which shows that the vast majority of H58 isolates tested fell into internal nodes (B, C, G). In contrast the leaf nodes were rarely found, except for E2 which contained nearly half of the isolates from the Mekong Delta study and was defined by SNPs detected in an isolate taken directly from the study population for sequencing (AG3 from An Giang, 2004). On the other hand, within the rest of the Typhi phylogeny there was a lot of redundancy, since the Typhi isolates tested clustered to some extent into clonal groups as opposed to a continuum of haplotypes (see Figures 6.8 and 6.9). Although this may be unsurprising, it could not have been known conclusively before the present study, and even now it is not entirely clear which SNPs along the longer internal branches of the phylogenetic tree would be the most informative for a lower-resolution set of target loci. Finally, it is difficult to interpret a lack of signal from the GoldenGate assay. For plasmid and resistance gene loci, a lack of signal was interpreted in this study as absence of the plasmid or target sequence from the isolate. This is reasonable as lack of signal at IncHI1 loci was strongly correlated between IncHI1 loci and with lack of MDR phenotype (see 6.10). However it is difficult to interpret lack of signal at chromosomal loci, which may be due to deletion of chromosomal sequence or simply mutations within oligonucleotide binding sites. This should not affect phylogenetic inference though, as a lack of signal was represented in the allele alignments as the gap character '-', which does not contribute to phylogenetic inference using maximum likelihood methods.

The collections of Typhi isolates that were SNP typed in this study have been collected by different research groups over different time periods, for different purposes and using different sampling approaches (see Table 6.1 and the introduction to this chapter 6.1). As such the Typhi populations under study are not directly comparable, although each of the regional collections gives a snapshot of the Typhi population in a defined time and place. In order to directly compare Typhi populations across time and space, a more consistent sampling approach would need to be used, and ideally would include a strategy to collect isolates from asymptomatic carriers as well as typhoid fever patients. The availability of spatial data associated with the Kolkata study provided the opportunity to examine spatial clustering of typhoid cases and Typhi haplotypes. The detection of haplotype-specific spatial clusters in this study suggests that spatial data will be useful in future studies of Typhi populations and potentially genomic epidemiology of typhoid; ideally high-resolution data would be collected using geographical positioning systems (GPS) which are increasingly cheap and easy to implement. The population-based sampling approach implemented in the Kolkata study should provide a fairly complete view of typhoid fever within the study site, although there still may be issues with residents reluctant to report illness to the health services involved in the study. In particular, clustering within households may be confounded by household preferences to report or not report febrile episodes. The analysis of Typhi haplotypes presented above from the Kolkata study is currently underpowered, as only half of available isolates were typed. This was due to problems with obtaining adequate yields from DNA extractions at the NICED laboratories in Kolkata, but repeat extractions will be performed to allow all isolates to be SNP typed and provide a complete set of SNP data for this collection. This will provide additional power to detect haplotype-specific differences and may clarify the apparent difference in vaccine efficacy on the Typhi H58-B haplotype. The Nepal study is much smaller and more specific than the Mekong Delta and Kolkata studies, focussing only on pediatric cases of typhoid fever requiring hospitalisation at Patan Hospital. This should be considered the 'tip of the iceberg' of typhoid fever in Kathmandu, as the majority of enteric fever patients presenting at the hospital are treated as out-patients (279) and pediatric enteric fever is rarely reported in Kathmandu (279, 764). The studies of local populations presented here provide proof-of-principle that high-resolution SNP typing, including maximum resolution in H58 and inclusion of IncHI1 plasmid SNPs, is informative for studying Typhi populations, which have been relatively impenetrable to low-resolution typing methods that have achieved success in studying other more diverse bacterial populations. These studies also demonstrate that there is substructure within the H58 population, which can be used to distinguish clones with distinct spatio-temporal distributions.

6.4.2 Typhi populations in endemic areas

This study provides high-resolution data on the structures of Typhi populations within high- and medium-incidence endemic areas. The only such study published to date used the Sequenom platform to target 88 SNPs within 140 Typhi isolates from Jakarta, Indonesia, which is also a high-incidence typhoid endemic area (256) (note I was involved in the analysis and am a co-author on this study). This revealed the presence of nine distinct Typhi haplotypes co-circulating within Jakarta, and geo-positioning data for 54 (39%) of the isolates revealed some spatial clustering within the city. However in Indonesia, typhoid fever follows quite different patterns to those observed in mainland Asia, including a lack of drug resistance (16, 327, 765, 766), prevalence of z66 antigen and haplotype H59 (2, 248, 251, 252, 253, 256, 767) and lack of prevalence of H58 (2, 256).

The present study focused on three high-incidence typhoid endemic regions in Asia and one medium-incidence endemic region in Africa (see Figure 6.1). Although the isolate collections from each region span different time periods and were collected via different sampling methods (Table 6.1), some clear similarities emerge from the SNP typing data. In each region multiple Typhi haplotypes were co-circulating, with H58 by far the most common group in each case (70-98%), see Table 6.5. Some of the haplotypes were shared between regions, while others appeared to be geographically restricted. The clusters of H42 isolates making up the second most frequent groups in Kolkata, Kathmandu and Kenya (labelled H42-A) were identical at all target loci. While H58 was dominant in each study site, the structure of the H58 populations varied and provided the strongest examples of geographical restriction (see Figure 6.27). There was usually a single dominant subtype of H58, which differed in each region (see Table 6.5), with the exception of the Mekong Delta isolates which were fairly evenly split between the closely related subtypes C and E2. However there was clearly variation from year to year, with different haplotypes appearing and disappearing over the course of the studies. This was particularly evident among isolates from Nairobi, which cover the longest time period (21 years). In the late 1980s to early 1990s, H58 made up less than 50% of isolates, although sample sizes were low (see Figure 6.26b). There is a gap in the sampling, but isolate numbers increased from 2001 and from this point H58 was clearly dominant (Figure 6.26b). In Nairobi H58-J1 was the only H58 subtype detected between 1988 and 1998 and the J1-derived J2 node was detected only in 2001 (see Figure 6.26d). These were the only H58 subtypes detected until the appearance of H58-B in 2004, which co-circulated with J1 for a few more years and had all but replaced it by 2008 (see Figure 6.26d).

Patterns of drug resistance varied markedly between the regions studied (see Table 6.5). In Kenya and Vietnam, there were high rates of MDR, associated with IncHI1 ST6 plasmids in H58, and the presence of the plasmids appeared to contribute to the success of the dominant Typhi haplotypes. In Vietnam the H58-E2 subtype, which was generally plasmid free, was replaced in 2005 by the H58-C subtype which had a high rate of the MDR plasmid (see Figures 6.14 and 6.16b-c). In Kenya nearly all H58 isolates carried the MDR plasmid and only two cases of the (plasmid-free) H58-G subtype were detected, possibly due to lack of the plasmid. In Nepal and Kolkata,

	Mekong Delta	**Kathmandu**	**Kolkata**	**Nairobi**
H58	98%	70%	74%	76%
Dominant H58	C (45%)	G (94%)	B (55%)	J1 (64%)
subtype(s)	E2 (41%)	-	G (31%)	B (30%)
MDR (overall)	58%	0	3%	69%
MDR (H58)	59%	0	4%	89%
Nal-R (overall)	96%	63%	nd	nd
Nal-R (H58)	98%	91%	nd	nd

Table 6.5: Summary of Typhi populations from endemic areas - MDR = multidrug resistant, Nal-R = nalidixic acid resistant.

however, MDR was very rare (0% - 3%, see Table 6.5). Nalidixic acid resistance was studied only in the collections from Nepal and Vietnam, and high rates were found in both locations. In Nepal, 91% of H58 isolates were Nal resistant and 97% carried the Phe83 mutation in *gyrA*. The dominant H58-G isolates, all of which carried this mutation, occurred at twice the rate of other, Nal-sensitive haplotypes. In Vietnam, Nal resistance was only observed among H58 isolates, at a rate of 98%.

The peaks observed in these studies are considered to be random (or seasonal) fluctuations in an endemic area, as opposed to outbreak events. In the Mekong Delta, Kathmandu and Kolkata studies, typhoid fever cases showed varying degrees of seasonality (see Figures 6.16, 6.19 and 6.21). In areas that experience a wet season, peaks in typhoid fever and other water-borne bacterial disease is usually associated with onset of the wet season (343, 354, 768, 769, 770). It has been suggested that such peaks tend to occur just before the beginning of the wet season, when water scarcity leads to difficulties in sourcing drinking water and compromised hygiene practices (354). Flooding, too, may be associated with water-borne disease as it becomes hard to keep drinking water and sewerage systems separate (343, 771). In countries with a dry climate, seasonal peaks in typhoid incidence are usually associated with high-temperature (772, 773). The seasonality of typhoid fever and other enteric bacterial disease in the Mekong Delta has been studied over much longer time scales, which revealed weak correlations between high rainfall and periods of high-incidence of typhoid fever, peaking between April and July (354). This is broadly consistent with the present study, in which typhoid fever peaked between March-July in 2004 and May-June in 2005 (see

Figure 6.16). Rainfall data was not available, but this timing corresponds to the usual onset of the wet season (April-November).

In the Kolkata study, similar seasonal patterns were visible, although less pronounced. The first major peak (1 in Figure 6.21) corresponded with the onset of the wet season, a pattern which was also evident in 2006 (5, 6 in Figure 6.21). However other peaks occurred at the beginning of the dry season, or during the wet season (2-4 in Figure 6.21), consistent with previous reports of a lack of seasonality to typhoid outbreaks in India (774). It is interesting to note that greater seasonality was observed in the rural setting of the Mekong Delta than the urban slum setting of Kolkata. This may be due to climatic differences between the two regions, but could also be associated with the different pressures on the water supply. In the Mekong Delta, water scarcity at the end of the dry season is a very real problem, as nearly half the population relies on rivers, streams and ponds for their drinking water and over two-thirds of the population have their toilet facilities directly over water (349). In the Kolkata slums drinking water is taken from public taps, but water pipes and sewerage pipes lie close together, so although contamination is more likely to occur as a result of flooding during the wet season, it is a risk all year round (343).

No seasonality was observed in the Kathmandu study (see Figure 6.19), although this may be due to the low number of cases included in the study, which were limited to children under 12 hospitalised at Patan Hospital with blood-culture confirmed typhoid fever. The majority of enteric fever patients presenting at Patan Hospital are treated as out-patients (279), and pediatric enteric fever is rarely reported in Kathmandu (279, 764) so the present study is unlikely to detect seasonal patterns in typhoid incidence that may exist in Kathmandu.

6.4.3 The evolution of drug resistance in Typhi

This study is the first to link IncHI1 plasmid variation with Typhi strain variation in a phylogenetically informative way. Most studies have been unable to detect differences between plasmids in distinct strain backgrounds (277, 735) or have addressed only plasmid differences (275) or strain differences (276) but not both. One study reported two distinct PFGE patterns for 200 kbp plasmids from strains that had closely related but distinct chromosomal PFGE patterns, and a third PFGE pattern for a 200 kbp plasmid in an unrelated strain type (736). This is consistent with either co-diversification of plasmid and strain type or independent acquisition of distinct plasmid types by distinct strain types, but these scenarios could not be differentiated due to the lack of phylogenetic informativeness of PFGE. By simultaneously targeting chromosomal and IncHI1 plasmid SNP loci, the GoldenGate assay enabled each IncHI1 and host strain to be positioned on phylogenetic trees, providing direct evidence of independent plasmid acquisition events (Figure 6.11). The plasmid-strain associations revealed instances of the same plasmid being present within distantly related strains (e.g. ST2, ST8, see Figure 6.11) which confirms multiple acquisition of highly similar plasmids by distinct Typhi lineages. In some cases these coincided in time and space, for example ST8 plasmids were found in distinct Typhi strains isolated from Peru in 1981, which although circumstantial does suggest direct transfer between Typhi cells or at least transfer from a single third-party bacteria into multiple Typhi cells at around the same time and place.

The most striking plasmid-strain association identified in this study was that between the ST6 IncHI1 plasmid and H58 Typhi. Although there were hundreds of examples of H58 isolates without plasmids, in particular in Nepal (100% plasmid-free) and Kolkata (96% plasmid-free), the presence of *IS*1 in all but the most ancestral node of the H58 phylogeny (see Figure 6.30) suggests that the plasmid was acquired early in the life of H58 and subsequently lost. The loss of the plasmid is presumably associated with reduced selective pressure following the switch to fluoroquinolones for the treatment of enteric fever. However the maintenance of the plasmid in the Mekong Delta (59% of H58 isolates) suggests it may still provide some selective advantage, at least in this region. The dominance of the plasmid-containing H58-C isolates in the second year of the study (see Figure 6.16) may be associated with such an advantage. Although inappropriate antibiotic prescribing and dispensing practices are common in Vietnam (775) and high rates of drug resistance have been reported among bacterial contaminants in the human food chain (776, 777), water systems in the Mekong Delta do not contain unusually high levels of antibiotics (778) and a survey of *Salmonella* isolates originating from food, animals and human diarrhea patients in the Mekong Delta found less than 10% of isolates were resistant to any of the drugs tested (including chloramphenicol, streptomycin, ampicillin and nalidixic acid among others) (779). The ST6 IncHI1 plasmid may provide some selective advantage unrelated to drug resistance, perhaps associated with osmoprotectant properties of *betU* encoded on *Tn*6062 (721) or mercury resistance encoded on *Tn*21 (although this was present in most IncHI1 plasmid types). *BetU* may be associated with urinary carriage (721). Successful culturing of Typhi from urine has been reported from typhoid fever patients and asymptomatic carriers, albeit at much lower rates than isolation from stool (295, 780, 781, 782), and urinary shedding has been linked to disease transmission (780).

Unfortunately, SNPs known to confer nalidixic adid resistance were unsuitable targets for the GoldenGate assay (see 6.2.2), and thus the present study offers few novel insights into the evolution of fluoroquinolone resistance in Typhi. This could be remedied in future studies by additional typing of Nal resistance loci using another method such as Luminex SNP typing (as was done for the Kathmandu study), Sequenom SNP typing or targeted resequencing. However, the study did provide increased resolution within the H58 cluster for isolates that had previously been typed at Nal resistance SNP loci in GyrA (the Pasteur collection), or sequenced (see Figure 6.31). The results add further support to earlier assertions that the prevalence of resistance-associated GyrA SNPs among H58 isolates is due to multiple independent mutations arising in different H58 subtypes and not the expansion of a single clone following acquisition of resistance mutations (2).

Chapter 7

Final discussion

In this thesis Typhi and Paratyphi A, the chief agents of enteric fever, were investigated at the population genomic level using novel technologies for sequencing and genotyping. Before the study began, there were two Typhi genome sequences and one Paratyphi A genome available (46, 47, 49). These finished, annotated sequences provided a wealth of information about the genomes of these organisms, facilitating the identification of novel pathogenicity islands (in particular Typhi's SPI7 (46, 92)), prophage insertions (47, 224) and patterns of gene inactivation (46, 49). Comparative analyses of the Typhi and Paratyphi A genomes revealed patterns of similarities, yet they were not the sorts of similarities one might have expected. The genomes shared several genes that were rare among *Salmonella*, yet there were no obvious "smoking guns" with regards to their shared ability to cause systemic infection in humans (49). They each contained over 200 pseudogenes, yet most of the inactivated genes were different (46, 47, 49). Evidence of large-scale recombination was detected, yet the timing and mechanism was unclear (56). The genome sequences, along with that of Typhimurium (50), allowed the development of oligonucleotide arrays, leading to the first attempts to compare gene content between different strains of Typhi or Paratyphi A (49, 511). These array studies, along with comparison of the two Typhi genomes, confirmed what had long been suspected: the Typhi and Paratyphi A populations were both remarkably monomorphic, displaying very few substitution, deletion or insertion mutations that could give clues as to the evolution or population dynamics of these pathogens, or be targeted in genomic epidemiology studies. While the introduction of MLST led to rapid progress in population genetics studies for many bacterial pathogens (460), the analysis of Typhi, Paratyphi A and other monomorphic pathogen populations lagged behind (1, 471). It became clear that to pursue population-level studies in these bacteria would require the interrogation of vast amounts of sequence. In 2006, a landmark study addressing the global population structure of Typhi was completed by Roumagnac *et al.* (2), using a technique analagous to MLST but involving more than 25 times the number of loci normally targeted in MLST schemes. The study revealed several important aspects of the Typhi population - it was clonal in structure, with no evidence of recombination between lineages; clones were globally distributed, persisting side-by-side over decades on every continent; and a single clone, H58, had recently come to dominate the South East Asian population (2). The <100 SNPs identified in the study were hard-won, with two-thirds of the examined loci yielding no variation at all (2). Fortunately, it was at this point that two novel sequencing technologies came on the scene, bringing with them the possibility of examining entire populations at the whole-genome level. This is where the present study began, with the aim of exploiting the new genomics technologies to study the Typhi and Paratyphi A populations at high resolution.

For Typhi, the 2006 study (2) provided an unbiased framework for selecting isolates for whole-genome comparison. Seventeen novel isolates were chosen for sequencing, to complement the CT18 and Ty2 genomes that were already finished (Chapter 2). The ∼2,000 SNPs identified among these genomes described a single phylogenetic tree, strengthening the suspicion that the Typhi population is essentially clonal and uncomplicated by recombination between lineages. The whole-genome comparisons identified novel prophage over and above those discovered in Typhi CT18 and Ty2 (47, 224), but found no other insertions - just a host of deletions. This is reminiscent of the *M. tuberculosis* population, in which insertions even of prophage or plasmids are virtually unheard of (783, 784) but deletions are quite common (512). In *M. tuberculosis* this genetic isolation is easy to explain, as the bacterium can be isolated for decades inside encapsulated lung granulomas during asymptomatic infection of human hosts. In Typhi, it points towards a similar explanation, suggesting that long-term carriage in the gall bladder is the niche that really matters for the long-term survival of Typhi. This niche is less isolated than the granulomas inhabited by *M. tuberculosis*, which may account for the mid-level frequency of genetic exchange in Typhi: higher than that of *M. tuberculosis*, in the form of phage and plasmids, but lower than that of other *S.*

enterica serovars which show signs of recombination and more extensive plasmid and phage variation (178, 245). The deletions identified in the *M. tuberculosis* genome are conserved within lineages and have proved useful targets for typing (493). A similar scheme is currently in development for Typhi using some of the deletions identified in Chapter 2, which may provide a low-tech way of discriminating between Typhi lineages. The novel prophage identified among the Typhi genomes also provide an opportunity for further research. These regions are currently divided into many contigs but with a little additional sequencing could be improved to the point where cargo genes could be identified, which may prove an interesting source of genetic variation in the Typhi population. The Typhi genome comparisons were not particularly fruitful in terms of identifying genes under selection. However a relatively high level of variation was detected in the genes *yehT* (four SNPs, all nonsynonymous) and *yehU* (nine SNPs, all nonsynonymous) which together encode a two-component regulatory system. In the study by Roumagnac *et al.* (2), 13 SNPs were identified among 105 strains in the 500 bp fragment analysed from the sensor kinase gene *yehU*, compared to 0-5 in other gene fragments of the same size. The function of this regulatory system is currently being investigated using a combination of mouse models (with Typhimurium), phenotypic screening and gene expression analysis.

Typing of the Typhi SNPs identified by comparative genome analysis was applied in Chapter 6 to the analysis of Typhi populations in localised endemic areas. This was the highest resolution sequence-based interrogation of Typhi ever performed, and offered novel insights into the structure and dynamics of the Typhi populations in each area, which could be directly compared. These studies confirmed that multiple lineages of Typhi co-circulate in very narrow windows of time and space, but also revealed fluctuations in the composition of the population over time. The former is attributable to the central role of asymptomatic carriers, each of whom provides an independent and persistent source of infection with their own particular Typhi strain. The latter demonstrates the importance of other factors for short-term changes in the Typhi population, likely including things like climate, human behaviour and immunity in the human population, as well as random chance. Geographically, there was evidence of global dissemination of Typhi haplotypes over the long term, but also rather localised expansions of H58 sublineages in the short term (evident in the localised studies in

Vietnam (sublineage C) and in Nepal and India (sublineage B) spanning the last five years). There was also evidence of long term trends in the Typhi population, with SNP typing in both the global and Kenyan collections showing H58 becoming more and more dominant over time. However, the resolution offered by the current set of SNPs is inadequate to study the expansion of H58 in really fine detail. For example, isolates assigned to node H58-B undoubtedly harbour additional variation that happened not to be present among the seven sequenced isolates. Thus by SNP typing, we are blinded to this variation and may be tempted to assume that the H58-B isolates identified as potentially escaping the effects of the Vi conjugate vaccine in India represent a single clone, which is very closely related to that recently arrived in Kenya. However this may well be wrong, and can only be resolved by additional sequencing. Until it becomes feasible to engage in whole-genome sequencing of entire collections of Typhi isolates (which may not be too distant a prospect, discussed below) a multi-step approach may be best, where isolates are typed first using a subset of phylogenetically informative SNPs to discriminate major clusters, followed by additional sequencing to identify additional SNPs able to discriminate within those clusters. For example, based on the Typhi populations studied here, a sensible approach may be to (i) type a few SNPs that can discriminate between H58, H42 and other lineages plus all SNPs known to discriminate within H58, then (ii) select a manageable subset of isolates within the H58 (and potentially H42) clusters (maximising variability by pre-screening with PFGE) and sequence them, perhaps using a pooled approach to limit costs, followed by (iii) additional typing within clusters at SNP loci identified in step ii. This approach is currently being trialled in a large study of Typhi isolates from Kathmandu, using Sequenom for SNP typing (which allows rapid design and execution of SNP typing assays in under two weeks) and Solexa sequencing for SNP detection.

One outstanding issue with respect to the expansion of Typhi H58 is the role of the ST6 IncHI1 plasmid. The association between strain and plasmid is remarkable, with every single ST6 plasmid identified in this study (N=235) found within H58 host strains. However at this point it is not clear whether the apparent success of the ST6 plasmid subtype is simply hitchhiking on the expansion of H58, or whether it played a role in the success of the clone. The evidence from *IS*1 insertions suggests that the common ancestor of extant H58 lineages carried the plasmid (6.3.7), so the latter is at least possible. One potential mechanism for selection of ST6-containing strains, apart from drug resistance which is essentially identical to that conferred by plasmids of other IncHI1 subtypes, is the *betU*-carrying transposon *Tn*6062. The osmoprotectant BetU may help host cells to survive in the environment or within the urinary tract of human hosts, facilitating long-term carriage and transmission. Competition experiments in a variety of different media may help to detect phenotypic differences between ST6-containing and ST1-containing Typhi strains, although care must be taken to control for Typhi strain background.

Prior to this study, there was barely any phylogenetic information available for Paratyphi A. Comparison of gene content using microarrays had detected five chromosomal deletions including three prophage deletions, but no SNPs were known. A second Paratyphi A genome sequence was finished at the Sanger Institute, providing the first opportunity for comparison at the nucleotide level. However by this time, 454 and Solexa sequencing was also available, thus the first comparative analysis of Paratyphi A genomes included not two but seven isolates (Chapter 3). Selection of these isolates was random and we still do not know how representative they really are, but it seems they do capture a lot of the structure present in the larger global sample of isolates sequenced in pools (3.4.1). The level of variation detected among these genomes was markedly less than that observed between Typhi genomes (4.3.1), supporting previous suggestions that the Paratyphi A population harbours even less genetic variability than Typhi (479). The pooling strategy was designed to maximise genome-wide variation detection in the absence of any guiding phylogenetic framework like that available for Typhi. This was essentially a compromise between individual interrogation of a smaller number of isolates, which provides not only variation detection but also haplotypes appropriate for phylogenetic analysis, and maximal sensitivity to detect variation around the entire chromosome, which was expected to be very low. Haplotypes and phylogenetic structure could then be resolved using high throughput SNP typing. This last step was not pursued during the course of this study, partly due to time constraints and partly due to rapid changes in the relative cost-effectiveness of sequencing and SNP typing. As of mid-2009, it is increasingly feasible to sequence tens to hundreds of individual bacterial genomes of 4-5 Mbp in a single Solexa run using indexed libraries (785). As throughput continues to increase, thanks to improvements in sequencing chemistry that allow ever-longer paired-end reads (recent runs at the Sanger Institute exceed 30 Gb), it is likely that sequencing of even more genomes will be feasible very soon. Given that sequencing combines detection of known and novel variants, the cost of high throughput SNP typing will need to be significantly lower than that of sequencing in order to be of any benefit. The further pursuit of population structure in Paratyphi A currently involves SNP typing of ∼100 of the SNPs identified in Chapter 3, and will soon move to individual sequencing of the isolates so far sequenced only in pools.

References

1. Kidgell C., Reichard U., Wain J., Linz B., Torpdahl M., Dougan G. and Achtman M. ***Salmonella typhi*, the causative agent of typhoid fever, is approximately 50,000 years old**. *Infection, Genetics and Evolution*, **2**(1):39–45, 2002. 1, 33, 34, 40, 47, 158, 214, 272, 282

2. Roumagnac P., Weill F.X., Dolecek C., Baker S., Brisse S., Chinh N.T., Le T.A., Acosta C.J., Farrar J., Dougan G. and Achtman M. **Evolutionary history of *Salmonella typhi***. *Science*, **314**(5803):1301–1304, 2006. 1, 33, 40, 41, 47, 48, 49, 50, 52, 69, 73, 74, 76, 88, 90, 95, 154, 158, 210, 215, 228, 232, 236, 237, 238, 243, 244, 252, 265, 266, 267, 270, 271, 272, 273, 275, 280, 282, 283

3. Parry C.M., Hien T.T., Dougan G., White N.J. and Farrar J.J. **Typhoid fever**. *The New England Journal of Medicine*, **347**(22):1770–1782, 2002. 1, 2, 22, 27, 74, 83, 213

4. Marr J. **Typhoid Mary**. *The Lancet*, **353**(9165):1714, 1999. 1

5. Mortimer P. **Mr N the milker, and Dr Koch's concept of the healthy carrier**. *The Lancet*, **353**(9161):1354–1356, 1999. 1

6. Navarro F., Llovet T., Echeita M.A., Coll P., Aladuena A., Usera M.A. and Prats G. **Molecular typing of *Salmonella enterica* serovar typhi**. *Journal of Clinical Microbiology*, **34**(11):2831–2834, 1996. 1, 15, 33

7. Bhunia R., Hutin Y., Ramakrishnan R., Pal N., Sen T. and Murhekar M. **A typhoid fever outbreak in a slum of South Dumdum municipality, West Bengal, India, 2007: evidence for foodborne and waterborne transmission.** *BMC Public Health*, **9**:115+, 2009. 1

8. Chandel D.S., Nisar N., Thong K.L., Pang T. and Chaudhry R. **Role of molecular typing in an outbreak of *Salmonella paratyphi A*.** *Tropical Gastroenterology*, **21**(3):121–123, 2000. 1, 33, 96

9. The Public Health Laboratory Service Standing Sub Committee On The Bacteriological Examination Of Water Supplies. **Waterborne Infectious Disease in Britain**. *The Journal of Hygiene*, **81**(1):139–149, 1978. 1, 24

10. Crump J.A., Luby S.P. and Mintz E.D. **The global burden of typhoid fever**. *Bulletin of the World Health Organization*, **82**(5):346–353, 2004. 1, 24, 25, 29, 216, 217, 254, 263

11. Ekdahl K., de Jong B. and Andersson Y. **Risk of travel-associated typhoid and paratyphoid fevers in various regions.** *Journal of Travel Medicine*, **12**(4):197–204, 2005. 1, 24, 216, 254

12. Health Protection Agency. **Foreign travel-associated illness in England, Wales and Northern Ireland: 2007 report**. Technical report, Health Protection Agency, London, UK, 2007. 1, 24, 25

13. Grabenstein J.D., Pittman P.R., Greenwood J.T. and Engler R.J. **Immunization to protect the US Armed Forces: heritage, current practice, and prospects.** *Epidemiologic Reviews*, **28**:3–26, 2006. 1, 29

14. Fraser A., Goldberg E., Acosta C.J., Paul M. and Leibovici L. **Vaccines for preventing typhoid fever.** *Cochrane Database of Systematic Reviews*, (3), 2007. 1, 2, 29

15. Whitaker J.A., Franco-Paredes C., del Rio C. and Edupuganti S. **Rethinking typhoid fever vaccines: implications for travelers and people living in highly endemic areas.** *Journal of Travel Medicine*, **16**:46–52, 2009. 2, 29, 187

16. Chau T.T., Campbell J.I., Galindo C.M., Hoang N.V.M., Diep T.S., Nga T.T., Chau N.V.V., Tuan P.Q., Page A.L., Ochiai R.L., Schultsz C., Wain J., Bhutta Z.A., Parry C.M., Bhattacharya S.K., Dutta S., Agtini M., Dong B., Honghui Y., Anh D.D., do G.C., Naheed A., Albert M.J., Phetsouvanh R., Newton P.N., Basnyat B., Arjyal A., La T.T., Rang N.N., le T.P., Bay P.V.B., von Seidlein L., Dougan G., Clemens J.D., Vinh H., Hien T.T., Chinh N.T., Acosta C.J., Farrar J. and Dolecek C. **Antimicrobial Drug Resistance of *Salmonella enterica* Serovar Typhi in Asia and Molecular Mechanism of Reduced Susceptibility to the Fluoroquinolones**. *Antimicrobial Agents and Chemotherapy*, **51**(12):4315–4323, 2007. 2, 18, 28, 47, 187, 188, 213, 243, 244, 252, 275

17. Parry C.M. and Threlfall E.J. **Antimicrobial resistance in typhoidal and nontyphoidal salmonellae.** *Current Opinion in Infectious Diseases*, **21**(5):531–538, 2008. 2

18. Brenner F.W., Villar R.G., Angulo F.J., Tauxe R. and Swaminathan B. ***Salmonella* Nomenclature**. *Journal of Clinical Microbiology*, **38**(7):2465–2467, 2000. 2, 16

19. Health Protection Agency. **Kauffmann-White Scheme - 2007**. Health Protection Agency, 2007. 2, 3, 4, 6, 8, 16, 19, 23, 156

20. Popoff M.Y., Bockemühl J. and Gheesling L.L. **Supplement 2002 (no. 46) to the Kauffmann-White scheme.** *Research in Microbiology*, **155**(7):568–570, 2004. 2, 3

21. Crosa J.H., Brenner D.J., Ewing W.H. and Falkow S. **Molecular Relationships Among the Salmonelleae**. *Journal of Bacteriology*, **115**(1):307–315, 1973. 2

22. Reeves M.W., Evins G.M., Heiba A.A., Plikaytis B.D. and Farmer. **Clonal nature of Salmonella typhi and its genetic relatedness to other salmonellae as shown by multilocus enzyme electrophoresis, and proposal of Salmonella bongori comb. nov.** *Journal of Clinical Microbiology*, **27**(2):313–320, 1989. 2

23. Falush D., Torpdahl M., Didelot X., Conrad D.F., Wilson D.J. and Achtman M. **Mismatch induced speciation in *Salmonella*: model and data.** *Philosophical transactions of the Royal Society of London Series B: Biological sciences*, **361**(1475):2045–2053, 2006. 2, 15, 156

24. Jones T., Ingram L., Cieslak P., Vugia D., TobinD'Angelo M., Hurd S., Medus C., Cronquist A. and Angulo F. **Salmonellosis Outcomes Differ Substantially by Serotype**. *The Journal of Infectious Diseases*, **198**(1):109–114, 2008. 3, 4, 31

25. Galanis E., Lo Fo Wong D.M., Patrick M.E., Binsztein N., Cieslik A., Chalermchikit T., Aidara-Kane A., Ellis A., Angulo F.J., Wegener H.C. and World Health Organization Global Salm-Surv. **Web-based surveillance and global *Salmonella* distribution, 2000-2002.** *Emerging Infectious Diseases*, **12**(3):381–388, 2006. 3, 31

26. Herikstad H., Motarjemi Y. and Tauxe R.V. ***Salmonella* surveillance: a global survey of public health serotyping**. *Epidemiology and Infection*, **129**(1):1–8, 2002. 3

27. Scuderi G., Fantasia M. and Niglio T. **Results of the three-year surveillance by the Italian SALM-NET System: human isolates of *Salmonella* serotypes.** *European Journal of Epidemiology*, **16**(4):377–383, 2000. 3

28. OzFoodNet Working Group. **Foodborne disease investigation across Australia: annual report of the OzFoodNet network, 2003.** *Communicable Diseases Intelligence*, **28**(3):359–389, 2004. 3

29. Fernandes S.A., Tavechio A.T., Ghilardi A.C., Dias A.M., Almeida I.A. and Melo L.C. ***Salmonella* serovars isolated from humans in São Paulo State, Brazil, 1996-2003.** *Revista do Instituto de Medicina Tropical de São Paulo*, **48**(4):179–184, 2006. 3

30. US Centers for Disease Control and Prevention. **Preliminary FoodNet Data on the incidence of infection with pathogens transmitted commonly through food–10 States, 2008.** *Morbidity and Mortality Weekly Report*, **58**(13):333–337, 2009. 3

31. Threlfall E.J., Hall M.L. and Rowe B. ***Salmonella* bacteraemia in England and Wales, 1981-1990.** *Journal of Clinical Pathology*, **45**(1):34–36, 1992. 3, 148

32. Chiu C.H., Su L.H. and Chu C. ***Salmonella enterica* serotype Choleraesuis: epidemiology, pathogenesis, clinical disease, and treatment.** *Clinical Microbiology Reviews*, **17**(2):311–322, 2004. 3, 148

33. Edsall G., Gaines S., Landy M., Tigertt W.D., Sprinz H., Trapani R.J., Mandel A.D. and Benenson A.S. **Studies on infection and immunity in experimental typhoid fever. I. Typhoid fever in chimpanzees orally infected with *Salmonella typhosa*.** *The Journal of Experimental Medicine*, **112**:143–166, 1960. 4, 148, 181, 182

34. Langridge G.C., Nair S. and Wain J. **Nontyphoidal *Salmonella* Serovars Cause Different Degrees of Invasive Disease Globally**. *The Journal of Infectious Diseases*, **199**(4):602–603, 2009. 4, 31

35. Gordon M.A., Banda H.T., Gondwe M., Gordon S.B., Boeree M.J., Walsh A.L., Corkill J.E., Hart C.A., Gilks C.F. and Molyneux M.E. **Non-typhoidal salmonella bacteraemia among HIV-infected Malawian adults: high mortality and frequent recrudescence.** *AIDS*, **16**(12):1633–1641, 2002. 4

36. Fierer J. ***Salmonella* Outcomes**. *The Journal of Infectious Diseases*, **198**(11):1724, 2008. 4

37. Fierer J., Krause M., Tauxe R. and Guiney D. ***Salmonella typhimurium* bacteremia: association with the virulence plasmid.** *The Journal of Infectious Diseases*, **166**(3):639–642, 1992. 4

38. Kariuki S., Revathi G., Kariuki N., Kiiru J., Mwituria J., Muyodi J., Githinji J.W., Kagendo D., Munyalo A. and Hart C.A. **Invasive multidrug-resistant non-typhoidal *Salmonella* infections in Africa: zoonotic or anthroponotic transmission?** *Journal of Medical Microbiology*, **55**(5):585–591, 2006. 4

39. Gordon M.A., Graham S.M., Walsh A.L., Wilson L., Phiri A., Molyneux E., Zijlstra E.E., Heyderman R.S., Hart C.A. and Molyneux M.E. **Epidemics of invasive *Salmonella enterica* serovar Enteritidis and *S. enterica* serovar Typhimurium infection associated with multidrug resistance among adults and children in Malawi.** *Clinical Infectious Diseases*, **46**(7):963–969, 2008. 4

40. Jean S.S., Wang J.Y. and Hsueh P.R. **Bacteremia caused by *Salmonella enterica* serotype Choleraesuis in Taiwan.** *Journal of Microbiology, Immunology, and Infection*, **39**(5):358–365, 2006. 4

41. Sharp P.M. **Determinants of DNA sequence divergence between Escherichia coli and Salmonella typhimurium: codon usage, map position, and concerted evolution.** *Journal of Molecular Evolution*, **33**(1):23–33, 1991. 4, 154, 158

42. Ochman H. and Wilson A.C. **Evolution in bacteria: evidence for a universal substitution rate in cellular genomes.** *Journal of Molecular Evolution*, **26**(1-2):74–86, 1987. 4, 154, 156, 158, 161

43. Achtman M., Morelli G., Zhu P., Wirth T., Diehl I., Kusecek B., Vogler A.J., Wagner D.M., Allender C.J., Easterday W.R., Chenal-Francisque V., Worsham P., Thomson N.R., Parkhill J., Lindler L.E., Carniel E. and Keim P. **Microevolution and history of the plague bacillus, *Yersinia pestis*.** *Proceedings of the National Academy of Sciences of the United States of America*, **101**(51):17837–17842, 2004. 4, 33, 34, 39, 154, 158, 161

44. Turner S.L. and Young J.P. **The glutamine synthetases of rhizobia: phylogenetics and evolutionary implications.** *Molecular Biology and Evolution*, **17**(2):309–319, 2000. 4, 154, 156, 158, 180

45. Battistuzzi F.U., Feijao A. and Hedges S.B. **A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land.** *BMC Evolutionary Biology*, **4**, 2004. 4, 154, 156, 158, 180

46. Parkhill J., Dougan G., James K.D., Thomson N.R., Pickard D., Wain J., Churcher C., Mungall K.L., Bentley S.D., Holden M.T., Sebaihia M., Baker S., Basham D., Brooks K., Chillingworth T., Connerton P., Cronin A., Davis P., Davies R.M., Dowd L., White N., Farrar J., Feltwell T., Hamlin N., Haque A., Hien T.T., Holroyd S., Jagels K., Krogh A., Larsen T.S., Leather S., Moule S., O'Gaora P., Parry C., Quail M., Rutherford K., Simmonds M., Skelton J., Stevens K., Whitehead S. and Barrell B.G. **Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18**. *Nature*, **413**(6858):848–852, 2001. 4, 8, 11, 14, 15, 16, 18, 37, 47, 66, 78, 80, 83, 149, 189, 195, 197, 200, 210, 281

47. Deng W., Liou S.R., 3rd G.P., Mayhew G.F., Rose D.J., Burland V., Kodoyianni V., Schwartz D.C. and Blattner F.R. **Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18**. *Journal of Bacteriology*, **185**(7):2330–2337, 2003. 4, 15, 16, 18, 36, 37, 47, 66, 67, 79, 83, 87, 149, 281, 282

48. Chen F., Poppe C., Liu G.R., Li Y.G., Peng Y.H., Sanderson K.E., Johnston R.N. and Liu S.L. **A genome map of *Salmonella enterica* serovar Agona: numerous insertions and deletions reflecting the evolutionary history of a human pathogen.** *FEMS Microbiology Letters*, **293**(2):188–195, 2009. 4

49. McClelland M., Sanderson K.E., Clifton S.W., Latreille P., Porwollik S., Sabo A., Meyer R., Bieri T., Ozersky P., McLellan M., Harkins C.R., Wang C., Nguyen C., Berghoff A., Elliott G., Kohlberg S., Strong C., Du F., Carter J., Kremizki C., Layman D., Leonard S., Sun H., Fulton L., Nash W., Miner T., Minx P., Delehaunty K., Fronick C., Magrini V., Nhan M., Warren W., Florea L., Spieth J. and Wilson R.K. **Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid**. *Nature Genetics*, **36**(12):1268–1274, 2004. 4, 15, 18, 19, 37, 96, 149, 150, 165, 168, 178, 185, 186, 281

50. McClelland M., Sanderson K.E., Spieth J., Clifton S.W., Latreille P., Courtney L., Porwollik S., Ali J., Dante M., Du F., Hou S., Layman D., Leonard S., Nguyen C., Scott K., Holmes A., Grewal N., Mulvaney E., Ryan E., Sun H., Florea L., Miller W., Stoneking T., Nhan M., Waterston R. and Wilson R.K. **Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2.** *Nature*, **413**(6858):852–856, 2001. 4, 5, 15, 16, 18, 37, 281

51. Hayashi K., Morooka N., Yamamoto Y., Fujita K., Isono K., Choi S., Ohtsubo E., Baba T., Wanner B.L., Mori H. and Horiuchi T. **Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110.** *Molecular Systems Biology*, **2**, 2006. 4

52. Mcclelland M., Florea L., Sanderson K., Clifton S.W., Parkhill J., Churcher C., Dougan G., Wilson R.K. and Miller W. **Comparison of the *Escherichia coli* K-12 genome with sampled genomes of a *Klebsiella pneumoniae* and three *Salmonella enterica* serovars, Typhimurium, Typhi and Paratyphi**. *Nucleic Acids Research*, **28**(24):4974–4986, 2000. 4

53. Bäumler A.J., Tsolis R.M., Ficht T.A. and Adams L.G. **Evolution of host adaptation in *Salmonella enterica*.** *Infection and Immunity*, **66**(10):4579–4587, 1998. 4, 5

54. Mills D.M., Bajaj V. and Lee C.A. **A 40 kb chromosomal fragment encoding *Salmonella typhimurium* invasion genes is absent from the corresponding region of the *Escherichia coli* K-12 chromosome.** *Molecular Microbiology*, **15**(4):749–759, 1995. 4, 9, 10

55. Shea J.E., Hensel M., Gleeson C. and Holden D.W. **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):2593–2597, 1996. 4, 9

56. Didelot X., Achtman M., Parkhill J., Thomson N.R. and Falush D. **A bimodal pattern of relatedness between the *Salmonella* Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination?** *Genome Research*, **17**(1):61–68, 2007. 4, 15, 19, 56, 76, 149, 153, 156, 167, 170, 179, 181, 281

57. Edwards R. **Comparative genomics of closely related salmonellae**. *Trends in Microbiology*, **10**(2):94–99, 2002. 4

58. Bäumler A.J. **The record of horizontal gene transfer in *Salmonella*.** *Trends in Microbiology*, **5**(8):318–322, 1997. 5, 11

59. Reeves P., Hobbs M., Valvano M., Skurnik M., Whitfield C., Coplin D., Kido N., Klena J., Maskell D. and Raetz C. **Bacterial polysaccharide synthesis and gene nomenclature**. *Trends in Microbiology*, **4**(12):495–503, 1996. 6

60. Brahmbhatt H.N., Wyk P., Quigley N.B. and Reeves P.R. **Complete physical map of the *rfb* gene cluster encoding biosynthetic enzymes for the O antigen of *Salmonella typhimurium* LT2.** *Journal of Bacteriology*, **170**(1):98–102, 1988. 6

61. Verma N.K., Quigley N.B. and Reeves P.R. **O-antigen variation in *Salmonella* spp.: *rfb* gene clusters of three strains.** *Journal of Bacteriology*, **170**(1):103–107, 1988. 6

62. Samuel G. **Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly**. *Carbohydrate Research*, **338**(23):2503–2519, 2003. 6

63. Liu D., Verma N.K., Romana L.K. and Reeves P.R. **Relationships among the *rfb* regions of *Salmonella* serovars A, B, and D.** *Journal of Bacteriology*, **173**(15):4814–4819, 1991. 6, 132

64. Xiang S.H., Haase A.M. and Reeves P.R. **Variation of the *rfb* gene clusters in *Salmonella enterica*.** *Journal of Bacteriology*, **175**(15):4877–4884, 1993. 6

65. Reeves P. **Evolution of *Salmonella* O antigen variation by interspecific gene transfer on a large scale**. *Trends in Genetics*, **9**(1):17–22, 1993. 6

66. Heinrichs D.E., Yethon J.A. and Whitfield C. **Molecular basis for structural diversity in the core regions of the lipopolysaccharides of *Escherichia coli* and *Salmonella enterica*.** *Molecular Microbiology*, **30**(2):221–232, 1998. 6, 162

67. Gibson D.L., White A.P., Snyder S.D., Martin S., Heiss C., Azadi P., Surette M. and Kay W.W. ***Salmonella* produces an O-antigen capsule regulated by AgfD and important for environmental persistence.** *Journal of Bacteriology*, **188**(22):7722–7730, 2006. 6

68. Crawford R.W., Gibson D.L., Kay W.W. and Gunn J.S. **Identification of a bile-induced exopolysaccharide required for *Salmonella* biofilm formation on gallstone surfaces.** *Infection and Immunity*, **76**(11):5341–5349, 2008. 6

69. Barak J.D., Jahn C.E., Gibson D.L. and Charkowski A.O. **The role of cellulose and O-antigen capsule in the colonization of plants by *Salmonella enterica*.** *Molecular Plant-Microbe Interactions*, **20**(9):1083–1091, 2007. 6

70. Harshey R. and Toguchi A. **Spinning tails: homologies among bacterial flagellar systems**. *Trends in Microbiology*, **4**(6):226–231, 1996. 8

71. Aldridge P. **Regulation of flagellar assembly**. *Current Opinion in Microbiology*, **5**(2):160–165, 2002. 8

72. Chilcott G.S. and Hughes K.T. **Coupling of Flagellar Gene Expression to Flagellar Assembly in *Salmonella enterica* Serovar Typhimurium and *Escherichia coli***. *Microbiology and Molecular Biology Reviews*, **64**(4):694–708, 2000. 8

73. Kutsukake K. and Iino T. **A trans-acting factor mediates inversion of a specific DNA segment in flagellar phase variation of *Salmonella*.** *Nature*, **284**(5755):479–481, 1980. 8

74. Simon M., Zieg J., Silverman M., Mandel G. and Doolittle R. **Phase variation: evolution of a controlling element.** *Science*, **209**(4463):1370–1374, 1980. 8

75. Fierer J. and Guiney D.G. **Diverse virulence traits underlying different clinical outcomes of *Salmonella* infection.** *The Journal of Clinical Investigation*, **107**(7):775–780, 2001. 8

76. Mortimer C., Gharbia S., Logan J., Peters T. and Arnold C. **Flagellin gene sequence evolution in *Salmonella***. *Infection, Genetics and Evolution*, **7**(4):411–415, 2007. 8

77. Cookson B.T. and Bevan M.J. **Identification of a natural T cell epitope presented by *Salmonella*-infected macrophages and recognized by T cells from orally immunized mice.** *Journal of Immunology*, **158**(9):4310–4319, 1997. 8

78. Kanto S., Okino H., Aizawa S. and Yamaguchi S. **Amino acids responsible for flagellar shape are distributed in terminal regions of flagellin**. *Journal of Molecular Biology*, **219**(3):471–480, 1991. 8

79. Smith N.H., Beltran P. and Selander R.K. **Recombination of *Salmonella* phase 1 flagellin genes generates new serovars.** *Journal of Bacteriology*, **172**(5):2209–2216, 1990. 8

80. Li J., Nelson K., McWhorter A.C., Whittam T.S. and Selander R.K. **Recombinational basis of serovar diversity in *Salmonella enterica*.** *Proceedings of the National Academy of Sciences of the United States of America*, **91**(7):2552–2556, 1994. 8

81. Felix A. **A new antigen of *B. typhosus*, it's relation to virulence and to active and passive immunisation**. *The Lancet*, **224**(5787):186–191, 1934. 8

82. Selander R.K., Smith N.H., Li J., Beltran P., Ferris K.E., Kopecko D.J. and Rubin F.A. **Molecular evolutionary genetics of the cattle-adapted serovar *Salmonella dublin*.** *Journal of Bacteriology*, **174**(11):3587–3592, 1992. 8, 11, 16

83. Daniels E.M., Schneerson R., Egan W.M., Szu S.C. and Robbins J.B. **Characterization of the *Salmonella paratyphi C* Vi polysaccharide.** *Infection and Immunity*, **57**(10):3159–3164, 1989. 8, 11, 16, 168

84. Snellings N.J., Johnson E.M., Kopecko D.J., Collins H.H. and Baron L.S. **Genetic regulation of variable Vi antigen expression in a strain of *Citrobacter freundii*.** *Journal of Bacteriology*, **145**(2):1010–1017, 1981. 8

85. Houng H.S., Noon K.F., Ou J.T. and Baron L.S. **Expression of Vi antigen in *Escherichia coli* K-12: characterization of ViaB from *Citrobacter freundii* and identity of ViaA with RcsB.** *Journal of Bacteriology*, **174**(18):5910–5915, 1992. 8, 11

86. JOHNSON E.M., KRAUSKOPF B. and BARON L.S. **Genetic mapping of Vi and somatic antigenic determinants in *Salmonella*.** *Journal of Bacteriology*, **90**:302–308, 1965. 8

87. Snellings N.J., Johnson E.M. and Baron L.S. **Genetic basis of Vi antigen expression in *Salmonella paratyphi C*.** *Journal of Bacteriology*, **131**(1):57–62, 1977. 8

88. Kolyva S., Waxin H. and Popoff M.Y. **The Vi antigen of *Salmonella typhi*: molecular analysis of the *viaB* locus.** *Journal of General Microbiology*, **138**(2):297–304, 1992. 8

89. Hashimoto Y., Li N., Yokoyama H. and Ezaki T. **Complete nucleotide sequence and molecular characterization of ViaB region encoding Vi antigen in *Salmonella typhi*.** *Journal of Bacteriology*, **175**(14):4456–4465, 1993. 8

90. Virlogeux I., Waxin H., Ecobichon C. and Popoff M.Y. **Role of the *viaB* locus in synthesis, transport and expression of *Salmonella typhi* Vi antigen.** *Microbiology*, **141**(12):3039–3047. 8

91. Johnson E.M. and Baron L.S. **Genetic transfer of the Vi antigen from *Salmonella typhosa* to *Escherichia coli*.** *Journal of Bacteriology*, **99**(1):358–359, 1969. 8

92. Pickard D., Wain J., Baker S., Line A., Chohan S., Fookes M., Barron A., Gaora P.O., Chabalgoity J.A., Thanky N., Scholes C., Thomson N., Quail M., Parkhill J. and Dougan G. **Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding *Salmonella enterica* pathogenicity island SPI-7.** *Journal of Bacteriology*, **185**(17):5055–5065, 2003. 8, 11, 13, 47, 74, 81, 162, 168, 281

93. Liu W.Q., Feng Y., Wang Y., Zou Q.H., Chen F., Guo J.T., Peng Y.H., Jin Y., Li Y.G., Hu S.N., Johnston R.N., Liu G.R. and Liu S.L. ***Salmonella paratyphi* C: Genetic Divergence from *Salmonella choleraesuis* and Pathogenic Convergence with *Salmonella typhi***. *PLoS ONE*, **4**(2):e4510+, 2009. 8, 11, 15, 150, 178

94. Pickard D., Li J., Roberts M., Maskell D., Hone D., Levine M., Dougan G. and Chatfield S. **Characterization of defined *ompR* mutants of *Salmonella typhi*: *ompR* is involved in the regulation of Vi polysaccharide expression.** *Infection and Immunity*, **62**(9):3984–3993, 1994. 8

95. Hornick R.B., Greisman S.E., Woodward T.E., DuPont H.L., Dawkins A.T. and Snyder M.J. **Typhoid fever: pathogenesis and immunologic control.** *The New England Journal of Medicine*, **283**(13):686–691, 1970. 8, 20

96. Robbins J.D. and Robbins J.B. **Reexamination of the protective role of the capsular polysaccharide (Vi antigen) of *Salmonella typhi*.** *The Journal of Infectious Diseases*, **150**(3):436–449, 1984. 8

97. Looney R.J. and Steigbigel R.T. **Role of the Vi antigen of *Salmonella typhi* in resistance to host defense in vitro.** *The Journal of Laboratory and Clinical Medicine*, **108**(5):506–516, 1986. 8

98. Sharma A. and Qadri A. **Vi polysaccharide of *Salmonella typhi* targets the prohibitin family of molecules in intestinal epithelial cells and suppresses early inflammatory responses**. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(50):17492–17497, 2004. 8

99. Thanassi D. **The chaperone/usher pathway: a major terminal branch of the general secretory pathway**. *Current Opinion in Microbiology*, **1**(2):223–231, 1998. 9

100. Weening E.H., Barker J.D., Laarakker M.C., Humphries A.D., Tsolis R.M. and Baumler A.J. **The *Salmonella enterica* Serotype Typhimurium *lpf, bcf, stb, stc, std,* and *sth* Fimbrial Operons Are Required for Intestinal Persistence in Mice**. *Infection and Immunity*, **73**(6):3358–3366, 2005. 9

101. van der Velden A.W..M., Baumler A.J., Tsolis R.M. and Heffron F. **Multiple Fimbrial Adhesins Are Required for Full Virulence of *Salmonella typhimurium* in Mice**. *Infection and Immunity*, **66**(6):2803–2808, 1998. 9

102. Althouse C., Patterson S., Fedorka-Cray P. and Isaacson R.E. **Type 1 Fimbriae of *Salmonella enterica* Serovar Typhimurium Bind to Enterocytes and Contribute to Colonization of Swine *In Vivo***. *Infection and Immunity*, **71**(11):6446–6452, 2003. 9

103. Porwollik S. **Lateral gene transfer in *Salmonella***. *Microbes and Infection*, **5**(11):977–989, 2003. 9, 37, 38

104. Townsend S.M., Kramer N.E., Edwards R., Baker S., Hamlin N., Simmonds M., Stevens K., Maloy S., Parkhill J., Dougan G. and Baumler A.J. ***Salmonella enterica* Serovar Typhi Possesses a Unique Repertoire of Fimbrial Gene Sequences**. *Infection and Immunity*, **69**(5):2894–2901, 2001. 9

105. Kuehn M.J., Normark S. and Hultgren S.J. **Immunoglobulin-like PapD chaperone caps and uncaps interactive surfaces of nascently translocated pilus subunits.** *Proceedings of the National Academy of Sciences of the United States of America*, **88**(23):10586–10590, 1991. 9

106. Sauer F.G., Barnhart M., Choudhury D., Knight S.D., Waksman G. and Hultgren S.J. **Chaperone-assisted pilus assembly and bacterial attachment.** *Current Opinion in Structural Biology*, **10**(5):548–556, 2000. 9

107. Norgren M., Båga M., Tennent J.M. and Normark S. **Nucleotide sequence, regulation and functional analysis of the *papC* gene required for cell surface localization of Pap pili of uropathogenic *Escherichia coli*.** *Molecular Microbiology*, **1**(2):169–178, 1987. 9

108. Klemm P. and Christiansen G. **The *fimD* gene required for cell surface localization of *Escherichia coli* type 1 fimbriae.** *Molecular and General Genetics*, **220**(2):334–338, 1990. 9

109. Jacob-Dubuisson F., Striker R. and Hultgren S.J. **Chaperone-assisted self-assembly of pili independent of cellular energy.** *The Journal of Biological Chemistry*, **269**(17):12447–12455, 1994. 9

110. Dodson K.W., Jacob-Dubuisson F., Striker R.T. and Hultgren S.J. **Outer-membrane PapC molecular usher discriminately recognizes periplasmic chaperone-pilus subunit complexes.** *Proceedings of the National Academy of Sciences of the United States of America*, **90**(8):3670–3674, 1993. 9

111. Huang Y., Smith B.S., Chen L.X., Baxter R.H. and Deisenhofer J. **Insights into pilus assembly and secretion from the structure and functional characterization of usher PapC.** *Proceedings of the National Academy of Sciences of the United States of America*, **106**(18):7403–7407, 2009. 9

112. Thanassi D.G., Saulino E.T., Lombardo M.J., Roth R., Heuser J. and Hultgren S.J. **The PapC usher forms an oligomeric channel: implications for pilus biogenesis across the outer membrane.** *Proceedings of the National Academy of Sciences of the United States of America*, **95**(6):3146–3151, 1998. 9

113. Craig L., Pique M.E. and Tainer J.A. **Type IV pilus structure and bacterial pathogenicity.** *Nature Reviews Microbiology*, **2**(5):363–378, 2004. 9

114. Zhang X.L., Tsui I.S.M., Yip C.M.C., Fung A.W.Y., Wong D.K.H., Dai X., Yang Y., Hackett J. and Morris C. ***Salmonella enterica* serovar Typhi uses type IVB pili to enter human intestinal epithelial cells**. *Infection and Immunity*, **68**(6):3067–3073, 2000. 9, 11

115. Hammar M., Bian Z. and Normark S. **Nucleator-dependent intercellular assembly of adhesive curli organelles in *Escherichia coli***. *Proceedings of the National Academy of Sciences of the United States of America*, **93**(13):6562–6566, 1996. 9

116. Sokurenko E.V., Chesnokova V., Dykhuizen D.E., Ofek I., Wu X.R., Krogfelt K.A., Struve C., Schembri M.A. and Hasty D.L. **Pathogenic adaptation of *Escherichia coli* by natural variation of the FimH adhesin**. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(15):8922–8926, 1998. 9

117. Duncan M.J., Mann E.L., Cohen M.S., Ofek I., Sharon N. and Abraham S.N. **The Distinct Binding Specificities Exhibited by Enterobacterial Type 1 Fimbriae Are Determined by Their Fimbrial Shafts**. *Journal of Biological Chemistry*, **280**(45):37707–37716, 2005. 9

118. Sokurenko E.V., Feldgarden M., Trintchina E., Weissman S.J., Avagyan S., Chattopadhyay S., Johnson J.R. and Dykhuizen D.E. **Selection Footprint in the FimH Adhesin Shows Pathoadaptive Niche Differentiation in *Escherichia coli*.** *Molecular Biology and Evolution*, **21**(7):1373–1383, 2004. 9

119. Wilson R.L., Elthon J., Clegg S. and Jones B.D. ***Salmonella enterica* Serovars Gallinarum and Pullorum Expressing *Salmonella enterica* Serovar Typhimurium Type 1 Fimbriae Exhibit Increased Invasiveness for Mammalian Cells.** *Infection and Immunity*, **68**(8):4782–4785, 2000. 9

120. Guo A., Cao S., Tu L., Chen P., Zhang C., Jia A., Yang W., Liu Z., Chen H. and Schifferli D.M. **FimH alleles direct preferential binding of *Salmonella* to distinct mammalian cells or to avian cells.** *Microbiology*, **155**(Pt 5):1623–1633, 2009. 9, 148, 177

121. Marcus S.L., Brumell J.H., Pfeifer C.G. and Finlay B.B. ***Salmonella* pathogenicity islands: big virulence in small packages.** *Microbes and Infection*, **2**(2):145–156, 2000. 9, 10

122. Hueck C.J. **Type III protein secretion systems in bacterial pathogens of animals and plants.** *Microbiology and Molecular Biology Reviews*, **62**(2):379–433, 1998. 9

123. Lee V.T. and Schneewind O. **Type III secretion machines and the pathogenesis of enteric infections caused by *Yersinia* and *Salmonella* spp.** *Immunological Reviews*, **168**:241–255, 1999. 9

124. McGhie E.J., Brawn L.C., Hume P.J., Humphreys D. and Koronakis V. ***Salmonella* takes control: effector-driven manipulation of the host.** *Current Opinion in Microbiology*, **12**(1):117–124, 2009. 9, 10

125. Collazo C.M. and Galán J.E. **The invasion-associated type-III protein secretion system in *Salmonella*-a review.** *Gene*, **192**(1):51–59, 1997. 9, 10

126. Hensel M., Nikolaus T. and Egelseer C. **Molecular and functional analysis indicates a mosaic structure of *Salmonella* pathogenicity island 2.** *Molecular Microbiology*, **31**(2):489–498, 1999. 9, 10

127. Ehrbar K., Friebel A., Miller S.I. and Hardt W.D. **Role of the *Salmonella* pathogenicity island 1 (SPI-1) protein InvB in type III secretion of SopE and SopE2, two *Salmonella* effector proteins encoded outside of SPI-1.** *Journal of Bacteriology*, **185**(23):6950–6967, 2003. 10

128. Ehrbar K. and Hardt W.D. **Bacteriophage-encoded type III effectors in *Salmonella enterica* subspecies 1 serovar Typhimurium.** *Infection, Genetics and Evolution*, **5**(1):1–9, 2005. 10, 15

129. Mirold S., Ehrbar K., Weissmüller A., Prager R., Tschäpe H., Rüssmann H. and Hardt W.D. ***Salmonella* host cell invasion emerged by acquisition of a mosaic of separate genetic elements, including *Salmonella* pathogenicity island 1 (SPI1), SPI5, and *sopE2*.** *Journal of Bacteriology*, **183**(7):2348–2358, 2001. 10

130. Ochman H. and Groisman E.A. **Distribution of pathogenicity islands in *Salmonella* spp.** *Infection and Immunity*, **64**(12):5410–5412, 1996. 10

131. Chan K., Baker S., Kim C.C., Detweiler C.S., Dougan G. and Falkow S. **Genomic Comparison of *Salmonella enterica* Serovars and *Salmonella bongori* by Use of an *S. enterica* Serovar Typhimurium DNA Microarray.** *Journal of Bacteriology*, **185**(2):553–563, 2003. 10

132. Daefler S. **Type III secretion by *Salmonella typhimurium* does not require contact with a eukaryotic host.** *Molecular Microbiology*, **31**(1):45–51, 1999. 10

133. Ellermeier C.D., Ellermeier J.R. and Slauch J.M. **HilD, HilC and RtsA constitute a feed forward loop that controls expression of the SPI1 type three secretion system regulator hilA in *Salmonella enterica* serovar Typhimurium.** *Molecular Microbiology*, **57**(3):691–705, 2005. 10

134. Ellermeier J.R. and Slauch J.M. **Adaptation to the host environment: regulation of the SPI1 type III secretion system in *Salmonella enterica* serovar Typhimurium.** *Current Opinion in Microbiology*, **10**(1):24–29, 2007. 10

135. Wallis T.S. and Galyov E.E. **Molecular basis of *Salmonella*-induced enteritis.** *Molecular Microbiology*, **36**(5):997–1005, 2000. 10

136. Finlay B.B., Ruschkowski S. and Dedhar S. **Cytoskeletal rearrangements accompanying *Salmonella* entry into epithelial cells.** *Journal of Cell Science*, **99**(2):283–296, 1991. 10

137. Francis C.L., Ryan T.A., Jones B.D., Smith S.J. and Falkow S. **Ruffles induced by *Salmonella* and other stimuli direct macropinocytosis of bacteria.** *Nature*, **364**(6438):639–642, 1993. 10

138. Jones B.D., Ghori N. and Falkow S. ***Salmonella typhimurium* initiates murine infection by penetrating and destroying the specialized epithelial M cells of the Peyer's patches.** *The Journal of Experimental Medicine*, **180**(1):15–23, 1994. 10

139. Hayward R. and Koronakiss V. **Direct modulation of the host cell cytoskeleton by *Salmonella* actin-binding proteins.** *Trends in Cell Biology*, **12**(1):15–20, 2002. 10

140. Hu Q., Coburn B., Deng W., Li Y., Shi X., Lan Q., Wang B., Coombes B.K. and Finlay B.B. ***Salmonella enterica* serovar Senftenberg human clinical isolates lacking SPI-1.** *Journal of Clinical Microbiology*, **46**(4):1330–1336, 2008. 10

141. Waterman S.R. and Holden D.W. **Functions and effectors of the *Salmonella* pathogenicity island 2 type III secretion system.** *Cellular Microbiology*, **5**:501–511, 2003. 10

142. Garmendia J., Beuzón C.R., Ruiz-Albert J. and Holden D.W. **The roles of SsrA-SsrB and OmpR-EnvZ in the regulation of genes encoding the *Salmonella typhimurium* SPI-2 type III secretion system.** *Microbiology*, **149**(9):2385–2396, 2003. 10

143. Bustamante V.H., Martínez L.C., Santana F.J., Knodler L.A., Steele-Mortimer O. and Puente J.L. **HilD-mediated transcriptional cross-talk between SPI-1 and SPI-2.** *Proceedings of the National Academy of Sciences of the United States of America*, **105**(38):14591–14596, 2008. 10

144. Fass E. and Groisman E.A. **Control of *Salmonella* pathogenicity island-2 gene expression.** *Current Opinion in Microbiology*, **12**(2):199–204, 2009. 10

145. Blanc-Potard A.B., Solomon F., Kayser J. and Groisman E.A. **The SPI-3 Pathogenicity Island of *Salmonella enterica*.** *Journal of Bacteriology*, **181**(3):998–1004, 1999. 10

146. Amavisit P., Lightfoot D., Browning G.F. and Markham P.F. **Variation between pathogenic serovars within *Salmonella* pathogenicity islands.** *Journal of Bacteriology*, **185**(12):3624–3635, 2003. 10, 11

147. China B. and Goffaux F. **Secretion of virulence factors by *Escherichia coli*.** *Veterinary research*, **30**(2-3):181–202, 1999. 11

148. Gerlach R.G., Jäckel D., Stecher B., Wagner C., Lupas A., Hardt W.D. and Hensel M. ***Salmonella* Pathogenicity Island 4 encodes a giant non-fimbrial adhesin and the cognate type 1 secretion system.** *Cellular Microbiology*, **9**(7):1834–1850, 2007. 11

149. Morgan E., Bowen A.J., Carnell S.C., Wallis T.S. and Stevens M.P. **SiiE is secreted by the *Salmonella enterica* serovar Typhimurium pathogenicity island 4-encoded secretion system and contributes to intestinal colonization in cattle.** *Infection and Immunity*, **75**(3):1524–1533, 2007. 11

150. Gerlach R.G., Cláudio N., Rohde M., Jäckel D., Wagner C. and Hensel M. **Cooperation of *Salmonella* pathogenicity islands 1 and 4 is required to breach epithelial barriers.** *Cellular Microbiology*, **10**(11):2364–2376, 2008. 11

151. Wong K.K., Mcclelland M., Stillwell L.C., Sisk E.C., Thurston S.J. and Saffer J.D. **Identification and Sequence Analysis of a 27-Kilobase Chromosomal Fragment Containing a *Salmonella* Pathogenicity Island Located at 92 Minutes on the Chromosome Map of *Salmonella enterica* Serovar Typhimurium LT2.** *Infection and Immunity*, **66**(7):3365–3371, 1998. 11

152. Wood M.W., Jones M.A., Watson P.R., Hedges S., Wallis T.S. and Galyov E.E. **Identification of a pathogenicity island required for *Salmonella* enteropathogenicity.** *Molecular Microbiology*, **29**(3):883–891, 1998. 11

153. Chiu C.H., Tang P., Chu C., Hu S., Bao Q., Yu J., Chou Y.Y., Wang H.S. and Lee Y.S. **The genome sequence of *Salmonella enterica* serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen.** *Nucleic Acids Research*, **33**(5):1690–1698, 2005. 11, 15

154. Shah D.H., Lee M.J., Park J.H., Lee J.H., Eo S.K., Kwon J.T. and Chae J.S. **Identification of *Salmonella gallinarum* virulence genes in a chicken infection model using PCR-based signature-tagged mutagenesis.** *Microbiology*, **151**(12):3957–3968, 2005. 11

155. Vernikos G.S. and Parkhill J. **Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands.** *Bioinformatics*, **22**(18):2196–2203, 2006. 11, 18, 79

156. Saroj S.D., Shashidhar R., Karani M. and Bandekar J.R. **Distribution of *Salmonella* pathogenicity island (SPI)-8 and SPI-10 among different serotypes of Salmonella.** *Journal of Medical Microbiology*, **57**(4), 2008. 11, 149

157. Jain R., Rivera M.C., Moore J.E. and Lake J.A. **Horizontal gene transfer in microbial genome evolution.** *Theoretical Population Biology*, **61**(4):489–495, 2002. 11

158. Ochman H., Lawrence J.G. and Groisman E.A. **Lateral gene transfer and the nature of bacterial innovation.** *Nature*, **405**(6784):299–304, 2000. 11

159. Miller R.V. **Microbial Evolution: Gene establishment, survival and exchange**, chapter 7-9. ASM Press, Washington DC, 2004. 11

160. Datta N. ***Salmonella typhi in vivo* acquires resistance to both chloramphenicol and co-trimoxazole.** *The Lancet*, **317**(8231):1181–1183, 1981. 11, 14, 27

161. Canchaya C., Fournous G., Chibani-Chennoufi S., Dillmann M.L. and Brüssow H. **Phage as agents of lateral gene transfer.** *Current Opinion in Microbiology*, **6**(4):417–424, 2003. 11, 14

162. Garcia-Quintanilla M., Ramos-Morales F. and Casadesus J. **Conjugal Transfer of the *Salmonella enterica* Virulence Plasmid in the Mouse Intestine.** *Journal of Bacteriology*, **190**(6):1922–1927, 2008. 11

163. Poole T. and Crippen T. **Conjugative plasmid transfer between *Salmonella enterica* Newport and *Escherichia coli* within the gastrointestinal tract of the lesser mealworm beetle, *Alphitobius diaperinus* (Coleoptera: Tenebrionidae).** *Poultry Science*, **88**(8):1553–1558, 2009. 11

164. Poppe C., Martin L.C., Gyles C.L., Reid-Smith R., Boerlin P., McEwen S.A., Prescott J.F. and Forward K.R. **Acquisition of resistance to extended-spectrum cephalosporins by *Salmonella enterica* subsp. *enterica* serovar Newport and *Escherichia coli* in the turkey poult intestinal tract.** *Applied and Environmental Microbiology*, **71**(3):1184–1192, 2005. 11

165. Redfield R.J. **Do bacteria have sex?** *Nature Reviews Genetics*, **2**(8):634–639, 2001. 12

166. Frost L.S., Ippen-Ihler K. and Skurray R.A. **Analysis of the sequence and gene products of the transfer region of the F sex factor.** *Microbiology and Molecular Biology Reviews*, **58**(2):162–210, 1994. 11

167. Franke A.E. and Clewell D.B. **Evidence for a chromosome-borne resistance transposon (*Tn*916) in *Streptococcus faecalis* that is capable of "conjugal" transfer in the absence of a conjugative plasmid.** *Journal of Bacteriology*, **145**(1):494–502, 1981. 11

168. Salyers A.A., Shoemaker N.B., Stevens A.M. and Li L.Y. **Conjugative transposons: an unusual and diverse set of integrated gene transfer elements**. *Microbiology Reviews*, **59**(4):579–590, 1995. 11

169. Novick R.P. and Hoppensteadt F.C. **On plasmid incompatibility.** *Plasmid*, **1**(4):421–434, 1978. 13

170. Novick R.P. **Plasmid incompatibility.** *Microbiological Reviews*, **51**(4):381–395, 1987. 13

171. Garcillán-Barcia M.P. and de la Cruz F. **Why is entry exclusion an essential feature of conjugative plasmids?** *Plasmid*, **60**(1):1–18, 2008. 13

172. Pembroke J.T., MacMahon C. and McGrath B. **The role of conjugative transposons in the *Enterobacteriaceae*.** *Cellular and Molecular Life Sciences*, **59**(12):2055–2064, 2002. 13

173. Baker S., Pickard D., Whitehead S., Farrar J. and Dougan G. **Mobilization of the incQ plasmid R300B with a chromosomal conjugation system in *Salmonella enterica* serovar typhi**. *Journal of Bacteriology*, **190**(11):4084–4087, 2008. 13

174. Boyd D.A., Peters G.A., Ng L. and Mulvey M.R. **Partial characterization of a genomic island associated with the multidrug resistance region of *Salmonella enterica* Typhymurium DT104**. *FEMS Microbiology Letters*, **189**(2):285–291, 2000. 13

175. Boyd D., Peters G.A., Cloeckaert A., Boumedine K.S., Chaslus-Dancla E., Imberechts H. and Mulvey M.R. **Complete nucleotide sequence of a 43-kilobase genomic island associated with the multidrug resistance region of *Salmonella enterica* serovar Typhimurium DT104 and its identification in phage type DT120 and serovar Agona**. *Journal of Bacteriology*, **183**(19):5725–5732, 2001. 13

176. Mulvey M.R., Boyd D.A., Olson A.B., Doublet B. and Cloeckaert A. **The genetics of *Salmonella* genomic island 1**. *Microbes and Infection*, **8**(7):1915–1922, 2006. 13

177. Doublet B., Boyd D., Mulvey M.R. and Cloeckaert A. **The *Salmonella* genomic island 1 is an integrative mobilizable element**. *Molecular Microbiology*, **55**(6):1911–1924, 2005. 13

178. Rychlik I., Gregorova D. and Hradecka H. **Distribution and function of plasmids in *Salmonella enterica***. *Veterinary Microbiology*, **112**(1):1–10, 2006. 13, 14, 283

179. Ou J. **The virulence plasmids of *Salmonella* serovars typhimurium, choleraesuis, dublin, and enteritidis, and the cryptic plasmids of *Salmonella* serovars copenhagen and sendai belong to the same incompatibility group, but not those of *Salmonella* serovars durban, gallinarum, give, infantis and pullorum**. *Microbial Pathogenesis*, **8**(2):101–107, 1990. 13

180. Jones C. and Stanley J. ***Salmonella* plasmids of the pre-antibiotic era**. *Journal of General Microbiology*, **138**(1):189–197, 1992. 13, 14, 18

181. Haneda T., Okada N., Nakazawa N., Kawakami T. and Danbara H. **Complete DNA sequence and comparative analysis of the 50-kilobase virulence plasmid of *Salmonella enterica* serovar Choleraesuis.** *Infection and Immunity*, **69**(4):2612–2620, 2001. 13

182. Hong S.F., Chiu C.H., Chu C., Feng Y. and Ou J.T. **Complete nucleotide sequence of a virulence plasmid of *Salmonella enterica* serovar Dublin and its phylogenetic relationship to the virulence plasmids of serovars Choleraesuis, Enteritidis and Typhimurium.** *FEMS Microbiology Letters*, **282**(1):39–43, 2008. 13

183. Chu C. and Chiu C.H. **Evolution of the virulence plasmids of non-typhoid *Salmonella* and its association with antimicrobial resistance.** *Microbes and Infection*, **8**(7):1931–1936, 2006. 13

184. Ahmer B.M., Tran M. and Heffron F. **The virulence plasmid of *Salmonella typhimurium* is self-transmissible.** *Journal of Bacteriology*, **181**(4):1364–1368, 1999. 13

185. Barrow P.A. and Lovell M.A. **Functional homology of virulence plasmids in *Salmonella gallinarum*, *S. pullorum*, and *S. typhimurium*.** *Infection and Immunity*, **57**(10):3136–3141, 1989. 13

186. Boyd E.F. and Hartl D.L. ***Salmonella* Virulence Plasmid: Modular Acquisition of the *spv* Virulence Region by an F-Plasmid in *Salmonella enterica* Subspecies I and Insertion Into the Chromosome of Subspecies II, IIIa, IV and VII Isolates**. *Genetics*, **149**(3):1183–1190, 1998. 13

187. Gulig P.A. **Virulence plasmids of *Salmonella typhimurium* and other salmonellae.** *Microbial Pathogenesis*, **8**(1):3–11, 1990. 13

188. Matsui H., Bacot C.M., Garlington W.A., Doyle T.J., Roberts S. and Gulig P.A. **Virulence plasmid-borne *spvB* and *spvC* genes can replace the 90-kilobase plasmid in conferring virulence to *Salmonella enterica* serovar Typhimurium in subcutaneously inoculated mice**. *Journal of Bacteriology*, **183**(15):4652–4658, 2001. 13

189. Montenegro M. **Heteroduplex analysis of *Salmonella* virulence plasmids and their prevalence in isolates of defined sources**. *Microbial Pathogenesis*, **11**(6):391–397, 1991. 13

190. Parry C.M. **Antimicrobial drug resistance in *Salmonella enterica*.** *Current Opinion in Infectious Diseases*, **16**(5):467–472, 2003. 14, 27

191. Carattoli A. **Plasmid-mediated antimicrobial resistance in *Salmonella enterica*.** *Current Issues in Molecular Biology*, **5**(4):113–122, 2003. 14

192. Bennett P.M. **Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria**. *British Journal of Pharmacology*, **153**(S1):S347–S357, 2008. 14

193. Ploy M.C., Lambert T., Couty J.P. and Denis F. **Integrons: an antibiotic resistance gene capture and expression system.** *Clinical Chemistry and Laboratory Medicine*, **38**(6):483–487, 2000. 14

194. Ghosh A. **Characterization of Large Plasmids Encoding Resistance to Toxic Heavy Metals in *Salmonella abortus equi*.** *Biochemical and Biophysical Research Communications*, **272**(1):6–11, 2000. 14

195. Silver S. **Bacterial silver resistance: molecular biology and uses and misuses of silver compounds.** *FEMS Microbiology Reviews*, **27**(2-3):341–353, 2003. 14

196. Huffman G.A. and Rownd R.H. **Transition of deletion mutants of the composite resistance plasmid NR1 in *Escherichia coli* and *Salmonella typhimurium*.** *Journal of Bacteriology*, **159**(2):488–498, 1984. 14

197. Tosini F., Visca P., Luzzi I., Dionisi A.M., Pezzella C., Petrucca A. and Carattoli A. **Class 1 integron-borne multiple-antibiotic resistance carried by IncFI and IncL/M plasmids in *Salmonella enterica* serotype typhimurium.** *Antimicrobial Agents and Chemotherapy*, **42**(12):3053–3058, 1998. 14

198. Guerra B., Soto S., Helmuth R. and Mendoza M.C. **Characterization of a self-transferable plasmid from *Salmonella enterica* serotype typhimurium clinical isolates carrying two integron-borne gene cassettes together with virulence and drug resistance genes.** *Antimicrobial Agents and Chemotherapy*, **46**(9):2977–2981, 2002. 14

199. Evershed N.J., Levings R.S., Wilson N.L., Djordjevic S.P. and Hall R.M. **Unusual class 1 integron-associated gene cassette configuration found in IncA/C plasmids from *Salmonella enterica*.** *Antimicrobial Agents and Chemotherapy*, **53**(6):2640–2642, 2009. 14

200. Chu C., Chiu C.H., Wu W.Y., Chu C.H., Liu T.P. and Ou J.T. **Large Drug Resistance Virulence Plasmids of Clinical Isolates of *Salmonella enterica* Serovar Choleraesuis**. *Antimicrobial Agents and Chemotherapy*, **45**(8):2299–2303, 2001. 14

201. Villa L. and Carattoli A. **Integrons and Transposons on the *Salmonella enterica* Serovar Typhimurium Virulence Plasmid**. *Antimicrobial Agents and Chemotherapy*, **49**(3):1194–1197, 2005. 14

202. Herrero A., Rodicio M.R., González-Hevia M.A. and Mendoza M.C. **Molecular epidemiology of emergent multidrug-resistant *Salmonella enterica* serotype Typhimurium strains carrying the virulence resistance plasmid pUO-StVR2.** *The Journal of Antimicrobial Chemotherapy*, **57**(1):39–45, 2006. 14

203. Herrero A., Mendoza M.C., Threlfall E.J. and Rodicio M.R. **Detection of *Salmonella enterica* serovar Typhimurium with pUO-StVR2-like virulence-resistance hybrid plasmids in the United Kingdom.** *European Journal of Clinical Microbiology and Infectious Diseases*, 2009. 14

204. Astill D. **Characterization of the Small Cryptic Plasmid, pIMVS1, of *Salmonella enterica* ser. Typhimurium**. *Plasmid*, **30**(3):258–267, 1993. 14

205. Bernardi A. and Bernardi F. **Complete sequence of pSC101.** *Nucleic Acids Research*, **12**(24):9415–9426, 1984. 14

206. Ridley A.M., Punia P., Ward L.R., Rowe B. and Threlfall E.J. **Plasmid characterization and pulsed-field electrophoretic analysis demonstrate that ampicillin-resistant strains of *Salmonella enteritidis* phage type 6a are derived from *Salm. enteritidis* phage type 4.** *The Journal of Applied Bacteriology*, **81**(6):613–618, 1996. 14

207. Rychlik I., Sebkova A., Gregorova D. and Karpiskova R. **Low-molecular-weight plasmid of *Salmonella enterica* serovar Enteritidis codes for retron reverse transcriptase and influences phage resistance.** *Journal of Bacteriology*, **183**(9):2852–2858, 2001. 14

208. Gregorova D., Pravcova M., Karpiskova R. and Rychlik I. **Plasmid pC present in *Salmonella enterica* serovar Enteritidis PT14b strains encodes a restriction modification system.** *FEMS Microbiology Letters*, **214**(2):195–198, 2002. 14

209. Threlfall E.J., Hampton M.D., Chart H. and Rowe B. **Use of plasmid profile typing for surveillance of *Salmonella enteritidis* phage type 4 from humans, poultry and eggs**. *Epidemiology and Infection*, **112**(01):25–31, 1994. 14

210. Rychlik I. **Subdivision of *Salmonella enterica* serovar enteritidis phage types PT14b and PT21 by plasmid profiling**. *Veterinary Microbiology*, **74**(3):217–225, 2000. 14

211. Rychlik I., Svestkova A. and Karpiskova R. **Subdivision of *Salmonella enterica* serovar enteritidis phage types PT14b and PT21 by plasmid profiling.** *Veterinary Microbiology*, **74**(3):217–225, 2000. 14

212. Johnson J.E. and Chiu W. **DNA packaging and delivery machines in tailed bacteriophages.** *Current Opinion in Structural Biology*, **17**(2):237–243, 2007. 14

213. Campbell A. **The future of bacteriophage biology**. *Nature Reviews Genetics*, **4**(6):471–477, 2003. 14

214. CAMPBELL A. **Transduction and segregation in *Escherichia coli* K12.** *Virology*, **4**(2):366–384, 1957. 14

215. Plunkett G., Rose D.J., Durfee T.J. and Blattner F.R. **Sequence of Shiga Toxin 2 Phage 933W from *Escherichia coli* O157:H7: Shiga Toxin as a Phage Late-Gene Product**. *Journal of Bacteriology*, **181**(6):1767–1778, 1999. 14, 15

216. Wagner P.L. and Waldor M.K. **Bacteriophage Control of Bacterial Virulence**. *Infection and Immunity*, **70**(8):3985–3993, 2002. 15

217. FREEMAN V.J. **Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*.** *Journal of Bacteriology*, **61**(6):675–688, 1951. 15

218. McDonough M.A. and Butterton J.R. **Spontaneous tandem amplification and deletion of the shiga toxin operon in *Shigella dysenteriae* 1.** *Molecular Microbiology*, **34**(5):1058–1069, 1999. 15

219. Davis B.M., Kimsey H.H., Kane A.V. and Waldor M.K. **A satellite phage-encoded antirepressor induces repressor aggregation and cholera toxin gene transfer.** *The EMBO Journal*, **21**(16):4240–4249, 2002. 15

220. Ubukata K., Konno M. and Fujii R. **Transduction of drug resistance to tetracycline, chloramphenicol, macrolides, lincomycin and clindamycin with phages induced from *Streptococcus pyogenes*.** *The Journal of Antibiotics*, **28**(9):681–688, 1975. 15

221. Mirold S., Rabsch W., Rohde M., Stender S., Tschäpe H., Rüssmann H., Igwe E. and Hardt W.D. **Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic *Salmonella typhimurium* strain**. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(17):9845–9850, 1999. 15

222. Hardt W.D., Urlaub H. and Galán J.E. **A substrate of the centisome 63 type III protein secretion system of *Salmonella typhimurium* is encoded by a cryptic bacteriophage**. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(5):2574–2579, 1998. 15

223. Hopkins K.L. and Threlfall E.J. **Frequency and polymorphism of *sopE* in isolates of *Salmonella enterica* belonging to the ten most prevalent serotypes in England and Wales.** *Journal of Medical Microbiology*, **53**(Pt 6):539–543, 2004. 15

224. Thomson N., Baker S., Pickard D., Fookes M., Anjum M., Hamlin N., Wain J., House D., Bhutta Z., Chan K., Falkow S., Parkhill J., Woodward M., Ivens A. and Dougan G. **The role of prophage-like elements in the diversity of *Salmonella enterica* serovars**. *Journal of Molecular Biology*, **339**(2):279–300, 2004. 15, 16, 17, 78, 79, 281, 282

225. Prager R., Rabsch W., Streckel W., Voigt W., Tietze E. and Tschäpe H. **Molecular properties of *Salmonella enterica* serotype paratyphi B distinguish between its systemic and its enteric pathovars**. *Journal of Clinical Microbiology*, **41**(9):4270–4278, 2003. 15, 147, 150, 183

226. Figueroa-Bossi N. and Bossi L. **Inducible prophages contribute to *Salmonella* virulence in mice.** *Molecular Microbiology*, **33**(1):167–176, 1999. 15

227. Thomson N.R., Clayton D.J., Windhorst D., Vernikos G., Davidson S., Churcher C., Quail M.A., Stevens M., Jones M.A., Watson M., Barron A., Layton A., Pickard D., Kingsley R.A., Bignell A., Clark L., Harris B., Ormond D., Abdellah Z., Brooks K., Cherevach I., Chillingworth T., Woodward J., Norberczak H., Lord A., Arrowsmith C., Jagels K., Moule S., Mungall K., Sanders M., Whitehead S., Chabalgoity J.A., Maskell D., Humphrey T., Roberts M., Barrow P.A., Dougan G. and Parkhill J. **Comparative genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways.** *Genome Research*, **18**(10):1624–1637, 2008. 15, 37

228. Lindberg A.A. and Hellerqvist C.G. **Bacteriophage attachment sites, serological specificity, and chemical composition of the lipopolysaccharides of semirough and rough mutants of *Salmonella typhimurium*.** *Journal of Bacteriology*, **105**(1):57–64, 1971. 15

229. Lindberg A.A. and Holme T. **Influence of O side chains on the attachment of the Felix O-1 bacteriophage to *Salmonella* bacteria.** *Journal of Bacteriology*, **99**(2):513–519, 1969. 15

230. Datta D.B., Arden B. and Henning U. **Major proteins of the *Escherichia coli* outer cell envelope membrane as bacteriophage receptors.** *Journal of Bacteriology*, **131**(3):821–829, 1977. 15

231. Pelludat C., Mirold S. and Hardt W.D. **The SopEPhi phage integrates into the *ssrA* gene of *Salmonella enterica* serovar Typhimurium A36 and is closely related to the Fels-2 prophage.** *Journal of Bacteriology*, **185**(17):5182–5191, 2003. 15

232. Bishop A.L., Baker S., Jenks S., Fookes M., Gaora P.O., Pickard D., Anjum M., Farrar J., Hien T.T., Ivens A. and Dougan G. **Analysis of the hypervariable region of the *Salmonella enterica* genome associated with tRNA(leuX)**. *Journal of Bacteriology*, **187**(7):2469–2482, 2005. 15, 79

233. Campbell A., Schneider S.J. and Song B. **Lambdoid phages as elements of bacterial genomes (integrase/phage21/*Escherichia coli* K-12/icd gene).** *Genetica*, **86**(1-3):259–267, 1992. 15

234. Campbell A.M. **Preferential orientation of natural lambdoid prophages and bacterial chromosome organization.** *Theoretical Population Biology*, **61**(4):503–507, 2002. 15

235. Pickard D., Thomson N.R., Baker S., Wain J., Pardo M., Goulding D., Hamlin N., Choudhary J., Threfall J. and Dougan G. **Molecular Characterization of the *Salmonella enterica* Serovar Typhi Vi-Typing Bacteriophage E1**. *Journal of Bacteriology*, **190**(7):2580–2587, 2008. 15

236. Echols H., Pilarski L. and Xcheng P.Y. **In vitro repression of phage lambda DNA transcription by a partially purified repressor from lysogenic cells.** *Proceedings of the National Academy of Sciences of the United States of America*, **59**(3):1016–1023, 1968. 15

237. Ranquet C., Toussaint A., de Jong H., Maenhaut-Michel G. and Geiselmann J. **Control of Bacteriophage Mu Lysogenic Repression**. *Journal of Molecular Biology*, **353**(1):186–195, 2005. 15

238. Cooke F.J., Wain J., Fookes M., Ivens A., Thomson N., Brown D.J., Threlfall E.J., Gunn G., Foster G. and Dougan G. **Prophage sequences defining hot spots of genome variation in *Salmonella enterica* serovar Typhimurium can be used to discriminate between field isolates.** *Journal of Clinical Microbiology*, **45**(8):2590–2598, 2007. 15

239. Rychlik I., Hradecka H. and Malcova M. ***Salmonella enterica* serovar Typhimurium typing by prophage-specific PCR.** *Microbiology*, **154**(5):1384–1389, 2008. 15

240. Nicolle P., Prunet J. and Diverneau G. **Phage typing of the typhus bacillus. I. General and technical considerations.** *Revue d'Hygiène et de Médecine Sociale*, **12**, 1964. 15

241. Threlfall E.J. and Frost J.A. **The identification, typing and fingerprinting of *Salmonella*: laboratory aspects and epidemiological applications.** *The Journal of Applied Bacteriology*, **68**(1):5–16, 1990. 15

242. Brown E.W., Mammel M.K., Leclerc J.E. and Cebula T.A. **Limited boundaries for extensive horizontal gene transfer among *Salmonella* pathogens**. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(26):15676–15681, 2003. 15

243. Kotetishvili M., Stine O.C., Kreger A., Morris J.G. and Sulakvelidze A. **Multilocus sequence typing for characterization of clinical and environmental salmonella strains**. *Journal of Clinical Microbiology*, **40**(5):1626–1635, 2002. 15

244. McQuiston J.R., Herrera-Leon S., Wertheim B.C., Doyle J., Fields P.I., Tauxe R.V. and Logsdon J.M. **Molecular phylogeny of the salmonellae: relationships among *Salmonella* species and subspecies determined from four housekeeping genes and evidence of lateral gene transfer events**. *Journal of Bacteriology*, **190**(21):7060–7067, 2008. 15

245. Octavia S. and Lan R. **Frequent recombination and low level of clonality within *Salmonella enterica* subspecies I**. *Microbiology*, **152**(4):1099–1108, 2006. 15, 31, 283

246. Parry C.M. ***Salmonella* Infections: Clinical, Immunological and Molecular Aspects**, chapter 1. Cambridge University Press, 2006. 16, 20, 21, 23

247. Baker S., Sarwar Y., Aziz H., Haque A., Ali A., Dougan G., Wain J. and Haque A. **Detection of Vi-negative *Salmonella enterica* serovar typhi in the peripheral blood of patients with typhoid fever in the Faisalabad region of Pakistan.** *Journal of Clinical Microbiology*, **43**(9):4418–4425, 2005. 16

248. Guinée P.A., Jansen W.H., Maas H.M., Le Minor L. and Beaud R. **An unusual H antigen (Z66) in strains of *Salmonella typhi*.** *Annales de Microbiologie*, **132**(3):331–334, 1981. 16, 275

249. Vieu J.F., Binette H. and Lehérissey M. **Absence of the antigen H:z66 in 2355 strains of *Salmonella typhi* from Madagascar and several countries of tropical Africa**. *Bulletin de la Société de Pathologie Exotique et de ses Filiales*, **79**(1):22–26, 1986. 16

250. Vieu J.F. and Lehérissey M. **The antigen H:z66 in 1,000 strains of Salmonella typhi from the Antilles, Central America and South America**. *Bulletin de la Société de Pathologie Exotique et de ses Filiales*, **81**(2):198–201, 1988. 16

251. Tamura K., Sakazaki R., Kuramochi S. and Nakamura A. **Occurrence of H-antigen Z66 of R phase in cultures of *Salmonella* serovar typhi originated from Indonesia.** *Epidemiology and Infection*, **101**(2):311–314, 1988. 16, 275

252. Moshitch S., Doll L., Rubinfeld B.Z., Stocker B.A., Schoolnik G.K., Gafni Y. and Frankel G. **Mono- and biphasic *Salmonella typhi*: genetic homogeneity and distinguishing characteristics.** *Molecular Microbiology*, **6**(18):2589–2597, 1992. 16, 275

253. Frankel G., Newton S.M., Schoolnik G.K. and Stocker B.A. **Intragenic recombination in a flagellin gene: characterization of the H1-j gene of *Salmonella typhi*.** *The EMBO Journal*, **8**(10):3149–3152, 1989. 16, 275

254. Huang X., Phung l.e.V., Dejsirilert S., Tishyadhigama P., Li Y., Liu H., Hirose K., Kawamura Y. and Ezaki T. **Cloning and characterization of the gene encoding the z66 antigen of *Salmonella enterica* serovar Typhi**. *FEMS Microbiology Letters*, **234**(2):239–246, 2004. 16

255. Baker S., Hardy J., Sanderson K.E., Quail M., Goodhead I., Kingsley R.A., Parkhill J., Stocker B. and Dougan G. **A Novel Linear Plasmid Mediates Flagellar Variation in *Salmonella* Typhi**. *PLoS Pathogens*, **3**(5):e59, 2007. 16, 80

256. Baker S., Holt K., van de Vosse E., Roumagnac P., Whitehead S., King E., Ewels P., Keniry A., Weill F.X., Lightfoot D., van Dissel J.T., Sanderson K.E., Farrar J., Achtman M., Deloukas P. and Dougan G. **High-Throughput Genotyping of *Salmonella* Typhi Allows Geographical Assignment of Haplotypes and Pathotypes within an Urban District of Jakarta, Indonesia**. *Journal of Cinical Microbiology*, **46**(5):1741–1746, 2008. 16, 42, 48, 49, 80, 215, 217, 243, 272, 273, 275

257. Liu S.L. and Sanderson K.E. **Rearrangements in the genome of the bacterium *Salmonella typhi***. *Proceedings of the National Academy of Sciences of the United States of America*, **92**(4):1018–1022, 1995. 18

258. Liu S.L. and Sanderson K.E. **Highly plastic chromosomal organization in *Salmonella typhi***. *Proceedings of the National Academy of Sciences of the United States of America*, **93**(19):10303–10308, 1996. 18

259. Liu G., Liu W.Q., Johnston R.N., Sanderson K.E., Li S.X. and Liu S.L. **Genome Plasticity and ori-ter Rebalancing in *Salmonella typhi***. *Molecular Biology and Evolution*, **23**(2):365–371, 2006. 18, 48

260. Liu S.L. and Sanderson K.E. **The chromosome of *Salmonella paratyphi A* is inverted by recombination between *rrnH* and *rrnG***. *Journal of Bacteriology*, **177**(22):6585–6592, 1995. 18, 19

261. Liu S.L. and Sanderson K.E. **I-CeuI reveals conservation of the genome of independent strains of *Salmonella typhimurium***. *Journal of Bacteriology*, **177**(11):3355–3357, 1995. 18

262. Liu S.L. and Sanderson K.E. **Homologous recombination between *rrn* operons rearranges the chromosome in host-specialized species of Salmonella**. *FEMS Microbiology Letters*, **164**(2):275–281, 1998. 18

263. Andersson J.O. and Andersson S.G.E. **Genome degradation is an ongoing process in Rickettsia**. *Molcular Biology and Evolution*, **16**(9):1178–1191, 1999. 18, 149

264. Cole S.T., Eiglmeier K., Parkhill J., James K.D., Thomson N.R., Wheeler P.R., Honore N., Garnier T., Churcher C., Harris D., Mungall K., Basham D., Brown D., Chillingworth T., Connor R., Davies R.M., Devlin K., Duthoy S., Feltwell T., Fraser A., Hamlin N., Holroyd S., Hornsby T., Jagels K., Lacroix C., Maclean J., Moule S., Murphy L., Oliver K., Quail M.A., Rajandream M.A., Rutherford K.M., Rutter S., Seeger K., Simon S., Simmonds M., Skelton J., Squares R., Squares S., Stevens K., Taylor K., Whitehead S., Woodward J.R. and Barrell B.G. **Massive gene decay in the leprosy bacillus**. *Nature*, **409**(6823):1007–1011, 2001. 18, 37, 149

265. Parkhill J., Sebaihia M., Preston A., Murphy L.D., Thomson N., Harris D.E., Holden M.T., Churcher C.M., Bentley S.D., Mungall K.L., Cerdeno-Tarraga A.M., Temple L., James K., Harris B., Quail M.A., Achtman M., Atkin R., Baker S., Basham D., Bason N., Cherevach I., Chillingworth T., Collins M., Cronin A., Davis P., Doggett J., Feltwell T., Goble A., Hamlin N., Hauser H., Holroyd S., Jagels K., Leather S., Moule S., Norberczak H., O'Neil S., Ormond D., Price C., Rabbinowitsch E., Rutter S., Sanders M., Saunders D., Seeger K., Sharp S., Simmonds M., Skelton J., Squares R., Squares S., Stevens K., Unwin L., Whitehead S., Barrell B.G. and Maskell D.J. **Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica***. *Nature Genetics*, **35**(1):32–40, 2003. 18, 37, 149

266. R. N., Howard S., Wren B.W., Holden M.T., Crossman L., Challis G.L., Churcher C., Mungall K., Brooks K., Chillingworth T., Feltwell T., Abdellah Z., Hauser H., Jagels K., Maddison M., Moule S., Sanders M., Whitehead S., Quail M.A., Dougan G., Parkhill J. and Prentice M.B. **The complete genome sequence and comparative genome analysis of the high pathogenicity *Yersinia enterocolitica* strain 8081**. *PLoS Genetics*, **2**(12):e206, 2006. 18, 36, 37, 83, 149

267. Riley M., Abe T., Arnaud M.B., Berlyn M.K., Blattner F.R., Chaudhuri R.R., Glasner J.D., Horiuchi T., Keseler I.M., Kosuge T., Mori H., Perna N.T., Plunkett G., Rudd K.E., Serres M.H., Thomas G.H., Thomson N.R., Wishart D. and Wanner B.L. ***Escherichia coli* K-12: a cooperatively developed annotation snapshot-2005**. *Nucleic Acids Research*, **34**(1):1–9, 2006. 18

268. Anderson E.S., Humphreys G.O. and Willshaw G.A. **The molecular relatedness of R factors in enterobacteria of human and animal origin.** *Journal of General Microbiology*, **91**(2):376–382, 1975. 18, 187, 188, 212

269. Taylor D.E. and Brose E.C. **Characterization of incompatibility group HI1 plasmids from *Salmonella typhi* by restriction endonuclease digestion and hybridization of DNA probes for Tn3, Tn9, and Tn10.** *Canadian Journal of Microbiology*, **31**(8), 1985. 18, 188, 189

270. Taylor D.E., Chumpitaz J.C. and Goldstein F. **Variability of IncHI1 plasmids from *Salmonella typhi* with special reference to Peruvian plasmids encoding resistance to trimethoprim and other antibiotics.** *Antimicrobial Agents and Chemotherapy*, **28**(3):452–455, 1985. 18, 188

271. Fica A., Fernandez-Beros M.E., Aron-Hott L., Rivas A., D'Ottone K., Chumpitaz J., Guevara J.M., Rodriguez M. and Cabello F. **Antibiotic-resistant *Salmonella typhi* from two outbreaks: few ribotypes and IS200 types harbor Inc HI1 plasmids.** *Microbial Drug Resistance*, **3**(4):339–343, 1997. 18, 27, 188

272. Shanahan P.M., Jesudason M.V., Thomson C.J. and Amyes S.G. **Molecular Analysis of and Identification of Antibiotic Resistance Genes in Clinical Isolates of *Salmonella typhi* from India**. *Journal of Clinical Microbiology*, **36**(6):1595–1600, 1998. 18, 188

273. Hermans P.W., Saha S.K., van Leeuwen W.J., Verbrugh H.A., van Belkum A. and Goessens W.H. **Molecular typing of *Salmonella typhi* strains from Dhaka (Bangladesh) and development of DNA probes identifying plasmid-encoded multidrug- resistant isolates**. *Journal of Clinical Microbiology*, **34**(6):1373–1379, 1996. 18, 188

274. Shanahan P.M., Karamat K.A., Thomson C.J. and Amyes S.G. **Characterization of multi-drug resistant *Salmonella typhi* isolated from Pakistan.** *Epidemiology and Infection*, **124**(1):9–16, 2000. 18, 188

275. Wain J., Nga L.T.D., Kidgell C., James K., Fortune S., Diep T.S., Ali T., Gaora P.O., Parry C., Parkhill J., Farrar J., White N.J. and Dougan G. **Molecular analysis of incHI1 antimicrobial resistance plasmids from *Salmonella* serovar Typhi strains associated with typhoid fever**. *Antimicrobial Agents and Chemotherapy*, **47**(9):2732–2739, 2003. 18, 47, 80, 188, 207, 210, 278

276. Kariuki S., Revathi G., Muyodi J., Mwituria J., Munyalo A., Mirza S. and Hart C.A. **Characterization of multidrug-resistant typhoid outbreaks in Kenya**. *Journal of Clinical Microbiology*, **42**(4):1477–1482, 2004. 18, 80, 188, 278

277. Mirza S., Kariuki S., Mamun K.Z., Beeching N.J. and Hart C.A. **Analysis of Plasmid and Chromosomal DNA of Multidrug-Resistant *Salmonella enterica* Serovar Typhi from Asia**. *Journal of Clinical Microbiology*, **38**(4):1449–1452, 2000. 18, 188, 278

278. Kidgell C., Pickard D., Wain J., James K., Nga L.T.D., Diep T.S., Levine M.M., O'Gaora P., Prentice M.B., Parkhill J., Day N., Farrar J. and Dougan G. **Characterisation and distribution of a cryptic *Salmonella typhi* plasmid pHCM2**. *Plasmid*, **47**(3):159–171, 2002. 18, 80, 243

279. Maskey A.P., Day J.N., Phung Q.T., Thwaites G.E., Campbell J.I., Zimmerman M., Farrar J.J. and Basnyat B. ***Salmonella enterica* serovar Paratyphi A and S. enterica serovar Typhi cause indistinguishable clinical syndromes in Kathmandu, Nepal**. *Clinical Infectious Diseases*, **42**(9):1247–1253, 2006. 18, 25, 28, 146, 188, 213, 275, 278

280. Fangtham M. and Wilde H. **Emergence of *Salmonella* paratyphi A as a major cause of enteric fever: need for early detection, preventive measures, and effective vaccines.** *Journal of Travel Medicine*, **15**(5):344–350, 2008. 18, 30, 146, 187

281. Janda J.M. **Enterobacteria**. American Society for Microbiology, 2nd edition edition, 2006. 18

282. Paramasivan C.N., Subramanian S. and Shanmugasundaram N. **Antimicrobial resistance and incidence of R factor among *Salmonella* isolated from patients with enteric fever and other clinical conditions in Madras, India (1975-1976)**. *The Journal of Infectious Diseases*, **136**(6):796–800, 1977. 20, 26, 27

283. Holt K.E., Thomson N.R., Wain J., Minh D.P., Nair S., Hasan R., Bhutta Z.A., Quail M.A., Norbertczak H., Walker D., Dougan G. and Parkhill J. **Multidrug-resistant *Salmonella enterica* serovar Paratyphi A harbors IncHI1 plasmids similar to those found in serovar Typhi**. *Journal of Bacteriology*, **189**(11):4257–4264, 2007. 20, 80, 188, 209

284. Hasan Z., Rahman K.M., Alam M.N., Afroza A., Asna Z.H., Ghosh P.K. and Alam N. **Role of a large plasmid in mediation of multiple drug resistance in *Salmonella* typhi and paratyphi A in Bangladesh**. *Bangladesh Medical Research Council Bulletin*, **21**(1):50–54, 1995. 20, 188, 209

285. Zhang Z.K., Huang Y.N., Guo B.C., Deng M.L., Yuan R.Z. and Wang Q.S. **Surveillance of the antibiotic resistance and plasmid of *Salmonella* paratyphoid**. *Chinese Journal of Antibiotics*, **29**(10):610, 2004. 20, 188, 209

286. Mandal S., Mandal M.D. and Pal N.K. **Antibiotic resistance of *Salmonella enterica* serovar Paratyphi A in India: Emerging and reemerging problem**. *Journal of Postgraduate Medicine*, **52**(3):163–166, 2006. 20, 27, 188

287. Huang H., Li J., Yang X., Wang Y., Wang Y., Tao J., Huang Y. and Zhang X. **Sequence Analysis of the Plasmid pGY1 Harbored in *Salmonella enterica* Serovar Paratyphi A**. *Biochemical Genetics*, **47**(3-4):191–197, 2009. 20, 105, 137

288. Panigrahi D., Chugh T.D., West P.W., Dimitrov T.Z., Groover S. and Mehta G. **Antimicrobial susceptibility, phage typing and plasmid profile of *Salmonella enterica* serotype paratyphi A strains isolated in Kuwait**. *Medical Principles and Practice*, **12**(4):252–255, 2003. 20

289. Tam M.A., Rydström A., Sundquist M. and Wick M.J. **Early cellular responses to *Salmonella* infection: dendritic cells, monocytes, and more**. *Immunological Reviews*, **225**(1):140–162, 2008. 20, 21

290. House D., Bishop A., Parry C., Dougan G. and Wain J. **Typhoid fever: pathogenesis and disease**. *Current Opinion in Infectious Diseases*, **14**(5):573–578, 2001. 21, 148

291. Wain J., Diep T.S., Ho V.A., Walsh A.M., Hoa N.T., Parry C.M. and White N.J. **Quantitation of Bacteria in Blood of Typhoid Fever Patients and Relationship between Counts and Clinical Features, Transmissibility, and Antibiotic Resistance**. *Journal of Cliical Microbiology*, **36**(6):1683–1687, 1998. 21

292. Wain J., Bay P.V., Vinh H., Duong N.M., Diep T.S., Walsh A.L., Parry C.M., Hasserjian R.P., Ho V.A., Hien T.T., Farrar J., White N.J. and Day N.P.J. **Quantitation of Bacteria in Bone Marrow from Patients with Typhoid Fever: Relationship between Counts and Clinical Features**. *Journal of Clinical Microbiology*, **39**(4):1571–1576, 2001. 21

293. Hathout S.E., El-Ghaffar Y.A., Awny A.Y. and Hassan K. **Relation between Urinary Schistosomiasis and Chronic Enteric Urinary Carrier State Among Egyptians**. *The American Journal of Tropical Medicine and Hygiene*, **15**(2):156–161, 1966. 21, 23

294. Vallenas C., Hernandez H., Kay B., Black R. and Gotuzzo E. **Efficacy of bone marrow, blood, stool and duodenal contents cultures for bacteriologic confirmation of typhoid fever in children.** *Pediatric Infectious Disease*, **4**(5):496–498, 1985. 21, 23

295. Gilman R.H., Terminel M., Levine M.M., Hernandez-Mendoza P. and Hornick R.B. **Relative efficacy of blood, urine, rectal swab, bone-marrow, and rose-spot cultures for recovery of *Salmonella typhi* in typhoid fever.** *The Lancet*, **1**(7918):1211–1213, 1975. 21, 23, 280

296. Haraga A., Ohlson M.B. and Miller S.I. **Salmonellae interplay with host cells**. *Nature Reviews Microbiology*, **6**(1):53–66, 2008. 21

297. Department of Vaccines and Biologicals, World Health Organization. **Background document: The diagnosis, treatment and prevention of typhoid fever**. Technical report, World Health Organization, 2003. 22, 23, 27, 28, 187, 213

298. van Basten J.P. and Stockenbrügger R. **Typhoid perforation. A review of the literature since 1960.** *Tropical and Geographical Medicine*, **46**(6):336–339, 1994. 22

299. Huang D. and Dupont H. **Problem pathogens: extra-intestinal complications of serotype Typhi infection**. *The Lancet Infectious Diseases*, **5**(6):341–348, 2005. 22

300. Hue N.T., Lanh M.N., Phuong L.T., Vinh H., Chinh N.T., Hien T.T., Hieu N.T., Farrar J.J. and Dunstan S.J. **Toll-Like Receptor 4 (TLR4) and Typhoid Fever in Vietnam**. *PLoS ONE*, **4**(3):e4800+, 2009. 22

301. Dunstan S.J., Stephens H.A., Blackwell J.M., Duc C.M., Lanh M.N., Dudbridge F., Phuong C..X..T., Luxemburger C., Wain J., Ho V.A., Hien T.t., Farrar J. and Dougan G. **Genes of the Class II and Class III Major Histocompatibility Complex Are Associated with Typhoid Fever in Vietnam**. *The Journal of Infectious Diseases*, **183**(2):261–268, 2001. 22

302. Dunstan S., Hue N., Rockett K., Forton J., Morris A., Diakite M., Lanh M., Phuong L., House D., Parry C., Vinh H., Hieu N., Dougan G., Hien T., Kwiatowski D. and Farrar J. **A TNF region haplotype offers protection from typhoid fever in Vietnamese patients**. *Human Genetics*, **122**(1):51–61, 2007. 22

303. Dunstan S.J., Ho V.A., Duc C.M., Lanh M.N., Phuong C.x.t., Luxemburger C., Wain J., Dudbridge F., Peacock C.s., House D., Parry C., Hien T.t., Dougan G., Farrar J. and Blackwell J.m. **Typhoid Fever and Genetic Polymorphisms at the Natural ResistanceAssociated Macrophage Protein 1**. *The Journal of Infectious Diseases*, **183**(7):1156–1160, 2001. 22

304. Dunstan S.J., Hawn T.r., Hue N.t., Parry C.p., Ho V.a., Vinh H., Diep T.s., House D., Wain J., Aderem A., Hien T.t. and Farrar J.j. **Host Susceptibility and Clinical Outcomes in Toll-Like Receptor 5-Deficient Patients with Typhoid Fever in Vietnam**. *The Journal of Infectious Diseases*, **191**(7):1068–1071, 2005. 22

305. Ali S., Vollaard A.M., Kremer D., de Visser A.W., Martina C.A.E., Widjaja S., Surjadi C., Slagboom E., van de Vosse E. and van Dissel J.T. **Polymorphisms in Proinflammatory Genes and Susceptibility to Typhoid Fever and Paratyphoid Fever**. *Journal of Interferon and Cytokine Research*, **27**(4):271–280, 2007. 22

306. Ali S., Vollaard A.M., Widjaja S., Surjadi C., van de Vosse E. and van Dissel J.T. **PARK2/PACRG polymorphisms and susceptibility to typhoid and paratyphoid fever**. *Clinical and Experimental Immunology*, **144**(3):425–431, 2006. 22

307. Stokes A. **A search for typhoid carriers among 800 convalescents.** *The Lancet*, **187**(4828):566–569, 1916. 22

308. Unattributed. **The problem of the typhoid carrier.** *The Lancet*, **182**(4698):821–822, 1913. 22

309. Levine M.M., Black R.E. and Lanata C. **Precise estimation of the number of chronic carriers of *Salmonella typhi* in Santiago, Chile, an endemic area**. *The Journal of Infectious Diseases*, **146**(6):724–726, 1982. 22, 95, 146, 148

310. Devi S. and Murray C.J. ***Salmonella* carriage rate amongst school children–a three year study.** *The Southeast Asian Journal of Tropical Medicine and Public Health*, **22**(3):357–361, 1991. 22

311. Khatri N.S., Maskey P., Poudel S., Jaiswal V.K., Karkey A., Koirala S., Shakya N., Agrawal K., Arjyal A., Basnyat B., Day J., Farrar J., Dolecek C. and Baker S. **Gallbladder carriage of *Salmonella* paratyphi A may be an important factor in the increasing incidence of this infection in South Asia.** *Annals of Internal Medicine*, **150**:567–568, 2009. 22, 23, 146, 148

312. Caygill C. **Cancer mortality in chronic typhoid and paratyphoid carriers**. *The Lancet*, **343**(8889):83–84, 1994. 23

313. Nath G., Singh H. and Shukla V.K. **Chronic typhoid carriage and carcinoma of the gallbladder.** *European Journal of Cancer Prevention*, **6**(6):557–559, 1997. 23, 146, 148

314. Dutta U., Garg P.K., Kumar R. and Tandon R.K. **Typhoid carriers among patients with gallstones are at increased risk for carcinoma of the gallbladder.** *The American Journal of Gastroenterology*, **95**(3):784–787, 2000. 23

315. Lanata C.F., Levine M.M., Ristori C., Black R.E., Jimenez L., Salcedo M., Garcia J. and Sotomayor V. **Vi serology in detection of chronic *Salmonella typhi* carriers in an endemic area.** *The Lancet*, **2**(8347):441–443, 1983. 23

316. Ferreccio C., Levine M., Astroza L., Berrios G., Solari V., Misraji A. and Pefaur C. **The detection of chronic *Salmonella typhi* carriers: a practical method applied to food handlers**. *Revista Médica de Chile*, **118**(1):33–37, 1990. 23

317. Vaishnavi C., Kochhar R., Singh G., Kumar S., Singh S. and Singh K. **Epidemiology of typhoid carriers among blood donors and patients with biliary, gastrointestinal and other related diseases.** *Microbiology and Immunology*, **49**(2):107–112, 2005. 23

318. Bhutta Z.A. **Current concepts in the diagnosis and treatment of typhoid fever**. *British Medical Journal*, **333**(7558):78–82, 2006. 23

319. Hoffman S.L., Edman D.C., Punjabi N.H., Lesmana M., Cholid A., Sundah S. and Harahap J. **Bone Marrow Aspirate Culture Superior to Streptokinase Clot Culture and 8 ml 1:10 Blood-to-Broth Ratio Blood Culture for Diagnosis of Typhoid Fever**. *The American Journal of Tropical Medicine and Hygiene*, **35**(4):836–839, 1986. 23

320. Olopoenia L.A. and King A.L. **Classic methods revisited: Widal agglutination test - 100 years later: still plagued by controversy**. *Postgraduate Medical Journal*, **76**(892):80–84, 2000. 23

321. Reynolds D.W., Carpenter R.L. and Simon W.H. **Diagnostic Specificity of Widal's Reaction for Typhoid Fever**. *Journal of the American Medical Association*, **214**(12):2192–2193, 1970. 23

322. Hatta M. and Smits H.L. **Detection of *Salmonella typhi* by nested polymerase chain reaction in blood, urine, and stool samples.** *The American Journal of Tropical Medicine and Hygiene*, **76**(1):139–143, 2007. 23

323. Prakash P., Mishra O.P., Singh A.K., Gulati A.K. and Nath G. **Evaluation of Nested PCR in Diagnosis of Typhoid Fever**. *Journal of Clinical Microbiology*, **43**(1):431–432, 2005. 23

324. Hirose K., Itoh K.I., Nakajima H., Kurazono T., Yamaguchi M., Moriya K., Ezaki T., Kawamura Y., Tamura K. and Watanabe H. **Selective Amplification of *tyv* (*rfbE*), *prt* (*rfbS*), *viaB*, and *fliC* Genes by Multiplex PCR for Identification of *Salmonella enterica* Serovars Typhi and Paratyphi A**. *Journal of Clinical Microbiology*, **40**(2):633–636, 2002. 23

325. Song J.H., Cho H., Park M.Y., Na D.S., Moon H.B. and Pai C.H. **Detection of Salmonella typhi in the blood of patients with typhoid fever by polymerase chain reaction.** *Journal of Clinical Microbiology*, **31**(6):1439–1443, 1993. 23

326. Levy H., Diallo S., Tennant S.M., Livio S., Sow S.O., Tapia M., Fields P.I., Mikoleit M., Tamboura B., Kotloff K.L., Lagos R., Nataro J.P., Galen J.E. and Levine M.M. **PCR Method To Identify *Salmonella enterica* Serovars Typhi, Paratyphi A, and Paratyphi B among Salmonella Isolates from the Blood of Patients with Clinical Enteric Fever**. *Journal of Clinical Microbiology*, **46**(5):1861–1866, 2008. 23

327. Ochiai R.L., Acosta C.J., Danovaro-Holliday M.C., Baiqing D., Bhattacharya S.K., Agtini M.D., Bhutta Z.A., do G.C., Ali M., Shin S., Wain J., Page A.L., Albert M.J., Farrar J., Abu-Elyazeed R., Pang T., Galindo C.M., von Seidlein L., Clemens J.D. and Group D.T.S. **A study of typhoid fever in five Asian countries: disease burden and implications for controls**. *Bulletin of the World Health Organization*, **86**(4):260–268, 2008. 24, 25, 28, 187, 218, 254, 275

328. Lin F.Y., Vo A.H., Phan V.B., Nguyen T.T., Bryla D., Tran C.T., Ha B.K., Dang D.T. and Robbins J.B. **The epidemiology of typhoid fever in the Dong Thap Province, Mekong Delta region of Vietnam.** *The American Journal of Tropical Medicine and Hygiene*, **62**(5):644–648, 2000. 24, 218, 243

329. Connor B. and Schwartz E. **Typhoid and paratyphoid fever in travellers**. *The Lancet Infectious Diseases*, **5**(10):623–628, 2005. 24, 25, 26

330. Health Protection Agency. ***Salmonella* Typhi and *Salmonella* Paratyphi Laboratory reports (cases only)**. Online, http://www.hpa.org.uk/web/HPAweb&HPAwebStandard/HPAweb_C/1195733753804. 24, 25

331. Lester A., Mygind O., Jensen K.T., Jarlov J.O. and Schonheyder H.C. **Typhoid and paratyphoid fever in Denmark 1986-1990. Epidemiologic aspects and the extent of bacteriological follow-up of patients**. *Ugeskr Laeger*, **156**(25):3770–3775, 1994. 24

332. Mermin J.H., Townes J.M., Gerber M., Dolan N., Mintz E.D. and Tauxe R.V. **Typhoid fever in the United States, 1985-1994: changing risks of international travel and increasing antimicrobial resistance**. *Archives of Internal Medicine*, **158**(6):633–638, 1998. 24

333. Yew F.S., Goh K.T. and Lim Y.S. **Epidemiology of typhoid fever in Singapore**. *Epidemiology and Infection*, **110**(01):63–70, 1993. 24

334. Gupta S.k., Medalla F., Omondi M.w., Whichard J.m., Fields P.i., Gernersmidt P., Patel N.j., Cooper K.l., Chiller T.m. and Mintz E.d. **LaboratoryBased Surveillance of Paratyphoid Fever in the United States: Travel and Antimicrobial Resistance**. *Clinical Infectious Diseases*, **46**(11):1656–1663, 2008. 24

335. Ochiai R.L., Wang X., von Seidlein L., Yang J., Bhutta Z.A., Bhattacharya S.K., Agtini M., Deen J.L., Wain J., Kim D.R., Ali M., Acosta C.J., Jodar L. and Clemens J.D. ***Salmonella* Paratyphi A Rates, Asia**. *Emerging Infectious Diseases*, **11**(11):1764–1766, 2005. 24, 146, 188

336. Maskey A.P., Basnyat B., Thwaites G.E., Campbell J.I., Farrar J.J. and Zimmerman M.D. **Emerging trends in enteric fever in Nepal: 9124 cases confirmed by blood culture 1993-2003**. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **102**(1):91–95, 2008. 24, 146, 188, 250

337. Shlim D.R., Schwartz E. and Eaton M. **Clinical Importance of *Salmonella* Paratyphi A Infection to Enteric Fever in Nepal**. *Journal of Travel Medicine*, **2**(3):165–168, 1995. 24

338. Woods C.W., Murdoch D.R., Zimmerman M.D., Glover W.A., Basnyat B., Wolf L., Belbase R.H. and Reller L.B. **Emergence of *Salmonella enterica* serotype Paratyphi A as a major cause of enteric fever in Kathmandu, Nepal**. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **100**(11):1063–1067, 2006. 24, 26, 28, 146

339. Palit A., Ghosh S., Dutta S., Sur D., Bhattacharya M.K. and Bhattacharya S.K. **Increasing prevalence of *Salmonella enterica* serotype Paratyphi-A in patients with enteric fever in a periurban slum setting of Kolkata, India**. *International Journal of Environmental Health Research*, **16**(6):455–459, 2006. 24

340. Vidyalakshmi K., Yashavanth R., Chakrapani M., Shrikala B., Bharathi B., Suchitra U., Dhanashree B. and Dominic R.M. **Epidemiological shift, seasonal variation and antimicrobial susceptibility patterns among enteric fever pathogens in South India**. *Tropical Doctor*, **38**(2):89–91, 2008. 24

341. Gupta V., Kaur J. and Chander J. **An increase in enteric fever cases due to *Salmonella* Paratyphi A in and around Chandigarh.** *The Indian Journal of Medical Research*, **129**(1):95–98, 2009. 24

342. Vollaard A.M., Ali S., van Asten H.A., Widjaja S., Visser L.G., Surjadi C. and van Dissel J.T. **Risk factors for typhoid and paratyphoid fever in Jakarta, Indonesia.** *Journal of the American Medical Association*, **291**(21):2607–2615, 2004. 25, 26, 146

343. Sur D., Ali M., von Seidlein L., Manna B., Deen J.L., Acosta C.J., Clemens J.D. and Bhattacharya S.K. **Comparisons of predictors for typhoid and paratyphoid fever in Kolkata, India.** *BMC Public Health*, **7**:289+, 2007. 25, 216, 218, 254, 277, 278

344. Sharma P.K., Ramakrishnan R., Hutin Y., Manickam P. and Gupte M.D. **Risk factors for typhoid in Darjeeling, West Bengal, India: evidence for practical action**. *Tropical Medicine and International Health*, **14**(6):696–702. 25, 26

345. Tran H.H., Bjune G., Nguyen B.M., Rottingen J.A., Grais R.F. and Guerin P.J. **Risk factors associated with typhoid fever in Son La province, northern Vietnam.** *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **99**(11):819–826, 2005. 25, 26

346. Hosoglu S., Celen M.K., Geyik M.F., Akalin S., Ayaz C., Acemoglu H. and Loeb M. **Risk factors for typhoid fever among adult patients in Diyarbakir, Turkey.** *Epidemiology and Infection*, **134**(3):612–616, 2006. 25, 26

347. Gasem M.H., Dolmans W.M., Keuter M.M. and Djokomoeljanto R.R. **Poor food hygiene and housing as risk factors for typhoid fever in Semarang, Indonesia.** *Tropical Medicine and International Health*, **6**(6):484–490, 2001. 25, 26

348. Ram P.K., Naheed A., Brooks W.A., Hossain M.A., Mintz E.D., Breiman R.F. and Luby S.P. **Risk factors for typhoid fever in a slum in Dhaka, Bangladesh.** *Epidemiology and Infection*, **135**(3):458–465, 2007. 25

349. Kelly-Hope L.A., Alonso W.J., Thiem V.D., Anh D.D., Canh d.o..G., Lee H., Smith D.L. and Miller M.A. **Geographical distribution and risk factors associated with enteric diseases in Vietnam.** *The American Journal of Tropical Medicine and Hygiene*, **76**(4):706–712, 2007. 25, 26, 243, 278

350. Srikantiah P., Vafokulov S., Luby S.P., Ishmail T., Earhart K., Khodjaev N., Jennings G., Crump J.A. and Mahoney F.J. **Epidemiology and risk factors for endemic typhoid fever in Uzbekistan.** *Tropical Medicine and International Health*, **12**(7):838–847, 2007. 25, 26

351. Luby S.P., Faizan M.K., Fisher-Hoch S.P., Syed A., Mintz E.D., Bhutta Z.A. and McCormick J.B. **Risk factors for typhoid fever in an endemic setting, Karachi, Pakistan.** *Epidemiology and Infection*, **120**(2):129–138, 1998. 26

352. Black R.E., Cisneros L., Levine M.M., Banfi A., Lobos H. and Rodriguez H. **Case-control study to identify risk factors for paediatric endemic typhoid fever in Santiago, Chile**. *Bulletin of the World Health Organisation*, **63**(5):899–904, 1985. 26

353. Luxemburger C., Chau M.C., Mai N.L., Wain J., Tran T.H., Simpson J.A., Le H.K., Nguyen T.T., White N.J. and Farrar J.J. **Risk factors for typhoid fever in the Mekong delta, southern Viet Nam: a case-control study.** *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **95**(1):19–23, 2001. 26

354. Kelly-Hope L.A., Alonso W.J., Thiem V.D., Canh d.o..G., Anh D.D., Lee H. and Miller M.A. **Temporal trends and climatic factors associated with bacterial enteric diseases in Vietnam, 1991-2001.** *Environmental Health Perspectives*, **116**(1):7–12, 2008. 26, 243, 277

355. Jelinek T., Nothdurft H.D., von Sonnenburg F and Löscher T. **Risk Factors for Typhoid Fever in Travelers.** *Journal of Travel Medicine*, **3**(4):200–203, 1996. 26

356. Angell S.Y. and Cetron M.S. **Health disparities among travelers visiting friends and relatives abroad.** *Annals of Internal Medicine*, **142**(1):67–72, 2005. 26

357. Usera M.A., Echeita A., Aladueña A., Alvarez J., Carreño C., Orcau A. and Planas C. **Investigation of an outbreak of water-borne typhoid fever in Catalonia in 1994**. *Enfermedades Infecciosas y Microbiología Clínica*, **13**(8):450–454, 1995. 26, 216

358. Stroffolini T., Manzillo G., De Sena R., Manzillo E., Pagliano P., Zaccarelli M., Russo M., Soscia M. and Giusti G. **Typhoid fever in the Neapolitan area: a case-control study.** *European Journal of Epidemiology*, **8**(4):539–542, 1992. 26

359. Echeita M.A. and Usera M.A. **Chromosomal Rearrangements in *Salmonella enterica* Serotype Typhi Affecting Molecular Typing in Outbreak Investigations**. *Journal of Clinical Microbiology*, **36**(7):2123–2126, 1998. 26, 33, 216

360. Tulchinsky T.H., Burla E., Clayman M., Sadik C., Brown A. and Goldberger S. **Safety of community drinking-water and outbreaks of waterborne enteric disease: Israel, 1976-97.** *Bulletin of the World Health Organization*, **78**(12):1466–1473, 2000. 26

361. Schoenen D. **Role of disinfection in suppressing the spread of pathogens with drinking water: possibilities and limitations**. *Water Research*, **36**(15):3874–3888, 2002. 26

362. Woodward T.E. and Smadel J.E. **Preliminary report on the beneficial effect of chloromycetin in the treatment of typhoid fever.** *Annals of Internal Medicine*, **29**(1):131–134, 1948. 26, 187

363. Colquhoun J. and Weetch R.S. **Resistance to chloramphenicol developing during treatment of typhoid fever**. *The Lancet*, **2**(6639):621–623, 1950. 26, 187, 269

364. Olarte J. and Galindo E. ***Salmonella typhi* resistant to chloramphenicol, ampicillin, and other antimicrobial agents: strains isolated during an extensive typhoid fever epidemic in Mexico.** *Antimicrobial Agents and Chemotherapy*, **4**(6):597–601, 1973. 26, 187

365. Paniker C.K. and Vimala K.N. **Transferable chloramphenicol resistance in *Salmonella typhi*.** *Nature*, **239**(5367):109–110, 1972. 26

366. Butler T., Linh N.N., Arnold K. and Pollack M. **Chloramphenicol-resistant typhoid fever in Vietnam associated with R factor.** *The Lancet*, **302**(7836):983–985, 1973. 26

367. Chun D., Seol S.Y., Cho D.T. and Tak R. **Drug resistance and R plasmids in *Salmonella typhi* isolated in Korea.** *Antimicrobial Agents and Chemotherapy*, **11**(2):209–213, 1977. 26

368. Alton N.K. and Vapnek D. **Nucleotide sequence analysis of the chloramphenicol resistance transposon Tn9**. *Nature*, **282**(5741):864–869, 1979. 26, 200

369. Shaw W.V. **Chloramphenicol acetyltransferase: enzymology and molecular biology.** *Critical Reviews in Biochemistry*, **14**(1):1–46, 1983. 26

370. Maddock C. **Ampicillin in the treatment of typhoid fever.** *The Lancet*, **279**(7235):918+, 1962. 26

371. Patel K. **Ampicillin in the treatment of typhoid**. *The Lancet*, **281**(7295):1378+, 1963. 26

372. Akinkugbe O.O., Lewis E.A., Montefiore D. and Okubadejo O.A. **Trimethoprim and sulphamethoxazole in typhoid.** *British Medical Journal*, **3**(5620):721–722, 1968. 26

373. Rao R.S., Sundararaj T., Subramanian S., Shankar V., Murty S.A. and Kapoor S.C. **A study of drug resistance among *Salmonella typhi* and *Salmonella paratyphi A* in an endemic area, 1977-79.** *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **75**(1):21–24, 1981. 26, 27

374. Ling J. and Chau P.Y. **Plasmids mediating resistance to chloramphenicol, trimethoprim, and ampicillin in *Salmonella typhi* strains isolated in the Southeast Asian region.** *The Journal of Infectious Diseases*, **149**(4), 1984. 27

375. Zhang L. **Mechanism of multiresistant *Salmonella typhi***. *Zhonghua Yi Xue Za Zhi*, **71**(6), 1991. 27

376. Smego R.A., Zaidi A.K., Mohammed Z., Bhutta Z.A. and Hafeez S. **Multiply-resistant *Salmonella* and *Shigella* isolates.** *APMIS. Supplementum*, **3**:65–67, 1988. 27

377. Anand A.C., Kataria V.K., Singh W. and Chatterjee S.K. **Epidemic multiresistant enteric fever in eastern India.** *The Lancet*, **335**(8685), 1990. 27

378. Threlfall E.J., Ward L.R., Rowe B., Raghupathi S., Chandrasekaran V., Vandepitte J. and Lemmens P. **Widespread occurrence of multiple drug-resistant *Salmonella typhi* in India.** *European Journal of Clinical Microbiology and Infectious Diseases*, **11**(11):990–993, 1992. 27

379. Phipps M., Pang T., Koh C.L. and Puthucheary S. **Plasmid incidence rate and conjugative chloramphenicol and tetracycline resistance plasmids in Malaysian isolates of *Salmonella typhi*.** *Microbiology and Immunology*, **35**(2):157–161, 1991. 27

380. Connerton P., Wain J., Hien T.T., Ali T., Parry C., Chinh N.T., Vinh H., Ho V.A., Diep T.S., Day N.P., White N.J., Dougan G. and Farrar J.J. **Epidemic typhoid in vietnam: molecular typing of multiple-antibiotic-resistant *Salmonella enterica* serotype typhi from four outbreaks.** *Journal of Clinical Microbiology*, **38**(2):895–897, 2000. 27, 216, 243

381. Saha S.K. and Saha S.K. **Antibiotic resistance of *Salmonella typhi* in Bangladesh.** *The Journal of Antimicrobial Chemotherapy*, **33**(1):190–191, 1994. 27

382. Mirza S.H., Beeching N.J. and Hart C.A. **Multi-drug resistant typhoid: a global problem.** *Journal of Medical Microbiology*, **44**(5):317–319, 1996. 27

383. Kariuki S., Gilks C., Revathi G. and Hart C.A. **Genotypic analysis of multidrug-resistant *Salmonella enterica* serovar Typhi, Kenya.** *Emerging Infectious Diseases*, **6**(6):649–651, 2000. 27, 263

384. Mirza S.H., Beeching N.J. and Hart C.A. **The prevalence and clinical features of multi-drug resistant *Salmonella typhi* infections in Baluchistan, Pakistan**. *Annals of Tropical Medicine and Parasitology*, **89**(5):515–519, 1995. 27, 188

385. Saxena S.N. and Sen R. ***Salmonella paratyphi A* infection in India: incidence and phage types**. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **60**(3):409–411, 1966. 27, 188

386. Butt T., Ahmad R.N., Salman M. and Kazmi S.Y. **Changing trends in drug resistance among typhoid salmonellae in Rawalpindi, Pakistan.** *Eastern Mediterranean Health Journal*, **11**(5-6):1038–1044, 2005. 27, 188

387. Chandel D.S., Chaudhry R., Dhawan B., Pandey A. and Dey A.B. **Drug-resistant *Salmonella enterica* serotype paratyphi A in India**. *Emerging Infectious Diseases*, **6**(4):420–421, 2000. 27, 188

388. Anjum P., Qureshi A.H., Parvez M.S., Haq M.Z.U. and Hamid M. **Increasing prevalence of multidrug resistant *Salmonella enterica* serotype paratyphi-A in patients with enteric fever**. *Pakistan Journal of Medical Research*, **43**(2), 2004. 27, 188

389. James D.G. **Therapeutic focus. The fluoroquinolones.** *The British Journal of Clinical Practice*, **43**(2):66–67, 1989. 27

390. Pithie A.D. and Wood M.J. **Treatment of typhoid fever and infectious diarrhoea with ciprofloxacin.** *The Journal of Antimicrobial Chemotherapy*, **26 Suppl F**:47–53, 1990. 27

391. Brown J.C., Shanahan P.M.A., Jesudason M.V., Thomson C.J. and Amyes S.G.B. **Mutations responsible for reduced susceptibility to 4-quinolones in clinical isolates of multi-resistant *Salmonella typhi* in India**. *The Journal of Antimicrobial Chemotherapy*, **37**(5):891–900, 1996. 28, 188

392. Brown N.M., Millar M.R., Frost J.A. and Rowe B. **Ciprofloxacin resistance in *Salmonella paratyphi A*.** *The Journal of Antimicrobial Chemotherapy*, **33**(6):1258–1259, 1994. 28

393. Wain J., Hoa N.T.T., Chinh N.T., Vinh H., Everett M.J., Diep T.S., Day N.P.J., Solomon T., White N.J., Piddock L.J.V. and Parry C.M. **Quinolone-resistant *Salmonella typhi* in Viet Nam: Molecular basis of resistance and clinical response to treatment**. *Clinical Infectious Diseases*, **25**(6):1404–1410, 1997. 28, 188

394. Dimitrov T., Udo E.E., Albaksami O., Kilani A.A. and Shehab e.l..D.M. **Ciprofloxacin treatment failure in a case of typhoid fever caused by *Salmonella enterica* serotype Paratyphi A with reduced susceptibility to ciprofloxacin.** *Journal of Medical Microbiology*, **56**(Pt 2):277–279, 2007. 28

395. Jacoby G.a. **Mechanisms of Resistance to Quinolones**. *Clinical Infectious Diseases*, **41**(s2):S120–S126, 2005. 28

396. Hopkins K.L., Davies R.H. and Threlfall E.J. **Mechanisms of quinolone resistance in *Escherichia coli* and *Salmonella*: recent developments.** *International Journal of Antimicrobial Agents*, **25**(5):358–373, 2005. 28, 213, 252

397. Turner A.K., Nair S. and Wain J. **The acquisition of full fluoroquinolone resistance in *Salmonella* Typhi by accumulation of point mutations in the topoisomerase targets**. *The Journal of Antimicrobial Chemotherapy*, **58**(4):733–740, 2006. 28, 47, 74, 213

398. Giraud E., Baucheron S. and Cloeckaert A. **Resistance to fluoroquinolones in *Salmonella*: emerging mechanisms and resistance prevention strategies.** *Microbes and Infection*, **8**(7):1937–1944, 2006. 28, 213

399. Butt T., Khan M.Y., Ahmad R.N., Salman M. and Afzal R.K. **Validity of nalidixic acid screening in fluoroquinolone-resistant typhoid salmonellae.** *Journal of the College of Physicians and Surgeons–Pakistan*, **16**(1):31–34, 2006. 28

400. Chuang C.H., Su L.H., Perera J., Carlos C., Tan B.H., Kumarasinghe G., So T., Van P.H., Chongthaleong A., Hsueh P.R., Liu J.W., Song J.H. and Chiu C.H. **Surveillance of antimicrobial resistance of *Salmonella enterica* serotype Typhi in seven Asian countries**. *Epidemiology and Infection*, **137**(02):266–269, 2009. 28

401. Dutta S., Sur D., Manna B., Sen B., Bhattacharya M., Bhattacharya S.K., Wain J., Nair S., Clemens J.D. and Ochiai R.L. **Emergence of highly fluoroquinolone-resistant *Salmonella enterica* serovar Typhi in a community-based fever surveillance from Kolkata, India.** *International Journal of Antimicrobial Agents*, **31**(4):387–389, 2008. 28, 213, 254

402. Ti T.Y., Monteiro E.H., Lam S. and Lee H.S. **Ceftriaxone therapy in bacteremic typhoid fever.** *Antimicrobial Agents and Chemotherapy*, **28**(4):540–543, 1985. 28

403. Wallace M. **Azithromycin and typhoid**. *The Lancet*, **343**(8911):1497–1498, 1994. 28

404. Effa E.E. and Bukirwa H. **Azithromycin for treating uncomplicated typhoid and paratyphoid fever (enteric fever).** *Cochrane Database of Systematic Reviews*, (4), 2008. 28

405. Saha S.K., Talukder S.Y., Islam M. and Saha S. **A highly ceftriaxone-resistant *Salmonella typhi* in Bangladesh.** *The Pediatric Infectious Disease Journal*, **18**(4), 1999. 28

406. Bhutta Z.A., Farooqui B.J. and Sturm A.W. **Eradication of a multiple drug resistant *Salmonella* paratyphi A causing meningitis with ciprofloxacin.** *The Journal of Infection*, **25**(2):215–219, 1992. 28

407. Wu J.J., Ko W.C., Chiou C.S., Chen H.M., Wang L.R. and Yan J.J. **Emergence of Qnr determinants in human *Salmonella* isolates in Taiwan**. *The Journal of Antimicrobial Chemotherapy*, **62**(6):1269–1272, 2008. 28, 213

408. Hopkins K.L., Wootton L., Day M.R. and Threlfall E.J. **Plasmid-mediated quinolone resistance determinant *qnrS1* found in *Salmonella enterica* strains isolated in the UK.** *The Journal of Antimicrobial Chemotherapy*, **59**(6):1071–1075, 2007. 28, 213

409. García-Fernández A., Fortini D., Veldman K., Mevius D. and Carattoli A. **Characterization of plasmids harbouring *qnrS1*, *qnrB2* and *qnrB19* genes in *Salmonella*.** *The Journal of Antimicrobial Chemotherapy*, **63**(2):274–281, 2009. 28, 213

410. Cheung T.K., Chu Y.W., Chu M.Y., Ma C.H., Yung R.W. and Kam K.M. **Plasmid-mediated resistance to ciprofloxacin and cefotaxime in clinical isolates of *Salmonella enterica* serotype Enteritidis in Hong Kong.** *The Journal of Antimicrobial Chemotherapy*, **56**(3):586–589, 2005. 28, 213

411. Veldman K., van Pelt W. and Mevius D. **First report of *qnr* genes in *Salmonella* in The Netherlands**. *The Journal of Antimicrobial Chemotherapy*, **61**(2):452–453, 2008. 28, 213

412. Cavaco L.M., Hendriksen R.S. and Aarestrup F.M. **Plasmid-mediated quinolone resistance determinant *qnrS1* detected in *Salmonella enterica* serovar Corvallis strains isolated in Denmark and Thailand**. *The Journal of Antimicrobial Chemotherapy*, **60**(3):704–706, 2007. 28, 213

413. Kehrenberg, Corinna, Friederichs, Sonja, Jong D., Anno, Michael, Brenner G., Schwarz and Stefan. **Identification of the plasmid-borne quinolone resistance gene *qnrS* in *Salmonella enterica* serovar Infantis**. *The Journal of Antimicrobial Chemotherapy*, **58**(1):18–22, 2006. 28, 213

414. Gay K., Robicsek A., Strahilevitz J., Park C.H., Jacoby G., Barrett T.J., Medalla F., Chiller T.M. and Hooper D.C. **Plasmid-mediated quinolone resistance in non-Typhi serotypes of *Salmonella enterica*.** *Clinical Infectious Diseases*, **43**(3):297–304, 2006. 28, 213

415. Cattoir V., Weill F.X., Poirel L., Fabre L., Soussy C.J. and Nordmann P. **Prevalence of *qnr* genes in *Salmonella* in France.** *The Journal of Antimicrobial Chemotherapy*, **59**(4):751–754, 2007. 28, 213

416. Weill F.X., Tran H.H., Roumagnac P., Fabre L., Minh N.B., Stavnes T.L., Lassen J., Bjune G., Grimont P.A. and Guerin P.J. **Clonal reconquest of antibiotic-susceptible Salmonella enterica serotype Typhi in Son La Province, Vietnam.** *The American Journal of Tropical Medicine and Hygiene*, **76**(6):1174–1181, 2007. 28

417. Dutta S., Sur D., Manna B., Bhattacharya S.K., Deen J.L. and Clemens J.D. **Rollback of *Salmonella enterica* serotype Typhi resistance to chloramphenicol and other antimicrobials in Kolkata, India.** *Antimicrobial Agents and Chemotherapy*, **49**(4):1662–1663, 2005. 28, 254

418. Mandal S., Mandal M.D. and Pal N.K. **Antimicrobial resistance pattern of Salmonella typhi isolates in Kolkata, India during 1991-2001: a retrospective study.** *Japanese Journal of Infectious Diseases*, **55**(2):58–59, 2002. 28, 254

419. Engels E.A., Falagas M.E., Lau J. and Bennish M.L. **Typhoid fever vaccines: a meta-analysis of studies on efficacy and toxicity**. *British Medical Journal*, **316**(7125):110–116, 1998. 29

420. Fraser A., Paul M., Goldberg E., Acosta C. and Leibovici L. **Typhoid fever vaccines: Systematic review and meta-analysis of randomised controlled trials**. *Vaccine*, **25**(45):7848–7857, 2007. 29

421. Sur D., Ochiai R.L., Bhattacharya S.K., Ganguly N.K., Ali M., Manna B., Dutta S., Donner A., Kanungo S., Park J.K., Puri M.K., Kim D.R., Dutta D., Bhaduri B., Acosta C.J. and Clemens J.D. **A Cluster-Randomized Effectiveness Trial of Vi Typhoid Vaccine in India**. *The New England Journal of Medicine*, **361**(4):335–344, 2009. 29, 216, 218, 254, 262

422. Lin F.Y., Ho V.A., Khiem H.B., Trach D.D., Bay P.V., Thanh T.C., Kossaczka Z., Bryla D.A., Shiloach J., Robbins J.B., Schneerson R. and Szu S.C. **The efficacy of a *Salmonella typhi* Vi conjugate vaccine in two-to-five-year-old children.** *The New England Journal of Medicine*, **344**(17):1263–1269, 2001. 29, 243

423. Steinberg E.B., Bishop R., Haber P., Dempsey A.F., Hoekstra R.M., Nelson J.M., Ackers M., Calugar A. and Mintz E.D. **Typhoid fever in travelers: who should be targeted for prevention?** *Clinical Infectious Diseases*, **39**(2):186–191, 2004. 29

424. Bodhidatta L., Taylor D.N., Thisyakorn U. and Echeverria P. **Control of typhoid fever in Bangkok, Thailand, by annual immunization of schoolchildren with parenteral typhoid vaccine.** *Reviews of Infectious Diseases*, **9**(4):841–845, 1987. 29, 30

425. World Health Organization. **Typhoid vaccines: WHO position paper**. *Weekly Epidemiological Record*, **83**(6):49–59, 2008. 29, 187

426. Poulos C., Bahl R., Whittington D., Bhan M.K., Clemens J.D. and Acosta C.J. **A cost-benefit analysis of typhoid fever immunization programmes in an Indian urban slum community.** *Journal of Health, Population, and Nutrition*, **22**(3):311–321, 2004. 29

427. Deroeck D., Jodar L. and Clemens J. **Putting Typhoid Vaccination on the Global Health Agenda**. *The New England Journal of Medicine*, **357**(11):1069–1071, 2007. 29

428. Cook J., Jeuland M., Whittington D., Poulos C., Clemens J., Sur D., Anh D.D., Agtini M., Bhutta Z. and . **The cost-effectiveness of typhoid Vi vaccination programs: calculations for four urban sites in four Asian countries.** *Vaccine*, **26**(50):6305–6316, 2008. 29

429. Mathew J.L. **Conjugate typhoid vaccine(s) in the Indian context.** *Indian Pediatrics*, **46**(2):182–184, 2009. 29, 187

430. Lauria D.T., Maskery B., Poulos C. and Whittington D. **An optimization model for reducing typhoid cases in developing countries without increasing public spending.** *Vaccine*, **27**(10):1609–1621, 2009. 29

431. Guzman C.A., Borsutzky S., Wenk G.M., Metcalfe I.C., Pearman J., Collioud A., Favre D. and Dietrich G. **Vaccines against typhoid fever**. *Vaccine*, **24**(18):3804–3811, 2006. 30, 187

432. Yang H.H., Wu C.G., Xie G.Z., Gu Q.W., Wang B.R., Wang L.Y., Wang H.F., Ding Z.S., Yang Y., Tan W.S., Wang W.Y., Wang X.C., Qin M., Wang J.H., Tang H.A., Jiang X.M., Li Y.H., Wang M.L., Zhang S.L. and Li G.L. **Efficacy trial of Vi polysaccharide vaccine against typhoid fever in south-western China.** *Bulletin of the World Health Organization*, **79**(7):625–631, 2001. 30

433. Acosta C.J., Hong-Hui Y., Ning W., Qion G., Qun D., Xiaolei M., Baode Z., Liu W., Danovaro-Holliday M.C., Ochiai R.L., Wang X.Y., Kim D.R., Zhi-Yi X., Bai-Qing D., Galindo C.M. and Clemens J.D. **Efficacy of a locally produced, Chinese Vi polysaccharide typhoid fever vaccine during six years of follow-up.** *Vaccine*, **23**(48-49):5618–5623, 2005. 30

434. Achtman M. and Wagner M. **Microbial diversity and the genetic nature of microbial species**. *Nature Reviews Microbiology*, **6**(6):431–440, 2008. 30

435. Fraser C., Alm E.J., Polz M.F., Spratt B.G. and Hanage W.P. **The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity**. *Science*, **323**(5915):741–746, 2009. 30

436. Maiden M.C.J. and Urwn R. **Evolution of microbial pathogens**, chapter 3. ASM Press, 2006. 31

437. Spratt B. **The relative contributions of recombination and point mutation to the diversification of bacterial clones**. *Current Opinion in Microbiology*, **4**(5):602–606, 2001. 31

438. Wirth T., Falush D., Lan R., Colles F., Mensa P., Wieler L.H., Karch H., Reeves P.R., Maiden M.C.J., Ochman H. and Achtman M. **Sex and virulence in *Escherichia coli*: an evolutionary perspective**. *Molecular Microbiology*, **60**(5):1136–1151, 2006. 31

439. Maiden M. **Population genomics: diversity and virulence in the *Neisseria***. *Current Opinion in Microbiology*, **11**(5):467–471, 2008. 31

440. Maynard Smith J. and Smith N.H. **Detecting recombination from gene trees.** *Molecular Biology and Evolution*, **15**(5):590–599, 1998. 31

441. Worobey M. **A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria.** *Molecular Biology and Evolution*, **18**(8):1425–1434, 2001. 31

442. Sinsheimer J.S., Suchard M.A., Dorman K.S., Fang F. and Weiss R.E. **Are you my mother? Bayesian phylogenetic inference of recombination among putative parental strains.** *Applied Bioinformatics*, **2**(3):131–144, 2003. 31

443. Didelot X. and Falush D. **Inference of bacterial microevolution using multilocus sequence data.** *Genetics*, **175**(3):1251–1266, 2007. 31, 39

444. Didelot X., Darling A. and Falush D. **Inferring genomic flux in bacteria**. *Genome Research*, **19**(2):306–317, 2009. 31

445. Sukhnanand S., Alcaine S., Warnick L.D., Su W.L., Hof J., Craver M.P., Mcdonough P., Boor K.J. and Wiedmann M. **DNA Sequence-Based Subtyping and Evolutionary Analysis of Selected *Salmonella enterica* Serotypes**. *Journal of Clinical Microbiology*, **43**(8):3688–3698, 2005. 31

446. Torpdahl M. and Ahrens P. **Population structure of *Salmonella* investigated by amplified fragment length polymorphism.** *Journal of Applied Microbiology*, **97**(3):566–573, 2004. 31, 156

447. Torpdahl M., Skov M., Sandvang D. and Baggesen D. **Genotypic characterization of by multilocus sequence typing, pulsed-field gel electrophoresis and amplified fragment length polymorphism**. *Journal of Microbiological Methods*, **63**(2):173–184, 2005. 31

448. Tankouo-Sandjong B., Sessitsch A., Liebana E., Kornschober C., Allerberger F., Hächler H. and Bodrossy L. **MLST-v, multilocus sequence typing based on virulence genes, for molecular typing of *Salmonella enterica* subsp. *enterica* serovars.** *Journal of Microbiological Methods*, **69**(1):23–36, 2007. 31, 156

449. Brisse S., Fevre C., Passet V., Issenhuth-Jeanjean S., Tournebize R., Diancourt L. and Grimont P. **Virulent clones of *Klebsiella pneumoniae*: identification and evolutionary scenario based on genomic and phenotypic characterization.** *PloS ONE*, **4**(3), 2009. 31

450. Perezlosada M., Porter M., Tazi L. and Crandall K. **New methods for inferring population dynamics from microbial sequences**. *Infection, Genetics and Evolution*, **7**(1):24–43, 2007. 32

451. Leopold S.R., Magrini V., Holt N.J., Shaikh N., Mardis E.R., Cagno J., Ogura Y., Iguchi A., Hayashi T., Mellmann A., Karch H., Besser T.E., Sawyer S.A., Whittam T.S. and Tarr P.I. **A precise reconstruction of the emergence and constrained radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis**. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(21):8713–8718, 2009. 32, 39

452. Laing C., Buchanan C., Taboada E., Zhang Y., Karmali M., Thomas J. and Gannon V. **In silico genomic analyses reveal three distinct lineages of *Escherichia coli* O157:H7, one of which is associated with hyper-virulence.** *BMC Genomics*, **10**(1):287+, 2009. 32

453. Manning S.D., Motiwala A.S., Springman A.C., Qi W., Lacher D.W., Ouellette L.M., Mladonicky J.M., Somsel P., Rudrik J.T., Dietrich S.E., Zhang W., Swaminathan B., Alland D. and Whittam T.S. **Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks**. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(12):4868–4873, 2008. 32, 38

454. Zhang W., Qi W., Albert T.J., Motiwala A.S., Alland D., Hyytia-Trees E.K., Ribot E.M., Fields P.I., Whittam T.S. and Swaminathan B. **Probing genomic diversity and evolution of *Escherichia coli* O157 by single nucleotide polymorphisms.** *Genome Research*, **16**(6):757–767, 2006. 32, 39

455. Boyd E.F. and Hartl D.L. **Diversifying Selection Governs Sequence Polymorphism in the Major Adhesin Proteins FimA, PapA, and SfaA of *Escherichia coli***. *Journal of Molecular Evolution*, **47**(3):258–267, 1998. 32

456. Wong W.S., Yang Z., Goldman N. and Nielsen R. **Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites.** *Genetics*, **168**(2):1041–1051, 2004. 32

457. Chen S.L., Hung C.S., Xu J., Reigstad C.S., Magrini V., Sabo A., Blasiar D., Bieri T., Meyer R.R., Ozersky P., Armstrong J.R., Fulton R.S., Latreille J.P., Spieth J., Hooton T.M., Mardis E.R., Hultgren S.J. and Gordon J.I. **Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: A comparative genomics approach**. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(15):5977–5982, 2006. 32, 38, 39

458. Maiden M.C.J., Bygraves J.A., Feil E., Morelli G., Russell J.E., Urwin R., Zhang Q., Zhou J., Zurth K., Caugant D.A., Feavers I.M., Achtman M. and Spratt B.G. **Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms**. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(6):3140–3145, 1998. 32, 214

459. Feil E.J., Li B.C., Aanensen D.M., Hanage W.P. and Spratt B.G. **eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data.** *Journal of Bacteriology*, **186**(5):1518–1530, 2004. 32

460. Maiden M.C. **Multilocus sequence typing of bacteria.** *Annual Review of Microbiology*, **60**(1):561–588, 2006. 32, 214, 281

461. Jolley K., Chan M.S. and Maiden M. **mlstdbNet - distributed multi-locus sequence typing (MLST) databases**. *BMC Bioinformatics*, **5**(1):86+, 2004. 32

462. **PubMLST**. Online, http://pubmlst.org/. 32

463. **Institut Pasteur MLST Databases**. Online, http://www.pasteur.fr/mlst/. 32

464. **UCC MLST Databases**. Online, http://mlst.ucc.ie/. 32, 97, 151, 161, 179, 214

465. **MLST Databases at Imperial College**. Online, http://www.mlst.net/. 32

466. **EcMLST - A multilocus sequence typing database system for pathogenic *E. coli***. Online, http://www.shigatox.net/cgi-bin/mlst7/index. 32

467. Brehony C., Jolley K.A. and Maiden M.C. **Multilocus sequence typing for global surveillance of meningococcal disease.** *FEMS Microbiology Reviews*, **31**(1):15–26, 2007. 32

468. Fan J., Shu M., Zhang G., Zhou W., Jiang Y., Zhu Y., Chen G., Peacock S.J., Wan C., Pan W. and Feil E.J. **Biogeography and Virulence of *Staphylococcus aureus***. *PLoS ONE*, **4**(7):e6216+, 2009. 32

469. **Mapping MLST**. Online, http://maps.mlst.net/. 32

470. Spratt B.G. and Maiden M.C. **Bacterial population genetics, evolution and epidemiology.** *Philosophical Transactions of the Royal Society of London Series B: Biological sciences*, **354**(1384):701–710, 1999. 32

471. Achtman M. **Evolution, Population Structure, and Phylogeography of Genetically Monomorphic Bacterial Pathogens**. *Annual Review of Microbiology*, **62**(1):53–70, 2008. 33, 38, 40, 282

472. Lazarus A.S. **Typing of Typhoid Bacilli in the Western States by Means of Bacteriophage.** *American Journal of Public Health and the Nation's Health*, **30**(10):1177–1182, 1940. 33

473. Nair S., Poh C.L., Lim Y.S., Tay L. and Goh K.T. **Genome fingerprinting of *Salmonella typhi* by pulsed-field gel electrophoresis for subtyping common phage types.** *Epidemiology and Infection*, **113**(3):391–402, 1994. 33

474. Thong K.L., Cheong Y.M., Puthucheary S., Koh C.L. and Pang T. **Epidemiologic analysis of sporadic *Salmonella typhi* isolates and those from outbreaks by pulsed-field gel electrophoresis.** *Journal of Clinical Microbiology*, **32**(5):1135–1141, 1994. 33, 216

475. Altwegg M., Hickman-Brenner F.W. and Farmer J.J. **Ribosomal RNA gene restriction patterns provide increased sensitivity for typing *Salmonella typhi* strains.** *The Journal of Infectious Diseases*, **160**(1):145–149, 1989. 33

476. Threlfall E.J., Torre E., Ward L.R., Dávalos-Pérez A., Rowe B. and Gibert I. **Insertion sequence IS200 fingerprinting of *Salmonella typhi*: an assessment of epidemiological applicability.** *Epidemiology and Infection*, **112**(2):253–261, 1994. 33

477. Sanborn W.R., Hablas R., Komalarini S., Sinta, Trenggonowati R., Sadjimin T., Atas and Sutrisna. **Salmonellosis in Indonesia: phage type distribution of *Salmonella paratyphi A*.** *The Journal of Hygiene*, **79**(1):1–4, 1977. 33, 96

478. Thong K.L., Nair S., Chaudhry R., Seth P., Kapil A., Kumar D., Kapoor H., Puthucheary S. and Pang T. **Molecular analysis of *Salmonella paratyphi A* from an outbreak in New Delhi, India.** *Emerging Infectious Diseases*, **4**(3):507–508, 1998. 33, 96

479. Matsumoto M., Miwa Y., Hiramatsu R., Yamazaki M., Saito M. and Suzuki Y. ***Salmonella paratyphi A* is more genetically homogeneous than *Salmonella typhi*, as indicated by pulsed-field gel electrophoresis**. *The Journal of the Japanese Association for Infectious Diseases*, **74**(2):143–149, 2000. 33, 96, 161, 285

480. Goh Y.L., Puthucheary S.D., Chaudhry R., Bhutta Z.A., Lesmana M., Oyofo B.A., Punjabi N.H., Ahmed A. and Thong K.L. **Genetic diversity of *Salmonella enterica* serovar Paratyphi A from different geographical regions in Asia**. *Journal of Applied Microbiology*, **92**(6):1167–1171, 2002. 33, 96

481. Ramisse V., Houssu P., Hernandez E., Denoeud F., Hilaire V., Lisanti O., Ramisse F., Cavallo J.D. and Vergnaud G. **Variable Number of Tandem Repeats in *Salmonella enterica* subsp. enterica for Typing Purposes**. *Journal of Clinical Microbiology*, **42**(12):5722–5730, 2004. 33, 34, 96, 105

482. Liu Y., Lee M.A., Ooi E.E., Mavis Y., Tan A.L. and Quek H.H. **Molecular Typing of *Salmonella enterica* Serovar Typhi Isolates from Various Countries in Asia by a Multiplex PCR Assay on Variable-Number Tandem Repeats**. *Journal of Clinical Microbiology*, **41**(9):4388–4394, 2003. 33

483. Octavia S. and Lan R. **Multiple locus variable number of tandem repeat analysis of *Salmonella enterica* serovar Typhi.** *Journal of Clinical Microbiology*, 2009. 33, 34

484. Sanger F., Air G.M., Barrell B.G., Brown N.L., Coulson A.R., Fiddes C.A., Hutchison C.A., Slocombe P.M. and Smith M. **Nucleotide sequence of bacteriophage phi X174 DNA.** *Nature*, **265**(5596):687–695, 1977. 34

485. Sanger F. **Sequences, Sequences, and Sequences.** *Annual Review of Biochemistry*, **57**(1):1–29, 1988. 34

486. Blattner F.R., Plunkett G., Bloch C.A., Perna N.T., Burland V., Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G., F., Gregor J., Davis N.W., Kirkpatrick H.A., Goeden M.A., Rose D.J., Mau B. and Shao Y. **The Complete Genome Sequence of *Escherichia coli* K-12**. *Science*, **277**(5331):1453–1462, 1997. 34

487. Fleischmann R.D., Adams M.D., White O., Clayton R.A., Kirkness E.F., Kerlavage A.R., Bult C.J., Tomb J.F., Dougherty B.A. and Merrick J.M. **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.** *Science*, **269**(5223):496–512, 1995. 34

488. **Genomes OnLine Database Statistics**. Online, http://genomesonline.org/gold_statistics.htm. 34, 35

489. Guzmán E., Romeu A. and Garcia-Vallve S. **Completely sequenced genomes of pathogenic bacteria: a review.** *Enfermedades Infecciosas y Microbiología Clínica*, **26**(2):88–98, 2008. 34

490. Whittam T. **Inferences from whole-genome sequences of bacterial pathogens**. *Current Opinion in Genetics and Development*, **12**(6):719–725, 2002. 34

491. Fraser-Liggett C.M. **Insights on biology and evolution from microbial genome sequencing**. *Genome Research*, **15**(12):1603–1610, 2005. 34, 36

492. Dorrell N., Hinchliffe S.J. and Wren B.W. **Comparative phylogenomics of pathogenic bacteria by microarray analysis.** *Current Opinion in Microbiology*, **8**(5):620–626, 2005. 34, 37

493. Kong Y., Cave M.D., Zhang L., Foxman B., Marrs C.F., Bates J.H. and Yang Z.H. **Population-Based Study of Deletions in Five Different Genomic Regions of *Mycobacterium tuberculosis* and Possible Clinical Relevance of the Deletions**. *Journal of Clinical Microbiology*, **44**(11):3940–3946, 2006. 34, 37, 283

494. Hutchison C.A. **DNA sequencing: bench to bedside and beyond.** *Nucleic Acids Research*, **35**(18):6227–6237, 2007. 34

495. Chan E. **Advances in sequencing technology**. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, **573**(1-2):13–40, 2005. 34, 35

496. **GenBank Growth**. Online, http://ncbi.nlm.nih.gov/Genbank/genbankstats.html. 34

497. Scheibye-Alsing K., Hoffmann S., Frankel A., Jensen P., Stadler P.F., Mang Y., Tommerup N., Gilchrist M.J., Nygård A.B. and Cirera S. **Sequence assembly**. *Computational Biology and Chemistry*, **33**(2):121–136, 2009. 35

498. Richterich P. **Estimation of errors in "raw" DNA sequences: a validation study.** *Genome Research*, **8**(3):251–259, 1998. 35

499. Baxevanis A.D. **An overview of gene identification: approaches, strategies, and considerations.** *Current Protocols in Bioinformatics*, **4**, 2004. 36

500. Delcher A.L., Bratke K.A., Powers E.C. and Salzberg S.L. **Identifying bacterial genes and endosymbiont DNA with Glimmer**. *Bioinformatics*, **23**(6):673–679, 2007. 36

501. Médigue C. and Moszer I. **Annotation, comparison and databases for hundreds of bacterial genomes**. *Research in Microbiology*, **158**(10):724–736, 2007. 36

502. Hardison R.C. **Comparative Genomics**. *PLoS Biology*, **1**(2):e58+, 2003. 36

503. Gill S.R., Fouts D.E., Archer G.L., Mongodin E.F., Deboy R.T., Ravel J., Paulsen I.T., Kolonay J.F., Brinkac L., Beanan M., Dodson R.J., Daugherty S.C., Madupu R., Angiuoli S.V., Durkin A.S., Haft D.H., Vamathevan J., Khouri H., Utterback T., Lee C., Dimitrov G., Jiang L., Qin H., Weidman J., Tran K., Kang K., Hance I.R., Nelson K.E. and Fraser C.M. **Insights on Evolution of Virulence and Resistance from the Complete Genome Analysis of an Early Methicillin-Resistant *Staphylococcus aureus* Strain and a Biofilm-Producing Methicillin-Resistant *Staphylococcus epidermidis* Strain**. *Journal of Bacteriology*, **187**(7):2426–2438, 2005. 36

504. Holden M.T., Heather Z., Paillot R., Steward K.F., Webb K., Ainslie F., Jourdan T., Bason N.C., Holroyd N.E., Mungall K., Quail M.A., Sanders M., Simmonds M., Willey D., Brooks K., Aanensen D.M., Spratt B.G., Jolley K.A., Maiden M.C., Kehoe M., Chanter N., Bentley S.D., Robinson C., Maskell D.J., Parkhill J. and Waller A.S. **Genomic evidence for the evolution of *Streptococcus equi*: host restriction, increased virulence, and genetic exchange with human pathogens.** *PLoS Pathogens*, **5**(3), 2009. 37

505. Hayashi T., Makino K., Ohnishi M., Kurokawa K., Ishii K., Yokoyama K., Han C.G., Ohtsubo E., Nakayama K., Murata T., Tanaka M., Tobe T., Iida T., Takami H., Honda T., Sasakawa C., Ogasawara N., Yasunaga T., Kuhara S., Shiba T., Hattori M. and Shinagawa H. **Complete Genome Sequence of Enterohemorrhagic *Escherichia coli* O157:H7 and Genomic Comparison with a Laboratory Strain K-12**. *DNA Research*, **8**(1):11–22, 2001. 37

506. Perna N.T., Plunkett G., Burland V., Mau B., Glasner J.D., Rose D.J., Mayhew G.F., Evans P.S., Gregor J., Kirkpatrick H.A., Posfai G., Hackett J., Klink S., Boutin A., Shao Y., Miller L., Grotbeck E.J., Davis N.W., Lim A., Dimalanta E.T., Potamousis K.D., Apodaca J., Anantharaman T.S., Lin J., Yen G., Schwartz D.C., Welch R.A. and Blattner F.R. **Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7**. *Nature*, **409**(6819):529–533, 2001. 37

507. Stinear T.P., Seemann T., Pidot S., Frigui W., Reysset G., Garnier T., Meurice G., Simon D., Bouchier C., Ma L., Tichit M., Porter J.L., Ryan J., Johnson P.D.R., Davies J.K., Jenkin G.A., Small P.L.C., Jones L.M., Tekaia F., Laval F., Daffé M., Parkhill J. and Cole S.T. **Reductive evolution and niche adaptation inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer**. *Genome Research*, **17**(2):192–200, 2007. 37

508. Chain P.S.G., Carniel E., Larimer F.W., Lamerdin J., Stoutland P.O., Regala W.M., Georgescu A.M., Vergez L.M., Land M.L., Motin V.L., Brubaker R.R., Fowler J., Hinnebusch J., Marceau M., Medigue C., Simonet M., Chenal-Francisque V., Souza B., Dacheux D., Elliott J.M., Derbise A., Hauser L.J. and Garcia E. **Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis***. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(38):13826–13831, 2004. 37

509. Ambur O.H., Davidsen T., Frye S.A., Balasingham S.V., Lagesen K., Rognes T. and Tønjum T. **Genome dynamics in major bacterial pathogens**. *FEMS Microbiology Reviews*, **33**(3):453–470, 2009. 37

510. Porwollik S., Wong R.M. and Mcclelland M. **Evolutionary genomics of *Salmonella*: Gene acquisitions revealed by microarray analysis**. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(13):8956–8961, 2002. 37, 183

511. Boyd E.F., Porwollik S., Blackmer F. and McClelland M. **Differences in gene content among *Salmonella enterica* serovar Typhi isolates**. *Journal of Clinical Microbiology*, **41**(8):3823–3828, 2003. 37, 38, 47, 80, 81, 281

512. Tsolaki A.G., Hirsh A.E., Deriemer K., Enciso J.A., Wong M.Z., Hannan M., de La Salmoniere Y.O.L., Aman K., Kato-Maeda M. and Small P.M. **Functional and evolutionary genomics of *Mycobacterium tuberculosis*: Insights from genomic deletions in 100 strains**. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(14):4865–4870, 2004. 37, 282

513. Alm R.A., Ling L.S.L., Moir D.T., King B.L., Brown E.D., Doig P.C., Smith D.R., Noonan B., Guild B.C., Dejonge B.L., Carmel G., Tummino P.J., Caruso A., Uria-Nickelsen M., Mills D.M., Ives C., Gibson R., Merberg D., Mills S.D., Jiang Q., Taylor D.E., Vovis G.F. and Trust T.J. **Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori***. *Nature*, **397**(6715):176–180, 1999. 38

514. Bille E., Zahar J.R., Perrin A., Morelle S., Kriz P., Jolley K.A., Maiden M.C.J., Dervin C., Nassif X. and Tinsley C.R. **A chromosomally integrated bacteriophage in invasive meningococci**. *Journal of Experimental Medicine*, **201**(12):1905–1913, 2005. 38

515. Tettelin H., Masignani V., Cieslewicz M.J., Donati C., Medini D., Ward N.L., Angiuoli S.V., Crabtree J., Jones A.L., Durkin A.S., Deboy R.T., Davidsen T.M., Mora M., Scarselli M., y Ros I.M., Peterson J.D., Hauser C.R., Sundaram J.P., Nelson W.C., Madupu R., Brinkac L.M., Dodson R.J., Rosovitz M.J., Sullivan S.A., Daugherty S.C., Haft D.H., Selengut J., Gwinn M.L., Zhou L., Zafar N., Khouri H., Radune D., Dimitrov G., Watkins K., O'Connor K.J., Smith S., Utterback T.R., White O., Rubens C.E., Grandi G., Madoff L.C., Kasper D.L., Telford J.L., Wessels M.R., Rappuoli R. and Fraser C.M. **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome"**. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(39):13950–13955, 2005. 38, 84

516. Touchon M., Hoede C., Tenaillon O., Barbe V., Baeriswyl S., Bidet P., Bingen E., Bonacorsi S., Bouchier C., Bouvet O., Calteau A., Chiapello H., Clermont O., Cruveiller S., Danchin A., Diard M., Dossat C., Karoui M.E., Frapy E., Garry L., Ghigo J.M., Gilles A.M., Johnson J., Le Bouguénec C., Lescat M., Mangenot S., Martinez-Jéhanne V., Matic I., Nassif X., Oztas S., Petit M.A., Pichon C., Rouy Z., Ruf C.S., Schneider D., Tourret J., Vacherie B., Vallenet D., Médigue C., Rocha E.P.C. and Denamur E. **Organised Genome Dynamics in the *Escherichia coli* Species Results in Highly Diverse Adaptive Paths**. *PLoS Genetics*, **5**(1):e1000344+, 2009. 38, 39

517. Bueno S.M., Santiviago C.A., Murillo A.A., Fuentes J.A., Trombert A.N., Rodas P.I., Youderian P. and Mora G.C. **Precise Excision of the Large Pathogenicity Island, SPI7, in *Salmonella enterica* Serovar Typhi**. *Journal of Bacteriology*, **186**(10):3202–3213, 2004. 38, 47, 82

518. Duncan B.K. and Miller J.H. **Mutagenic deamination of cytosine residues in DNA.** *Nature*, **287**(5782):560–561, 1980. 39, 70

519. Hudson R.E., Bergthorsson U. and Ochman H. **Transcription increases multiple spontaneous point mutations in Salmonella enterica.** *Nucleic acids research*, **31**(15):4517–4522, 2003. 39

520. Pearson T., Busch J.D., Ravel J., Read T.D., Rhoton S.D., U'Ren J.M., Simonson T.S., Kachur S.M., Leadem R.R., Cardon M.L., Ert M.N.V., Huynh L.Y., Fraser C.M. and Keim P. **Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing**. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(37):13536–13541, 2004. 39, 40, 41

521. Baker L., Brown T., Maiden M.C. and Drobniewski F. **Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*.** *Emerging Infectious Diseases*, **10**(9):1568–1577, 2004. 39

522. Ogura M., Perez J.C., Mittl P.R.E., Lee H.K., Dailide G., Tan S., Ito Y., Secka O., Dailidiene D., Putty K., Berg D.E. and Kalia A. ***Helicobacter pylori* Evolution: Lineage–Specific Adaptations in Homologs of Eukaryotic Sel1-Like Genes**. *PLoS Computational Biology*, **3**(8):e151+, 2007. 39

523. Yang Z. **Inference of selection from multiple species alignments.** *Current Opinion in Genetics and Development*, **12**(6):688–694, 2002. 39

524. Fry A.J. and Wernegreen J.J. **The roles of positive and negative selection in the molecular evolution of insect endosymbionts.** *Gene*, **355**:1–10, 2005. 39

525. Kryazhimskiy S. and Plotkin J.B. **The Population Genetics of dN/dS**. *PLoS Genetics*, **4**(12):e1000304+, 2008. 39, 92, 94

526. Rocha E.P.C., Smith J.M., Hurst L.D., Holden M.T.G., Cooper J.E., Smith N.H. and Feil E.J. **Comparisons of dN/dS are time dependent for closely related bacterial genomes**. *Journal of Theoretical Biology*, **239**(2):226–235, 2006. 39, 70, 73, 94, 130, 181

527. Liò P. and Bishop M. **Bioinformatics**, chapter 3. Humana Press, 2004. 39

528. Kosiol C., Bofkin L. and Whelan S. **Phylogenetics by likelihood: Evolutionary modeling as a tool for understanding the genome**. *Journal of Biomedical Informatics*, **39**(1):51–61, 2006. 39

529. Delport W., Scheffler K. and Seoighe C. **Models of coding sequence evolution**. *Briefings in Bioinformatics*, **10**(1):97–109, 2009. 39

530. Kwok P.Y. and Chen X. **Detection of single nucleotide polymorphisms.** *Current Issues in Molecular Biology*, **5**(2):43–60, 2003. 40, 41

531. Alland D., Whittam T.S., Murray M.B., Cave M.D., Hazbon M.H., Dix K., Kokoris M., Duesterhoeft A., Eisen J.A., Fraser C.M. and Fleischmann R.D. **Modeling Bacterial Evolution with Comparative-Genome-Based Marker Systems: Application to *Mycobacterium tuberculosis* Evolution and Pathogenesis**. *Journal of Bacteriology*, **185**(11):3392–3399, 2003. 40

532. Gutacker M.M., Smoot J.C., Migliaccio C.A., Ricklefs S.M., Hua S., Cousins D.V., Graviss E.A., Shashkina E., Kreiswirth B.N. and Musser J.M. **Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains.** *Genetics*, **162**(4):1533–1543, 2002. 40, 42

533. Hershberg R., Lipatov M., Small P.M., Sheffer H., Niemann S., Homolka S., Roach J.C., Kremer K., Petrov D.A., Feldman M.W. and Gagneux S. **High Functional Diversity in *Mycobacterium tuberculosis* Driven by Genetic Drift and Human Demography**. *PLoS Biology*, **6**(12):e311+, 2008. 40

534. Jolley K.A., Wilson D.J., Kriz P., Mcvean G. and Maiden M.C.J. **The Influence of Mutation, Recombination, Population History, and Selection on Patterns of Genetic Diversity in *Neisseria meningitidis***. *Molecular Biology and Evolution*, **22**(3):562–569, 2005. 40

535. Sheppard S.K., Dallas J.F., Strachan N.J., MacRae M., McCarthy N.D., Wilson D.J., Gormley F.J., Falush D., Ogden I.D., Maiden M.C. and Forbes K.J. ***Campylobacter* genotyping to determine the source of human infection.** *Clinical Infectious Diseases*, **48**(8):1072–1078, 2009. 40

536. Octavia S. and Lan R. **Single Nucleotide Polymorphism Typing and Genetic Relationships of *Salmonella enterica* serovar Typhi Isolates**. *Journal of Clinical Microbiology*, **45**(11):3795–3801, 2007. 41, 42

537. Black W.C. and Vontas J.G. **Affordable assays for genotyping single nucleotide polymorphisms in insects**. *Insect Molecular Biology*, **16**(4):377–387, 2007. 41

538. Lee J.E. **High-throughput genotyping.** *Forum of Nutrition*, **60**:97–101, 2007. 42

539. Ward T.J., Ducey T.F., Usgaard T., Dunn K.A. and Bielawski J.P. **Multilocus Genotyping Assays for Single Nucleotide Polymorphism-Based Subtyping of *Listeria monocytogenes* Isolates**. *Applied Environmental Microbiology*, **74**(24):7629–7642, 2008. 42

540. Sakamuri R.M.U.R.T.H.Y., Kimura M., Li W.E.I., Kim H.C., Lee H., Madanahally K., Blackiv W.C., Balagon M., Gelber R., Cho S.N., Brennan P.J. and Vissa V. **Population based molecular epidemiology of leprosy in Cebu, Philippines**. *Journal of Clinical Microbiology*, 2009. 42

541. Mcdonald M., Dougall A., Holt D., Huygens F., Oppedisano F., Giffard P.M., Inman-Bamber J., Stephens A.J., Towers R., Carapetis J.R. and Currie B.J. **Use of a Single-Nucleotide Polymorphism Genotyping System To Demonstrate the Unique Epidemiology of Methicillin-Resistant *Staphylococcus aureus* in Remote Aboriginal Communities**. *Journal of Clinical Microbiology*, **44**(10):3720–3727, 2006. 42

542. Foster J.T., Okinaka R.T., Svensson R., Shaw K., De B.K., Robison R.A., Probert W.S., Kenefic L.J., Brown W.D. and Keim P. **Real-Time PCR Assays of Single-Nucleotide Polymorphisms Defining the Major *Brucella* Clades**. *Journal of Clinical Microbiology*, **46**(1):296–301, 2008. 42

543. Papp A.C., Pinsonneault J.K., Cooke G. and Sadée W. **Single nucleotide polymorphism genotyping using allele-specific PCR and fluorescence melting curves.** *BioTechniques*, **34**(5):1068–1072, 2003. 42

544. Vogler A.J., Birdsell D., Price L.B., Bowers J.R., Beckstrom-Sternberg S.M., Auerbach R.K., Beckstrom-Sternberg J.S., Johansson A., Clare A., Buchhagen J.L., Petersen J.M., Pearson T., Vaissaire J., Dempsey M.P., Foxall P., Engelthaler D.M., Wagner D.M. and Keim P. **Phylogeography of *Francisella tularensis*: Global Expansion of a Highly Fit Clone**. *Journal of Bacteriology*, **191**(8):2474–2484, 2009. 42

545. Tang K., Fu D.J., Julien D., Braun A., Cantor C.R. and Köster H. **Chip-based genotyping by mass spectrometry**. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(18):10016–10020, 1999. 42

546. Van Ert M.N., Hofstadler S.A., Jiang Y., Busch J.D., Wagner D.M., Drader J.J., Ecker D.J., Hannis J.C., Huynh L.Y., Schupp J.M., Simonson T.S. and Keim P. **Mass spectrometry provides accurate characterization of two genetic marker types in *Bacillus anthracis*.** *BioTechniques*, **37**(4), 2004. 42

547. Filliol I., Motiwala A.S., Cavatore M., Qi W., Hazbon M.H., Bobadilla Del Valle M., Fyfe J., Garcia-Garcia L., Rastogi N., Sola C., Zozio T., Guerrero M.I., Leon C.I., Crabtree J., Angiuoli S., Eisenach K.D., Durmaz R., Joloba M.L., Rendon A., Sifuentes-Osornio J., de Leon A., Cave M.D., Fleischmann R., Whittam T.S. and Alland D. **Global Phylogeny of *Mycobacterium tuberculosis* Based on Single Nucleotide Polymorphism (SNP) Analysis: Insights into Tuberculosis Evolution, Phylogenetic Accuracy of Other DNA Fingerprinting Systems, and Recommendations for a Minimal Standard SNP Set**. *Journal of Bacteriology*, **188**(2):759–772, 2006. 42

548. Hazbon M.H. and Alland D. **Hairpin Primers for Simplified Single-Nucleotide Polymorphism Analysis of *Mycobacterium tuberculosis* and Other Organisms**. *Journal of Clinical Microbiology*, **42**(3):1236–1242, 2004. 42

549. Wahab T., Hjalmarsson S., Wollin R. and Engstrand L. **Pyrosequencing *Bacillus anthracis*.** *Emerging Infectious Diseases*, **11**(10):1527–1531, 2005. 42

550. Clarke S.C. **Pyrosequencing: nucleotide sequencing technology with bacterial genotyping applications.** *Expert Review of Molecular Diagnostics*, **5**(6):947–953, 2005. 42

551. Hardenbol P., Baner J., Jain M., Nilsson M., Namsaraev E.A., Karlin-Neumann G.A., Fakhrai-Rad H., Ronaghi M., Willis T.D., Landegren U. and Davis R.W. **Multiplexed genotyping with sequence-tagged molecular inversion probes**. *Nature Biotechnology*, **21**(6):673–678, 2003. 42

552. Shendure J. and Ji H. **Next-generation DNA sequencing**. *Nature Biotechnology*, **26**(10):1135–1145, 2008. 42, 43, 44

553. Tettelin H. and Feldblyum T. **Molecular Epidemiology of Microorganisms**, pages 231–247. 2009. 42, 45

554. Pop M. and Salzberg S. **Bioinformatics challenges of new sequencing technology**. *Trends in Genetics*, **24**(3):142–149, 2008. 42, 44

555. Margulies M., Egholm M., Altman W.E., Attiya S., Bader J.S., Bemben L.A., Berka J., Braverman M.S., Chen Y.J., Chen Z., Dewell S.B., Du L., Fierro J.M., Gomes X.V., Godwin B.C., He W., Helgesen S., Ho C.H., Irzyk G.P., Jando S.C., Alenquer M.L., Jarvie T.P., Jirage K.B., Kim J.B., Knight J.R., Lanza J.R., Leamon J.H., Lefkowitz S.M., Lei M., Li J., Lohman K.L., Lu H., Makhijani V.B., McDade K.E., McKenna M.P., Myers E.W., Nickerson E., Nobile J.R., Plant R., Puc B.P., Ronan M.T., Roth G.T., Sarkis G.J., Simons J.F., Simpson J.W., Srinivasan M., Tartaro K.R., Tomasz A., Vogt K.A., Volkmer G.A., Wang S.H., Wang Y., Weiner M.P., Yu P., Begley R.F. and Rothberg J.M. **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature*, **437**(7057):376–380, 2005. 42, 43, 45, 49, 67, 83

556. Leamon J.H., Lee W.L., Tartaro K.R., Lanza J.R., Sarkis G.J., Dewinter A.D., Berka J. and Lohman K.L. **A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions**. *Electrophoresis*, **24**(21):3769–3777, 2003. 43

557. Droege M. and Hill B. **The Genome Sequencer FLX System–Longer reads, more applications, straight forward bioinformatics and more complete data sets**. *Journal of Biotechnology*, **136**(1-2):3–10, 2008. 43

558. Andries K., Verhasselt P., Guillemont J., Gohlmann H.W.H., Neefs J.M., Winkler H., Van Gestel J., Timmerman P., Zhu M., Lee E., Williams P., de Chaffoy D., Huitric E., Hoffner S., Cambau E., Truffot-Pernot C., Lounis N. and Jarlier V. **A Diarylquinoline Drug Active on the ATP Synthase of *Mycobacterium tuberculosis***. *Science*, **307**(5707):223–227, 2005. 43

559. Hofreuter D., Tsai J., Watson R.O., Novik V., Altman B., Benitez M., Clark C., Perbost C., Jarvie T., Du L. and Galan J.E. **Unique Features of a Highly Pathogenic *Campylobacter jejuni* Strain**. *Infection and Immunity*, **74**(8):4694–4707, 2006. 43, 45

560. Smith M.G., Gianoulis T.A., Pukatzki S., Mekalanos J.J., Ornston L.N., Gerstein M. and Snyder M. **New insights into *Acinetobacter baumannii* pathogenesis revealed by high-density pyrosequencing and transposon mutagenesis**. *Genes and Development*, **21**(5):601–614, 2007. 43, 45

561. Quail M.A., Kozarewa I., Smith F., Scally A., Stephens P.J., Durbin R., Swerdlow H. and Turner D.J. **A large genome center's improvements to the Illumina sequencing system**. *Nature Methods*, **5**(12):1005–1010, 2008. 43, 44, 100

562. Bentley D.R., Balasubramanian S., Swerdlow H.P., Smith G.P., Milton J., Brown C.G., Hall K.P., Evers D.J., Barnes C.L., Bignell H.R., Boutell J.M., Bryant J., Carter R.J., Keira C., Cox A.J., Ellis D.J., Flatbush M.R., Gormley N.A., Humphray S.J., Irving L.J., Karbelashvili M.S., Kirk S.M., Li H., Liu X., Maisinger K.S., Murray L.J., Obradovic B., Ost T., Parkinson M.L., Pratt M.R., Rasolonjatovo I.M., Reed M.T., Rigatti R., Rodighiero C., Ross M.T., Sabot A., Sankar S.V., Scally A., Schroth G.P., Smith M.E., Smith V.P., Spiridou A., Torrance P.E., Tzonev S.S., Vermaas E.H., Walter K., Wu X., Zhang L., Alam M.D., Anastasi C., Aniebo I.C., Bailey D.M., Bancarz I.R., Banerjee S., Barbour S.G., Baybayan P.A., Benoit V.A., Benson K.F., Bevis C., Black P.J., Boodhun A., Brennan J.S., Bridgham J.A., Brown R.C., Brown A.A., Buermann D.H., Bundu A.A., Burrows J.C., Carter N.P., Castillo N., Chiara, Chang S., Neil C., Crake N.R., Dada O.O., Diakoumakos K.D., Dominguez-Fernandez B., Earnshaw D.J., Egbujor U.C., Elmore D.W., Etchin S.S., Ewan M.R., Fedurco M., Fraser L.J., Fuentes F., Scott F., George D., Gietzen K.J., Goddard C.P., Golda G.S., Granieri P.A., Green D.E., Gustafson D.L., Hansen N.F., Harnish K., Haudenschild C.D., Heyer N.I., Hims M.M., Ho J.T., Horgan A.M., Hoschler K., Hurwitz S., Ivanov D.V., Johnson M.Q., James T., Huw J., Kang G.D., Kerelska T.H., Kersey A.D., Khrebtukova I., Kindwall A.P., Kingsbury Z., Kokko-Gonzales P.I., Kumar A., Laurent M.A., Lawley C.T., Lee S.E., Lee X., Liao A.K., Loch J.A., Lok M., Luo S., Mammen R.M., Martin J.W., Mccauley P.G., Mcnitt P., Mehta P., Moon K.W., Mullens J.W., Newington T., Ning Z., Ling N., Novo S.M., O'Neill M.J., Osborne M.A., Osnowski A., Ostadan O., Paraschos L.L., Pickering L., Pike A.C., Pike A.C., Chris P., Pliskin D.P., Podhasky J., Quijano V.J., Raczy C., Rae V.H., Rawlings S.R., Chiva R., Roe P.M., Rogers J., Rogert B., Romanov N., Romieu A., Roth R.K., Rourke N.J., Ruediger S.T., Rusman E., Sanches-Kuiper R.M., Schenker M.R., Seoane J.M., Shaw R.J., Shiver M.K., Short S.W., Sizto N.L., Sluis J.P., Smith M.A., Ernest S., Spence E.J., Stevens K., Sutton N., Szajkowski L., Tregidgo C.L., Turcatti G., Vandevondele S., Verhovsky Y., Virk S.M., Wakelin S., Walcott G.C., Wang J., Worsley G.J., Yan J., Yau L., Zuerlein M., Rogers J., Mullikin J.C., Hurles M.E., Mccooke N.J., West J.S., Oaks F.L., Lundberg P.L., Klenerman D., Durbin R. and Smith A. **Accurate whole human genome sequencing using reversible terminator chemistry**. *Nature*, **456**(7218):53–59, 2008. 43, 44, 49

563. Smith A., Xuan Z. and Zhang M. **Using quality scores and longer reads improves accuracy of Solexa read mapping**. *BMC Bioinformatics*, **9**(1):128+, 2008. 44

564. Li H., Ruan J. and Durbin R. **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. *Genome Research*, **18**(11):1851–1858, 2008. 44, 63, 98, 191

565. Li H. and Durbin R. **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics*, **25**(14):1754–1760, 2009. 44

566. Warren R.L., Sutton G.G., Jones S.J.M. and Holt R.A. **Assembling millions of short DNA sequences using SSAKE**. *Bioinformatics*, **23**(4):500–501, 2007. 44

567. Zerbino D. and Birney E. **Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs**. *Genome Research*, **18**(5):821829, 2008. 44, 99

568. Nair S., Alokam S., Kothapalli S., Porwollik S., Proctor E., Choy C., McClelland M., Liu S.L. and Sanderson K.E. ***Salmonella enterica* Serovar Typhi Strains from Which SPI7, a 134-Kilobase Island with Genes for Vi Exopolysaccharide and Other Functions, Has Been Deleted**. *Journal of Bacteriology*, **186**(10):3214–3223, 2004. 47, 82

569. Kothapalli S., Nair S., Alokam S., Pang T., Khakhria R., Woodward D., Johnson W., Stocker B.A.D., Sanderson K.E. and Liu S.L. **Diversity of genome structure in *Salmonella enterica* serovar Typhi populations**. *Journal of Bacteriology*, **187**(8):2638–2650, 2005. 47

570. Le T.A., Fabre L., Roumagnac P., Grimont P.A., Scavizzi M.R. and Weill F.X. **Clonal Expansion and Microevolution of Quinolone-Resistant Salmonella enterica Serotype Typhi in Vietnam from 1996 to 2004**. *Journal of Clinical Microbiology*, **45**(11):3485–3492, 2007. 49, 210, 272

571. Kado C.I. and Liu S.T. **Rapid procedure for detection and isolation of large and small plasmids**. *Journal of Bacteriology*, **145**(3):1365–1373, 1981. 53

572. Taghavi S., van der Lelie D. and Mergeay M. **Electroporation of *Alcaligenes eutrophus* with (mega) plasmids and genomic DNA fragments**. *Applied and Environmental Microbiology*, **60**(10):3585–3591, 1994. 53

573. Felsenstein J. **PHYLIP (Phylogeny Inference Package)**, 2005. 55, 57, 101

574. Larkin M.A., Blackshields G., Brown N.P., Chenna R., Mcgettigan P.A., Mcwilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J. and Higgins D.G. **Clustal W and Clustal X version 2.0**. *Bioinformatics*, **23**(21):2947–2948, 2007. 57, 153, 155

575. Porwollik S., Frye J., Florea L.D., Blackmer F. and McClelland M. **A non-redundant microarray of genes for two related bacteria**. *Nucleic Acids Research*, **31**(7):1869–1876, 2003. 57

576. Kurtz S., Phillippy A., Delcher A.L., Smoot M., Shumway M., Antonescu C. and Salzberg S.L. **Versatile and open software for comparing large genomes**. *Genome Biology*, **5**(2), 2004. 58, 99, 191

577. Rice P., Longden I. and Bleasby A. **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends in Genetics*, **16**(6):276–277, 2000. 59, 99, 190

578. Darling A.C., Mau B., Blattner F.R. and Perna N.T. **Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements**. *Genome Research*, **14**(7):1394–1403, 2004. 59, 151

579. Holt K.E., Parkhill J., Mazzoni C.J., Roumagnac P., Weill F.X., Goodhead I., Rance R., Baker S., Maskell D.J., Wain J., Dolecek C., Achtman M. and Dougan G. **High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi**. *Nature Genetics*, **40**(8):987–993, 2008. 69

580. Hudson R.E., Bergthorsson U. and Ochman H. **Transcription increases multiple spontaneous point mutations in *Salmonella enterica***. *Nucleic Acids Research*, **31**(15):4517–4522, 2003. 70, 180

581. Haraga A., Ohlson M.B. and Miller S.I. **Salmonellae interplay with host cells**. *Nature Reviews Microbiology*, **6**(1):53–66, 2008. 76

582. Falush D. and Bowden R. **Genome-wide association mapping in bacteria?** *Trends in Microbiology*, **14**(8):353–355, 2006. 76

583. Liu T. and Haggard-Ljungquist E. **The transcriptional switch of bacteriophage WPhi, a P2-related but heteroimmune coliphage**. *Journal of Virology*, **73**(12):9816–9826, 1999. 79

584. Bertram G., Innes S., Minella O., Richardson J.P. and Stansfield I. **Endless possibilities: translation termination and stop codon recognition**. *Microbiology*, **147**(2):255–269, 2001. 82

585. Hiller N.L., Janto B., Hogg J.S., Boissy R., Yu S., Powell E., Keefe R., Ehrlich N.E., Shen K., Hayes J., Barbadora K., Klimke W., Dernovoy D., Tatusova T., Parkhill J., Bentley S.D., Post J.C., Ehrlich G.D. and Hu F.Z. **Comparative Genomic Analyses of Seventeen *Streptococcus pneumoniae* Strains: Insights into the Pneumococcal Supragenome**. *Journal of Bacteriology*, **189**(22):8186–8195, 2007. 84

586. Kunkel T.A. and Bebenek K. **DNA Replication Fidelity**. *Annual Review of Biochemistry*, **69**:497–529, 2000. 87

587. Aury J.M., Cruaud C., Barbe V., Rogier O., Mangenot S., Samson G., Poulain J., Anthouard V., Scarpelli C., Artiguenave F. and Wincker P. **High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies**. *BMC Genomics*, **9**:603, 2008. 88

588. la Cruz M.A.D., Fernández-Mora M., Guadarrama C., Flores-Valdez M.A., Bustamante V.H., Vázquez A. and Calva E. **LeuO antagonizes H-NS and StpA-dependent repression in *Salmonella enterica ompS1*.** *Molecular Microbiology*, **66**(3):727–743, 2007. 91

589. Müller C.M., Dobrindt U., Nagy C., Emöd L., Uhlin B. and Hacker J. **Role of histone-like proteins H-NS and StpA in expression of virulence determinants of uropathogenic *Escherichia coli*.** *Journal of Bacteriology*, **188**(15):5428–5438, 2006. 91

590. Romeo T. **Global regulation by the small RNA-binding protein CsrA and the non-coding RNA molecule CsrB.** *Molecular Microbiology*, **29**(6):1321–1330, 1998. 92

591. Lawhon S.D., Frye J.G., Suyemoto M., Porwollik S., McClelland M. and Altier C. **Global regulation by CsrA in *Salmonella typhimurium***. *Molecular Microbiology*, **48**(6):1633–1645, 2003. 92

592. Saurin W., Hofnung M. and Dassa E. **Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters**. *Journal of Molecular Evolution*, **48**(1):22–41, 1999. 92

593. Webber M.A. and Piddock L.J. **The importance of efflux pumps in bacterial antibiotic resistance**. *The Journal of Antimicrobial Chemotherapy*, **51**(1):9–11, 2003. 93

594. Suerbaum S., Smith J.M., Bapumia K., Morelli G., Smith N.H., Kunstmann E., Dyrek I. and Achtman M. **Free recombination within *Helicobacter pylori***. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(21):12619–12624, 1998. 93

595. Gomes J.P., Bruno W.J., Nunes A., Santos N., Florindo C., Borrego M.J. and Dean D. **Evolution of *Chlamydia trachomatis* diversity occurs by widespread interstrain recombination involving hotspots**. *Genome Research*, **17**(1):50–60, 2007. 93

596. Pittius N.C.G.V., Sampson S.L., Lee H., Kim Y., Helden P.D.V. and Warren R.M. **Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (*esx*) gene cluster regions**. *BMC Evolutionary Biology*, **6**, 2006. 93

597. Vaishnavi C., Kochhar R., Singh G., Kumar S., Singh S. and Singh K. **Epidemiology of typhoid carriers among blood donors and patients with biliary, gastrointestinal and other related diseases**. *Microbiology and Immunology*, **49**(2):107–112, 2005. 95

598. Lewis M.D., Serichantalergs O., Pitarangsi C., Chuanak N., Mason C.J., Regmi L.R., Pandey P., Laskar R., Shrestha C.D. and Malla S. **Typhoid fever: A massive, single-point source, multidrug-resistant outbreak in Nepal**. *Clinical Infectious Diseases*, **40**(4):554–561, 2005. 95, 252

599. Sears S.D., Ferreccio C. and Levine M.M. **Sensitivity of Moore sewer swabs for isolating *Salmonella typhi***. *Applied and Environmental Microbiology*, **51**(2):425–426, 1986. 95

600. Cho J.C. and Kim S.J. **Viable, but non-culturable, state of a green fluorescence protein-tagged environmental isolate of *Salmonella typhi* in groundwater and pond water**. *FEMS Microbiology Letters*, **170**(1):257–264, 1999. 95

601. Sokurenko E.V., Gomulkiewicz R. and Dykhuizen D.E. **Source-sink dynamics of virulence evolution**. *Nature Reviews Microbiology*, **4**(7):548–555, 2006. 95

602. Boyd E.F., Wang F.S., Beltran P., Plock S.A., Nelson K. and Selander R.K. ***Salmonella* reference collection B (SARB): strains of 37 serovars of subspecies I.** *Journal of General Microbiology*, **139 Pt 6**:1125–1132, 1993. 96, 156

603. Huson D.H. and Bryant D. **Application of Phylogenetic Networks in Evolutionary Studies**. *Molecular Biology and Evolution*, **23**(2):254–267, 2006. 98, 151, 192

604. Carver T.J., Rutherford K.M., Berriman M., Rajandream M.A., Barrell B.G. and Parkhill J. **ACT: the Artemis Comparison Tool**. *Bioinformatics*, **21**(16):3422–3423, 2005. 99, 151, 155, 191

605. Beissbarth T. and Speed T.P. **GOstat: find statistically overrepresented Gene Ontologies within a group of genes**. *Bioinformatics*, **20**(9):1464–1465, 2004. 100

606. Bösl M. and Kersten H. **A novel RNA product of the *tyrT* operon of *Escherichia coli*.** *Nucleic Acids Research*, **19**(21):5863–5870, 1991. 105

607. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup. **The Sequence Alignment/Map (SAM) Format and SAMtools.** *Bioinformatics*, **25**(16):2078–2079, 2009. 107

608. Notley-McRobb L., King T. and Ferenci T. ***rpoS* mutations and loss of general stress resistance in *Escherichia coli* populations as a consequence of conflict between competing stress responses.** *Journal of Bacteriology*, **184**(3):806–811, 2002. 144

609. Dardonville B. and Raibaud O. **Characterization of *malT* mutants that constitutively activate the maltose regulon of *Escherichia coli*.** *Journal of Bacteriology*, **172**(4):1846–1852, 1990. 144

610. Schnaitman C.A. and Klena J.D. **Genetics of lipopolysaccharide biosynthesis in enteric bacteria**. *Microbiological Reviews*, **57**(3):655–682, 1993. 145

611. Wang L., Liu D. and Reeves P.R. **C-terminal half of Salmonella enterica WbaP (RfbP) is the galactosyl-1-phosphate transferase domain catalyzing the first step of O-antigen synthesis.** *Journal of Bacteriology*, **178**(9):2598–2604, 1996. 145

612. Saldías M.S., Patel K., Marolda C.L., Bittner M., Contreras I. and Valvano M.A. **Distinct functional domains of the Salmonella enterica WbaP transferase that is involved in the initiation reaction for synthesis of the O antigen subunit.** *Microbiology*, **154**(2):440–453, 2008. 145

613. Coburn B., Grassl G.A. and Finlay B.B. ***Salmonella*, the host and disease: a brief review.** *Immunology and Cell Biology*, **85**(2):112–118, 2007. 146

614. Bhan M.K., Bahl R. and Bhatnagar S. **Typhoid and paratyphoid fever**. *The Lancet*, **366**(9487):749–762, 2005. 146

615. Abdel Wahab M.F., Haseeb A.N., Hamdy H.S. and Awadalla Y.A. **Comparative study between paratyphoid A and typhoid fever cases.** *The Journal of the Egyptian Public Health Association*, **71**:539–551, 1996. 146

616. Vollaard A.M., Ali S., Widjaja S., Asten H.A., Visser L.G., Surjadi C. and van Dissel J.T. **Identification of typhoid fever and paratyphoid fever cases at presentation in outpatient clinics in Jakarta, Indonesia.** *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **99**(6):440–450, 2005. 146

617. Sinnott C.R. and Teall A.J. **Persistent gallbladder carriage of *Salmonella typhi*.** *The Lancet*, **1**(8539), 1987. 146

618. Brodie J., Macqueen I.A. and Livingstone D. **Effect of trimethoprim-sulphamethoxazole on typhoid and salmonella carriers.** *British Medical Journal*, **3**(5718):318–319, 1970. 146

619. Jacobs M.R., Koornhof H.J., Crisp S.I., Palmhert H.L. and Fitzstephens A. **Enteric fever caused by *Salmonella paratyphi C* in South and South West Africa.** *South African Medical Journal*, **54**(11):434–438, 1978. 147, 179

620. Selander R.K., Beltran P., Smith N.H., Barker R.M., Crichton P.B., Old D.C., Musser J.M. and Whittam T.S. **Genetic population structure, clonal phylogeny, and pathogenicity of *Salmonella paratyphi B*.** *Infection and Immunity*, **58**(6):1891–1901, 1990. 147

621. Cherry W.B., Davis B.R. and Edwards P.R. **Observations on the types and typing of *Salmonella paratyphi B* cultures in the United States.** *American Journal of Public Health and the Nation's Health*, **43**(10):1280–1286, 1953. 147

622. Le Minor L., Beaud R., Laurent B. and Monteil V. ***Salmonella* possessing the 6,7:c:1,5 antigenic factors**. *Annales de l'Institut Pasteur. Microbiologie*, **136B**(2):225–234, 1985. 147, 178

623. Barker R.M. **Utilization of d-tartaric acid by *Salmonella paratyphi B* and *Salmonella java*: comparison of anaerobic plate test, lead acetate test and turbidity test.** *The Journal of Hygiene*, **95**(1):107–114, 1985. 147

624. Giglioli G. **Agglutinins for the Typhoid-Paratyphoid Group in a Random Sample of the Population of British Guiana**. *The Journal of Hygiene*, **33**(3):379–386, 1933. 147, 179

625. Old D.C., Yakubu D.E. and Senior B.W. **Characterisation of a Fimbrial, Mannose-Resistant and Eluting Haemagglutinin (MREHA) Produced by Strains of *Salmonella* of Serotype Sendai**. *Journal of Medical Microbiology*, **30**(1):59–68, 1989. 147

626. Chau P.Y. and Huang C.T. **Biochemical characterization of H2S-positive *Salmonella* sendai strains isolated in Hong Kong.** *Microbiology and Immunology*, **23**:125–129, 1979. 147

627. **FoodNet-Foodborne Diseases Active Surveillance Network**. Online, http://www.cdc.gov/foodnet/. 147

628. Selander R.K., Beltran P., Smith N.H., Helmuth R., Rubin F.A., Kopecko D.J., Ferris K., Tall B.D., Cravioto A. and Musser J.M. **Evolutionary genetic relationships of clones of *Salmonella* serovars that cause human typhoid and other enteric fevers.** *Infection and Immunity*, **58**(7):2262–2275, 1990. 147

629. Kingsley R.A. and Bäumler A.J. **Host adaptation and the emergence of infectious disease: the *Salmonella* paradigm.** *Molecular Microbiology*, **36**(5):1006–1014, 2000. 147, 148

630. Sivapalasingam S., Friedman C.R., Cohen L. and Tauxe R.V. **Fresh produce: a growing cause of outbreaks of foodborne illness in the United States, 1973 through 1997.** *Journal of Food Protection*, **67**(10):2342–2353, 2004. 148

631. Thorns C.J. **Bacterial food-borne zoonoses.** *Revue Scientifique et Technique*, **19**(1):226–239, 2000. 148

632. Dupont H.L. **Food Safety: The Growing Threat of Foodborne Bacterial Enteropathogens of Animal Origin**. *Clinical Infectious Diseases*, **45**(10):1353–1361, 2007. 148

633. Steere A.C., Hall W.J., Wells J.G., Craven P.J., Leotsakis N., Farmer J.J. and Gangarosa E.J. **Person-to-person spread of *Salmonella typhimurium* after a hospital common-source outbreak.** *The Lancet*, **1**(7902):319–322, 1975. 148

634. Loewenstein M.S. **An outbreak of salmonellosis propagated by person-to-person transmission on an Indian reservation.** *American Journal of Epidemiology*, **102**(3):257–262, 1975. 148

635. Vladoianu I.R., Chang H.R. and Pechère J.C. **Expression of host resistance to *Salmonella typhi* and *Salmonella typhimurium*: bacterial survival within macrophages of murine and human origin.** *Microbial Pathogenesis*, **8**(2):83–90, 1990. 148

636. Ishibashi Y. and Arai T. **A possible mechanism for host-specific pathogenesis of *Salmonella* serovars.** *Microbial Pathogenesis*, **21**(6):435–446, 1996. 148

637. Schwan W.R., Huang X.Z., Hu L. and Kopecko D.J. **Differential bacterial survival, replication, and apoptosis-inducing ability of *Salmonella* serovars within human and murine macrophages**. *Infection and Immunity*, **68**(3):1005–1013, 2000. 148

638. Weinstein D.L., O'Neill B.L., Hone D.M. and Metcalf E.S. **Differential early interactions between *Salmonella enterica* serovar Typhi and two other pathogenic *Salmonella* serovars with intestinal epithelial cells.** *Infection and Immunity*, **66**(5):2310–2318, 1998. 148

639. Bäumler A.J., Tsolis R.M. and Heffron F. **Contribution of fimbrial operons to attachment to and invasion of epithelial cell lines by *Salmonella typhimurium*.** *Infection and Immunity*, **64**(5):1862–1865, 1996. 148

640. Forest C., Faucher S.P., Poirier K., Houle S., Dozois C.M. and Daigle F. **Contribution of the *stg* fimbrial operon of *Salmonella enterica* serovar Typhi during interaction with human cells.** *Infection and Immunity*, **75**(11):5264–5271, 2007. 148, 162, 177

641. Kaiser P., Rothwell L., Galyov E.E., Barrow P.A., Burnside J. and Wigley P. **Differential cytokine expression in avian cells in response to invasion by *Salmonella typhimurium, Salmonella enteritidis* and *Salmonella gallinarum*.** *Microbiology*, **146 Pt 12**:3217–3226, 2000. 148

642. Nielsen L.R., Schukken Y.H., Gröhn Y.T. and Ersbøll A.K. ***Salmonella* Dublin infection in dairy cattle: risk factors for becoming a carrier.** *Preventive Veterinary Medicine*, **65**(1-2):47–62, 2004. 148

643. Carver T., Thomson N., Bleasby A., Berriman M. and Parkhill J. **DNAPlotter: Circular and linear interactive genome visualisation.** *Bioinformatics*, **25**(1):119–120, 2009. 151

644. Stamatakis and Alexandros. **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models**. *Bioinformatics*, **22**(21):2688–2690, 2006. 151, 192, 226

645. Posada D. and Crandall K. **MODELTEST: testing the model of DNA substitution**. *Bioinformatics*, **14**(9):817–818, 1998. 153, 226

646. Tao N., Richardson R., Bruno W. and Kuiken C. **FindModel**. Online, http://www.hiv.lanl.gov/content/sequence/findmodel/findmodel.html. 153, 226

647. Drummond A.J. and Rambaut A. **BEAST: Bayesian evolutionary analysis by sampling trees**. *BMC Evolutionary Biology*, **7**:214+, 2007. 153, 154, 158, 161

648. Drummond A.J., Ho S.Y., Phillips M.J. and Rambaut A. **Relaxed Phylogenetics and Dating with Confidence**. *PLoS Biology*, **4**(5), 2006. 153

649. Suchard M.A., Weiss R.E. and Sinsheimer J.S. **Bayesian selection of continuous-time Markov chain evolutionary models.** *Molecular Biology and Evolution*, **18**(6):1001–1013, 2001. 153

650. Yang Z. and Nielsen R. **Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models.** *Molecular Biology and Evolution*, **17**(1):32–43, 2000. 155, 180

651. Yang Z. **Phylogenetic Analysis by Maximum Likelihood (PAML)**. Online, http://abacus.gene.ucl.ac.uk/software/paml.html. 155

652. Christensen H., Nordentoft S. and Olsen J.E. **Phylogenetic relationships of *Salmonella* based on rRNA sequences.** *International Journal of Systematic Bacteriology*, **48 Pt 2**:605–610, 1998. 156

653. Scott F., Threlfall J. and Arnold C. **Genetic structure of *Salmonella* revealed by fragment analysis.** *International Journal of Systematic and Evolutionary Microbiology*, **52**(5):1701–1713, 2002. 156

654. Campbell J.W., Morgan-Kiss R.M. and Cronan J.E. **A new *Escherichia coli* metabolic competency: growth on fatty acids by a novel anaerobic beta-oxidation pathway**. *Molecular Microbiology*, **47**(3):793–805, 2003. 166

655. Freeman J.A., Ohl M.E. and Miller S.I. **The *Salmonella enterica* serovar typhimurium translocated effectors SseJ and SifB are targeted to the *Salmonella*-containing vacuole.** *Infection and Immunity*, **71**(1):418–427, 2003. 166

656. Schlumberger M.C. and Hardt W.D. ***Salmonella* type III secretion effectors: pulling the host cell's strings**. *Current Opinion in Microbiology*, **9**(1):46–54, 2006. 176

657. Krogh A., Larsson B., von Heijne G. and Sonnhammer E.L. **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *Journal of Molecular Biology*, **305**(3):567–580, 2001. 177, 190

658. Uzzau S., Hovi M. and Stocker B.A. **Application of ribotyping and IS200 fingerprinting to distinguish the five *Salmonella* serotype O6,7:c:1,5 groups: Choleraesuis sensu stricto, Choleraesuis var. Kunzendorf, Choleraesuis var. Decatur, Paratyphi C, and Typhisuis.** *Epidemiology and Infection*, **123**(1):37–46, 1999. 179

659. Leminor L., Charie Marsaines C., Zajc Satler J., Delage R., Bories S., Perpezat A. and Segonne J. **New *Salmonella* serotypes identified in 1963**. *Annales de l'Institut Pasteur*, **106**:931–934, 1964. 179

660. Kariuki S., Cheesbrough J., Mavridis A.K. and Hart C.A. **Typing of *Salmonella enterica* serotype paratyphi C isolates from various countries by plasmid profiles and pulsed-field gel electrophoresis.** *Journal of Clinical Microbiology*, **37**(6):2058–2060, 1999. 179

661. Stoletzki N. and Eyre-Walker A. **Synonymous codon usage in *Escherichia coli*: Selection for translational accuracy**. *Molecular Biology and Evolution*, **24**(2):374–381, 2007. 180

662. Moran N.A., Munson M.A., Baumann P. and Ishikawa H. **A Molecular Clock in Endosymbiotic Bacteria is Calibrated Using the Insect Hosts**. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **253**(1337):167–171, 1993. 180

663. Moodley Y., Linz B., Yamaoka Y., Windsor H.M., Breurec S., Wu J.Y., Maady A., Bernhoft S., Thiberge J.M., Phuanukoonnon S., Jobb G., Siba P., Graham D.Y., Marshall B.J. and Achtman M. **The Peopling of the Pacific from a Bacterial Perspective**. *Science*, **323**(5913):527–530, 2009. 180

664. Zlateva K.T., Lemey P., Vandamme A.M. and Van Ranst M. **Molecular evolution and circulation patterns of human respiratory syncytial virus subgroup a: positively selected sites in the attachment g glycoprotein.** *Journal of Virology*, **78**(9):4675–4683, 2004. 180

665. Rambaut A., Pybus O.G., Nelson M.I., Viboud C., Taubenberger J.K. and Holmes E.C. **The genomic and epidemiological dynamics of human influenza A virus**. *Nature*, 2008. 180

666. Hasegawa M. and Kishino H. **Heterogeneity of tempo and mode of mitochondrial DNA evolution among mammalian orders.** *Japanese Journal of Genetics*, **64**(4):243–258, 1989. 180

667. Ayala F.J. **Vagaries of the molecular clock.** *Proceedings of the National Academy of Sciences of the United States of America*, **94**(15):7776–7783, 1997. 180

668. Britten R.J. **Rates of DNA sequence evolution differ between taxonomic groups**. *Science*, **231**(4744):1393–1398, 1986. 180

669. Singh N.D., Arndt P.F., Clark A.G. and Aquadro C.F. **Strong Evidence for Lineage- and Sequence-Specificity of Substitution Rates and Patterns in Drosophila**. *Molecular Biology and Evolution*, **26**(7):1591–1605, 2009. 180

670. Ho S.Y., Shapiro B., Phillips M.J., Cooper A. and Drummond A.J. **Evidence for time dependency of molecular rate estimates.** *Systematic Biology*, **56**(3):515–522, 2007. 181

671. Holmes E.C. **Patterns of Intra- and Interhost Nonsynonymous Variation Reveal Strong Purifying Selection in Dengue Virus**. *Journal of Virology*, **77**(20):11296–11298, 2003. 181

672. Sharp P.M., Bailes E., Chaudhuri R.R., Rodenburg C.M., Santiago M.O. and Hahn B.H. **The origins of acquired immune deficiency syndrome viruses: where and when?** *Philosophical transactions of the Royal Society of London Series B: Biological sciences*, **356**(1410):867–876, 2001. 181

673. Wood, Bernard, Lonergan and Nicholas. **The hominin fossil record: taxa, grades and clades**. *Journal of Anatomy*, **212**(4):354–376, 2008. 181

674. **Time Tree–The Timescale of Life**. Online, http://www.timetree.org. 181

675. Jiang X., Rossanese O.W., Brown N.F., Kujat-Choy S., Galán J.E., Finlay B.B. and Brumell J.H. **The related effector proteins SopD and SopD2 from *Salmonella enterica* serovar Typhimurium contribute to virulence during systemic infection of mice**. *Molecular Microbiology*, **54**(5):1186–1198, 2004. 183

676. Brumell J.H., Kujat-Choy S., Brown N.F., Vallance B.A., Knodler L.A. and Finlay B.B. **SopD2 is a novel type III secreted effector of *Salmonella typhimurium* that targets late endocytic compartments upon delivery into host cells.** *Traffic*, **4**(1):36–48, 2003. 183

677. Halici S., Zenk S.F., Jantsch J. and Hensel M. **Functional analysis of the *Salmonella* pathogenicity island 2-mediated inhibition of antigen presentation in dendritic cells.** *Infection and Immunity*, **76**(11):4924–4933, 2008. 183

678. Zhang Y., Higashide W., Dai S., Sherman D.M. and Zhou D. **Recognition and Ubiquitination of *Salmonella* Type III Effector SopA by a Ubiquitin E3 Ligase, HsRMA1**. *Journal of Biological Chemistry*, **280**(46):38682–38688, 2005. 184

679. Zhang, Ying, Higashide, Wendy M., Mccormick, Beth A., Chen, Jue, Zhou and Daoguo. **The inflammation-associated *Salmonella* SopA is a HECT-like E3 ubiquitin ligase**. *Molecular Microbiology*, **62**(3):786–793, 2006. 184

680. Diao J., Zhang Y., Huibregtse J.M., Zhou D. and Chen J. **Crystal structure of SopA, a *Salmonella* effector protein mimicking a eukaryotic ubiquitin ligase**. *Nature Structural and Molecular Biology*, **15**(1):65–70, 2007. 184

681. Raffatellu M., Wilson R.P., Chessa D., Andrews-Polymenis H., Tran Q.T., Lawhon S., Khare S., Adams L.G. and Baumler A.J. **SipA, SopA, SopB, SopD, and SopE2 Contribute to *Salmonella enterica* Serotype Typhimurium Invasion of Epithelial Cells**. *Infection and Immunity*, **73**(1):146–154, 2005. 184

682. Zhang S., Santos R.L., Tsolis R.M., Stender S., Hardt W.D., Baumler A.J. and Adams L.G. **The *Salmonella enterica* Serotype Typhimurium Effector Proteins SipA, SopA, SopB, SopD, and SopE2 Act in Concert To Induce Diarrhea in Calves**. *Infection and Immunity*, **70**(7):3843–3855, 2002. 184

683. Kingsley R.A., Humphries A.D., Weening E.H., De Zoete M.R., Winter S., Papaconstantinopoulou A., Dougan G. and Bäumler A.J. **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):629–640, 2003. 185

684. Mered B., Poittevin F. and Louli B. **Surveillance of drug resistance in pathogenic enterobacteria in Algeria. I. Study of the resistance of major and minor *Salmonella* species to antibiotics in 1973-1974**. *Archives. Institut Pasteur d'Algerie*, **52**:17–35, 1977. 187, 212

685. Asperilla M.O., Smego R.A. and Scott L.K. **Quinolone antibiotics in the treatment of *Salmonella* infections.** *Reviews of Infectious Diseases*, **12**(5):873–889, 1990. 187, 213

686. Joshi S. and Amarnath S.K. **Fluoroquinolone resistance in *Salmonella typhi* and *S. paratyphi A* in Bangalore, India**. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **101**(3):308–310, 2007. 188

687. Pokharel B.M., Koirala J., Dahal R.K., Mishra S.K., Khadga P.K. and Tuladhar N.R. **Multidrug-resistant and extended-spectrum beta-lactamase (ESBL)-producing *Salmonella enterica* (serotypes Typhi and Paratyphi A) from blood isolates in Nepal: surveillance of resistance and a search for newer alternatives**. *International Journal of Infectious Diseases*, **10**(6):434–438, 2006. 188, 252

688. Goh Y.L., Puthucheary S.D., Chaudhry R., Bhutta Z.A., Lesmana M., Oyofo B.A., Punjabi N.H., Ahmed A. and Thong K.L. **Genetic diversity of *Salmonella enterica* serovar Paratyphi A from different geographical regions in Asia.** *Journal of Applied Microbiology*, **92**(6):1167–1171, 2002. 188

689. Maher D. and Taylor D.E. **Host range and transfer efficiency of incompatibility group HI plasmids.** *Canadian Journal of Microbiology*, **39**(6):581–587, 1993. 188, 189, 212

690. Taylor D.E., Brose E.C., Kwan S. and Yan W. **Mapping of transfer regions within incompatibility group HI plasmid R27.** *Journal of Bacteriology*, **162**(3):1221–1226, 1985. 188, 195, 197

691. Rooker M.M., Sherburne C., Lawley T.D. and Taylor D.E. **Characterization of the Tra2 region of the IncHI1 plasmid R27**. *Plasmid*, **41**(3):226–239, 1999. 188, 195, 197

692. Sherburne C.K., Lawley T.D., Gilmour M.W., Blattner F.R., Burland V., Grotbeck E., Rose D.J. and Taylor D.E. **The complete DNA sequence and analysis of R27, a large IncHI plasmid from *Salmonella typhi* that is temperature sensitive for transfer**. *Nucleic Acids Research*, **28**(10):2177–2186, 2000. 188, 189, 195, 197, 212

693. Finn R.D., Tate J., Mistry J., Coggill P.C., Sammut S.J., Hotz H.R., Ceric G., Forslund K., Eddy S.R., Sonnhammer E.L. and Bateman A. **The Pfam protein families database.** *Nucleic Acids Research*, **36**(Database issue), 2008. 190

694. Nielsen H. and Krogh A. **Prediction of signal peptides and signal anchors by a hidden Markov model.** *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, **6**:122–130, 1998. 190

695. Bendtsen J.D., Nielsen H., von Heijne G. and Brunak S. **Improved prediction of signal peptides: SignalP 3.0.** *Journal of Molecular Biology*, **340**(4):783–795, 2004. 190

696. Dodd I.B. and Egan. **Improved detection of helix-turn-helix DNA-binding motifs in protein sequences**. *Nucleic Acids Research*, **18**(17):5019–5026, 1990. 190

697. Rutherford K., Parkhill J., Crook J., Horsnell T., Rice P., Rajandream M.A. and Barrell B. **Artemis: sequence visualization and annotation**. *Bioinformatics*, **16**(10):944–945, 2000. 190

698. Shine J. and Dalgarno L. **Determinant of cistron specificity in bacterial ribosomes**. *Nature*, **254**(5495):34–38, 1975. 190

699. **IS Finder**. Online, http://www-is.biotoul.fr/. 190

700. Rozen S. and Skaletsky H. **Primer3 on the WWW for general users and for biologist programmers.** *Methods in Molecular Biology*, **132**:365–386, 2000. 192

701. Grindley N.D.F., Humphreys G.O. and Anderson E.S. **Molecular studies of R factor compatibility groups**. *Journal of Bacteriology*, **115**(1):387–398, 1973. 195

702. Gabant P., Newnham P., Taylor D. and Couturier M. **Isolation and location on the R27 map of two replicons and an incompatibility determinant specific for IncHI1 plasmids**. *Journal of Bacteriology*, **175**(23):7697–7701, 1993. 195

703. Gilmour M.W., Thomson N.R., Sanders M., Parkhill J. and Taylor D.E. **The complete nucleotide sequence of the resistance plasmid R478: defining the backbone components of incompatibility group H conjugative plasmids through comparative genomics**. *Plasmid*, **52**(3):182–202, 2004. 195

704. Foster T.J., Davis M.A., Roberts D.E., Takeshita K. and Kleckner N. **Genetic organization of transposon Tn10.** *Cell*, **23**(1):201–213, 1981. 197

705. Jorgensen R.A., Berg D.E., Allet B. and Reznikoff W.S. **Restriction enzyme cleavage map of Tn10, a transposon which encodes tetracycline resistance.** *Journal of Bacteriology*, **137**(1):681–685, 1979. 197

706. Lawley T.D., Burland V. and Taylor D.E. **Analysis of the complete nucleotide sequence of the tetracycline-resistance transposon Tn10**. *Plasmid*, **43**(3):235–239, 2000. 197

707. Kleckner N. **DNA sequence analysis of Tn10 insertions: origin and role of 9 bp flanking repetitions during Tn10 translocation.** *Cell*, **16**(4):711–720, 1979. 197

708. Johnsrud L., Calos M.P. and Miller J.H. **The transposon Tn9 generates a 9 bp repeated sequence during integration.** *Cell*, **15**(4):1209–1219, 1978. 200

709. Liebert C.A., Hall R.M. and Summers A.O. **Transposon Tn21, flagship of the floating genome**. *Microbiology and Molecular Biology Reviews*, **63**(3):507–522, 1999. 200, 201

710. Sundström L., Swedberg G. and Sköld O. **Characterization of transposon Tn5086, carrying the site-specifically inserted gene *dhfrVII* mediating trimethoprim resistance.** *Journal of Bacteriology*, **175**(6):1796–1805, 1993. 200

711. Sköld O. **Resistance to trimethoprim and sulfonamides.** *Veterinary Research*, **32**(3-4):261–273, 2001. 200

712. Sundström L., Rådström P., Swedberg G. and Sköld O. **Site-specific recombination promotes linkage between trimethoprim- and sulfonamide resistance genes. Sequence characterization of *dhfrV* and *sulI* and a recombination active locus of Tn21.** *Molecular and General Genetics*, **213**(2-3):191–201, 1988. 200

713. Rådström P. and Swedberg G. **RSF1010 and a conjugative plasmid contain *sulII*, one of two known genes for plasmid-borne sulfonamide resistance dihydropteroate synthase.** *Antimicrobial Agents and Chemotherapy*, **32**(11):1684–1692, 1988. 200

714. Sundin G.W. and Bender C.L. **Dissemination of the *strA-strB* streptomycin-resistance genes among commensal and pathogenic bacteria from humans, animals, and plants.** *Molecular Ecology*, **5**(1):133–143, 1996. 200

715. Szczepanowski R., Braun S., Riedel V., Schneiker S., Krahn I., Puhler A. and Schluter A. **The 120 592 bp IncF plasmid pRSB107 isolated from a sewage-treatment plant encodes nine different antibiotic-resistance determinants, two iron-acquisition systems and other putative virulence-associated functions**. *Microbiology*, **151**(4):1095–1111, 2005. 200

716. Chen C.Y., Nace G.W., Solow B. and Fratamico P. **Complete nucleotide sequences of 84.5- and 3.2-kb plasmids in the multi-antibiotic resistant *Salmonella enterica* serovar Typhimurium U302 strain G8430**. *Plasmid*, 2006. 200

717. Daly M., Villa L., Pezzella C., Fanning S. and Carattoli A. **Comparison of multidrug resistance gene regions between two geographically unrelated *Salmonella* serotypes**. *The Journal of Antimicrobial Chemotherapy*, **55**(4):558–561, 2005. 200, 213

718. Iida S., Mollet B., Meyer J. and Arber W. **Functional characterization of the prokaryotic mobile genetic element IS26.** *Molecular and General Genetics*, **198**:84–89, 1984. 200

719. Partridge S.R. and Hall R.M. **The IS1111 family members IS4321 and IS5075 have subterminal inverted repeats and target the terminal inverted repeats of Tn21 family transposons**. *Journal of Bacteriology*, **185**(21):6371–6384, 2003. 202

720. Brzezinska M. and Davies J. **Two enzymes which phosphorylate neomycin and kanamycin in *Escherichia coli* strains carrying R factors.** *Antimicrobial Agents and Chemotherapy*, **3**(2):266–269, 1973. 202

721. Culham D.E., Lu A., Jishage M., Krogfelt K.A., Ishihama A. and Wood J.M. **The osmotic stress response and virulence in pyelonephritis isolates of *Escherichia coli*: contributions of RpoS, ProP, ProU and other systems.** *Microbiology*, **147**(6):1657–1670, 2001. 202, 280

722. Ly A., Henderson J., Lu A., Culham D.E. and Wood J.M. **Osmoregulatory systems of *Escherichia coli*: identification of betaine-carnitine-choline transporter family member BetU and distributions of *betU* and *trkG* among pathogenic and nonpathogenic isolates.** *Journal of Bacteriology*, **186**(2):296–306, 2004. 202

723. Smith H.W., Parsell Z. and Green P. **Thermosensitive H1 plasmids determining citrate utilization.** *Journal of General Microbiology*, **109**(2):305–311, 1978. 203

724. Taylor D.E. and Brose E.C. **Location of plasmid-mediated citrate utilization determinant in R27 and incidence in other H incompatibility group plasmids.** *Applied and Environmental Microbiology*, **52**(6):1394–1397, 1986. 203

725. Phan M.D., Kidgell C., Nair S., Holt K.E., Turner A.K., Hinds J., Butcher P., Cooke F.J., Thomson N.R., Titball R., Bhutta Z.A., Hasan R., Dougan G. and Wain J. **Variation in *Salmonella enterica* serovar typhi IncHI1 plasmids during the global spread of resistant typhoid fever.** *Antimicrobial Agents and Chemotherapy*, **53**(2):716–727, 2009. 204, 207, 209, 210, 212, 232, 240, 241, 272

726. Lowbury E.J. and Hurst L. **Atypical anaerobic forms of *Streptococcus pyogenes* associated with tetracycline resistance.** *Journal of Clinical Pathology*, **9**(1):59–65, 1956. 212

727. Lee S.H. and Lewis R.G. **Transmissible resistance factors in isolates of enteropathogenic bacteria.** *Canadian Medical Association Journal*, **100**(3):105–109, 1969. 212

728. De Bruijn F.J. and Bukhari A.I. **Analysis of transposable elements inserted in the genomes of bacteriophages Mu and P1.** *Gene*, **3**(4):315–331, 1978. 213

729. Lautenbach E., Strom B.L., Nachamkin I., Bilker W.B., Marr A.M., Larosa L.A. and Fishman N.O. **Longitudinal trends in fluoroquinolone resistance among Enterobacteriaceae isolates from inpatients and outpatients, 1989-2000: differences in the emergence and epidemiology of resistance across organisms.** *Clinical Infectious Diseases*, **38**(5):655–662, 2004. 213

730. Martínez-Martínez L., Pascual A. and Jacoby G.A. **Quinolone resistance from a transferable plasmid.** *The Lancet*, **351**(9105):797–799, 1998. 213

731. Robicsek A., Jacoby G.A. and Hooper D.C. **The worldwide emergence of plasmid-mediated quinolone resistance**. *The Lancet Infectious Diseases*, **6**(10):629–640, 2006. 213

732. Nordmann P. and Poirel L. **Emergence of plasmid-mediated resistance to quinolones in Enterobacteriaceae**. *The Journal of Antimicrobial Chemotherapy*, **56**(3):463–469, 2005. 213

733. Fan J.B., Chee M.S. and Gunderson K.L. **Highly parallel genomic assays.** *Nature Reviews Genetics*, **7**(8):632–644, 2006. 215

734. Fan J.B., Oliphant A., Shen R., Kermani B.G., Garcia F., Gunderson K.L., Hansen M., Steemers F., Butler S.L., Deloukas P., Galver L., Hunt S., McBride C., Bibikova M., Rubano T., Chen J., Wickham E., Doucet D., Chang W., Campbell D., Zhang B., Kruglyak S., Bentley D., Haas J., Rigault P., Zhou L., Stuelpnagel J. and Chee M.S. **Highly parallel SNP genotyping.** *Cold Spring Harbor Aymposia on Quantitative Biology*, **68**:69–78, 2003. 215

735. Le T.A., Lejay-Collin M., Grimont P.A.D., Hoang T.L., Nguyen T.V., Grimont F. and Scavizzi M.R. **Endemic, Epidemic Clone of *Salmonella enterica* Serovar Typhi Harboring a Single Multidrug-Resistant Plasmid in Vietnam between 1995 and 2002**. *Journal of Clinical Microbiology*, **42**(7):3094–3099, 2004. 215, 278

736. Ling J.M., Lo N.W.S., Ho Y.M., Kam K.M., Hoa N.T., Phi L.T. and Cheng A.F. **Molecular Methods for the Epidemiological Typing of *Salmonella enterica* Serotype Typhi from Hong Kong and Vietnam**. *Journal of Clinical Microbiology*, **38**(1):292–300, 2000. 215, 279

737. Swaddiwudhipong W. and Kanlayanaphotporn J. **A common-source water-borne outbreak of multidrug-resistant typhoid fever in a rural Thai community.** *Journal of the Medical Association of Thailand*, **84**(11):1513–1517, 2001. 216

738. Lewis M.D., Serichantalergs O., Pitarangsi C., Chuanak N., Mason C.J., Regmi L.R., Pandey P., Laskar R., Shrestha C.D. and Malla S. **Typhoid fever: a massive, single-point source, multidrug-resistant outbreak in Nepal.** *Clinical Infectious Diseases*, **40**(4):554–561, 2005. 216

739. Ben Saida N., Mhalla S., Bouzouïa N. and Boukadida J. **Genotypic analysis of *Salmonella enterica* serovar Typhi collected during two successive autumnal typhoid outbreaks in southeast Tunisia.** *Pathologie-biologie*, **55**(7):336–339, 2007. 216

740. Fica A.E., Prat-Miranda S., Fernandez-Ricci A., D'Ottone K. and Cabello F.C. **Epidemic typhoid in Chile: analysis by molecular and conventional methods of *Salmonella typhi* strain diversity in epidemic (1977 and 1981) and nonepidemic (1990) years**. *Journal of Clinical Microbiology*, **34**(7):1701–1707, 1996. 216

741. Cabello F. and Springer A.D. **Typhoid fever in Chile 1977-1990: an emergent disease**. *Revista Médica de Chile*, **125**(4):474–482, 1997. 216

742. Murdoch D., Banatvala N., Bone A., Shoismatulloev B., Ward L. and Threlfall E. **Epidemic ciprofloxacin-resistant in Tajikistan**. *The Lancet*, **351**(9099):339, 1998. 216

743. Hampton M.D., Ward L.R., Rowe B. and Threlfall E.J. **Molecular fingerprinting of multidrug-resistant *Salmonella enterica* serotype Typhi.** *Emerging Infectious Diseases*, **4**(2):317–320, 1998. 216

744. Rathish K.C., Chandrashekar M.R. and Nagesha C.N. **An outbreak of multidrug resistant typhoid fever in Bangalore.** *Indian Journal of Pediatrics*, **62**(4):445–448, 1995. 216

745. Dolecek C., Phi, Rang N.N., Phuong L.T., Vinh H., Tuan P.Q., Cong, Be, Long D.T., Ha L.B., Binh N.T., Anh, Dung P.N., Lanh M.N., Be, Ho V.A., Minh V., Nga T.T., Chau T.T., Schultsz C., Dunstan S.J., Stepniewska K., Campbell J.I., Song, Basnyat B., Vinh, Van Sach N., Chinh N.T., Hien T.T. and Farrar J. **A Multi-Center Randomised Controlled Trial of Gatifloxacin versus Azithromycin for the Treatment of Uncomplicated Typhoid Fever in Children and Adults in Vietnam**. *PLoS ONE*, **3**(5):e2188+, 2008. 216, 218, 244, 247

746. Kelly D.F., Thorson S., Maskey M., Mahat S., Shrestha U., Hamaluba M., Williams E., Dongol S., Werno A.M., Portess H., Yadav B.K., Adhikari N., Guiver M., Thomas K., Murdoch D.R. and Pollard A.J. **The Burden of Vaccine-Preventable Invasive Bacterial Infections in Nepali Children**. Unpublished, 2009. 216, 218, 250

747. Sur D., von Seidlein L., Manna B., Dutta S., Deb A.K., Sarkar B.L., Kanungo S., Deen J.L., Ali M., Kim D.R., Gupta V.K., Ochiai R.L., Tsuzuki A., Acosta C.J., Clemens J.D. and Bhattacharya S.K. **The malaria and typhoid fever burden in the slums of Kolkata, India: data from a prospective community-based study**. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **100**(8):725–733, 2006. 216, 218, 254

748. Teo Y.Y., Inouye M., Small K.S., Gwilliam R., Deloukas P., Kwiatkowski D.P. and Clark T.G. **A genotype calling algorithm for the Illumina BeadArray platform**. *Bioinformatics*, **23**(20):2741–2746, 2007. 223, 228

749. Hornik K. **The R FAQ**, 2009. ISBN 3-900051-08-9. http://CRAN.R-project.org/doc/FAQ/R-FAQ.html. 226

750. Openshaw S., Charlton M., Wymer C. and Craft A. **A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets**. *International Journal of Geographical Information Science*, **1**(4):335–358, 1987. 226

751. Lopez-Qulez V.G.R.J.F.F.A. **Detecting clusters of disease with R**. *Journal of Geographical Systems*, **7**(2):189–206, 2005. 226, 227

752. Carattoli A., Bertini A., Villa L., Falbo V., Hopkins K. and Threlfall E. **Identification of plasmids by PCR-based replicon typing**. *Journal of Microbiological Methods*, **63**(3):219–228, 2005. 236

753. Murdoch D.R., Woods C.W., Zimmerman M.D., Dull P.M., Belbase R.H.A.R.I., Keenan A.J., Scott R.M.C.N.A.I.R., Basnyat B., Archibald L.K. and Reller L.B. **The etiology of febrile illness in adults presenting to Patan Hospital in Kathmandu, Nepal**. *The American Journal of Tropical Medicine and Hygiene*, **70**(6):670–675, 2004. 250

754. Khanal B., Sharma S.K., Bhattacharya S.K., Bhattarai N.R., Deb M. and Kanungo R. **Antimicrobial susceptibility patterns of *Salmonella enterica* Serotype Typhi in eastern Nepal**. *Journal of Health, Population and Nutrition*, **25**(1):82–87, 2007. 252

755. Dunbar S.A. **Applications of Luminex xMAP technology for rapid, high-throughput multiplexed nucleic acid detection.** *Clinical Chimica Acta*, **363**(1-2):71–82, 2006. 252

756. Karmaker S., Biswas D., Shaikh N.M., Chatterjee S.K., Kataria V.K. and Kumar R. **Role of a large plasmid of *Salmonella typhi* encoding multiple drug resistance**. *Journal of Medical Microbiology*, **34**(3):149–151, 1991. 254

757. Sen B., Dutta S., Sur D., Manna B., Deb A.K., Bhattacharya S.K. and Niyogi S.K. **Phage typing, biotyping and antimicrobial resistance profile of *Salmonella enterica* serotype Typhi from Kolkata**. *Indian Journal of Medical Research*, **125**(5):685–688, 2007. 254

758. Martin D. and Austin H. **An efficient program for computing conditional maximum likelihood estimates and exact confidence limits for a common odds ratio.** *Epidemiology*, **2**(5):359–362, 1991. 262

759. **OpenEpi–Epidemiological Calculators 2.3**. Online, http://www.openepi.com/. 262

760. Kariuki S., Mwituria J., Munyalo A., Revathi G. and Onsongo J. **Typhoid is over-reported in Embu and Nairobi, Kenya.** *African Journal of Health Sciences*, **11**(3-4):103–110, 2004. 263

761. Mweu E. and English M. **Typhoid fever in children in Africa.** *Tropical Medicine and International Health*, **13**(4):532–540, 2008. 263

762. Kakai R. **Laboratory diagnostic services in rural health centres, western Kenya.** *East African Medical Journal*, **78**(7):S34–35, 2001. 263

763. Kariuki S., Revathi G., Muyodi J., Mwituria J., Munyalo A., Mirza S. and Hart C.A. **Characterization of multidrug-resistant typhoid outbreaks in Kenya.** *Journal of Clinical Microbiology*, **42**(4):1477–1482, 2004. 263

764. Malla T., Malla K.K., Thapalial A. and Shaw C. **Enteric fever: a retrospective 6-year analysis of 82 paediatric cases in a teaching hospital.** *Kathmandu University Medical Journal*, **5**:181–187, 2007. 275, 278

765. Tjaniadi P., Lesmana M., Subekti D., Machpud N., Komalarini S., Santoso W., Simanjuntak C.H., Punjabi N., Campbell J.R., Alexander W.K., Beecham H.J., Corwin A.L. and Oyofo B.A. **Antimicrobial resistance of bacterial pathogens associated with diarrheal patients in Indonesia.** *The American Journal of Tropical Medicine and Hygiene*, **68**(6):666–670, 2003. 275

766. Isbandrio B.B., Gasem M.H., Dolmans W.M. and Hoogkamp-Korstanje J.A. **Comparative activities of three quinolones and seven comparison standard drugs against *Salmonella typhi* from Indonesia.** *The Journal of Antimicrobial Chemotherapy*, **33**(5):1055–1056, 1994. 275

767. Song J.H., Cho H., Park M.Y., Kim Y.S., Moon H.B., Kim Y.K. and Pai C.H. **Detection of the H1-j strain of *Salmonella typhi* among Korean isolates by the polymerase chain reaction.** *The American Journal of Tropical Medicine and Hygiene*, **50**(5):608–611, 1994. 275

768. Saha S.K., Baqui A.H., Hanif M., Darmstadt G.L., Ruhulamin M., Nagatake T., Santosham M. and Black R.E. **Typhoid fever in Bangladesh: implications for vaccination policy.** *The Pediatric Infectious Disease Journal*, **20**(5):521–524, 2001. 277

769. Ismail A. **An update on diarrhoeal diseases in Malaysia.** *The Southeast Asian Journal of Tropical Medicine and Public Health*, **19**(3):397–400, 1988. 277

770. Pinfold J.V., Horan N.J. and Mara D.D. **Seasonal effects on the reported incidence of acute diarrhoeal disease in northeast Thailand.** *International Journal of Epidemiology*, **20**(3):777–786, 1991. 277

771. Tarr P.E., Kuppens L., Jones T.C., Ivanoff B., Aparin P.G. and Heymann D.L. **Considerations regarding mass vaccination against typhoid fever as an adjunct to sanitation and public health measures: potential use in an epidemic in Tajikistan.** *The American Journal of Tropical Medicine and Hygiene*, **61**(1):163–170, 1999. 277

772. Battikhi M.N. **Occurrence of *Salmonella typhi* and *Salmonella paratyphi* in Jordan.** *The New Microbiologica*, **26**(4):363–373, 2003. 277

773. Siddiqui F.J., Rabbani F., Hasan R., Nizami S.Q. and Bhutta Z.A. **Typhoid fever in children: some epidemiological considerations from Karachi, Pakistan.** *International Journal of Infectious Diseases*, **10**(3):215–222, 2006. 277

774. Vidyalakshmi K., Yashavanth R., Chakrapani M., Shrikala B., Bharathi B., Suchitra U., Dhanashree B. and Dominic R.M. **Epidemiological shift, seasonal variation and antimicrobial susceptibility patterns among enteric fever pathogens in South India.** *Tropical Doctor*, **38**(2):89–91, 2008. 278

775. Hoa N.Q.Q., Larson M., Chuc N.T.K.T., Eriksson B., Trung N.V.V. and Stålsby C.L.L. **Antibiotics and paediatric acute respiratory infections in rural Vietnam: health-care providers' knowledge, practical competence and reported practice.** *Tropical Medicine and International Health*, **14**(5):546–555, 2009. 279

776. Van T.T., Moutafis G., Tran L.T. and Coloe P.J. **Antibiotic resistance in food-borne bacterial contaminants in Vietnam.** *Applied and Environmental Microbiology*, **73**(24):7906–7911, 2007. 279

777. Van T.T., Moutafis G., Istivan T., Tran L.T. and Coloe P.J. **Detection of *Salmonella* spp. in retail raw food samples from Vietnam and characterization of their antibiotic resistance.** *Applied and Environmental Microbiology*, **73**(21):6885–6890, 2007. 279

778. Managaki S., Murata A., Takada H., Tuyen B.C. and Chiem N.H. **Distribution of macrolides, sulfonamides, and trimethoprim in tropical waters: ubiquitous occurrence of veterinary antibiotics in the Mekong Delta.** *Environmental Science and Technology*, **41**(23):8004–8010, 2007. 279

779. Ogasawara N., Tran T.P., Ly T.L., Nguyen T.T., Iwata T., Okatani A.T., Watanabe M., Taniguchi T., Hirota Y. and Hayashidani H. **Antimicrobial susceptibilities of**

Salmonella from domestic animals, food and human in the Mekong Delta, Vietnam. *The Journal of Veterinary Medical Science*, **70**(11):1159–1164, 2008. 279

780. Chadwick P., Groves J.T. and Naylor G.R. **Intermittent urinary carrier of _Salmonella typhi_ detected by urinary antibody determinations and associated case of typhoid fever.** *The Lancet*, **266**(6807):344–345, 1954. 280

781. Mathai E., John T.J., Rani M., Mathai D., Chacko N., Nath V. and Cherian A.M. **Significance of _Salmonella typhi_ bacteriuria.** *Journal of Clinical Microbiology*, **33**(7):1791–1792, 1995. 280

782. Braddick M.R., Crump B.J. and Yee M.L. **How long should patients with _Salmonella typhi_ or _Salmonella paratyphi_ be followed-up? A comparison of published guidelines.** *Journal of Public Health Medicine*, **13**(2):101–107, 1991. 280

783. Dale J.W. **Mobile genetic elements in mycobacteria.** *The European Respiratory Journal*, **20**:633s–648s, 1995. 282

784. Bibb L.A. and Hatfull G.F. **Integration and excision of the _Mycobacterium tuberculosis_ prophage-like element, phiRv1.** *Molecular Microbiology*, **45**(6):1515–1526, 2002. 282

785. Illumina, Inc. **Multiplexed Sequencing with the Illumina Genome Analyzer System**. Technical Report 770-2008-011, Illumina, Inc., 2008. 285

Appendix A

Inactivating mutations in Typhi

Nonsense SNPs are labelled numerically, mutation column gives with the amino acid residue which is mutated. Deletions are numbered alphabetically corresponding to details in Table 2.11; *=deletion is not consistent with single event on phylogenetic tree. Expression gives maximal percentile rank of gene expression level in microarray experiments in the GEO database.

Strain key is:

A - E00-7866
B - E01-6750
C = E02-1180
D = E98-0664
E = E98-2068
F = J185SM
G = M223
H = 404ty
I = AG3
J = E98-3139
K = 150(98)S
L = 8(04)N
M = CT18
N = E02-2759
O = E03-4983
P = E03-9804
Q = ISP-03-07467
R = ISP-04-06979
S = Ty2

ID	Gene ID	Name	Product	Mutation
A	STY0068	*citD2*	citrate lyase acyl carrier protein	C-term del
A	STY0069	*citE2*	citrate lyase beta chain	N-term del
1	STY0089	*yaaU*	putative metabolite transport protein	387
2	STY0212	*sfsA*	sugar fermentation stimulation protein	132
3	STY0336	*safC*	outer-membrane fimbrial usher protein (SPI6)	347
4	STY0349	*tinR*	transcriptional regulator (SPI6)	141
5	STY0371	*stbC*	outer membrane fimbrial usher protein	155
6	STY0428	*araJ*	transport protein, potentially sugar efflux protein	202
7	STY0519	*acrB*	acriflavin resistance protein B	769
8	STY0519	*acrB*	acriflavin resistance protein B	
9	STY0590	*fimI*	fimbrin-like protein	47
10	STY0682		sec-independent protein translocase protein	68
11	STY0971	*sopD2*	secreted effector protein sopD homolog	144
12	STY0992	*ycbC*	putative membrane protein	174
B	STY1131	*hpaB*	4-hydroxyphenylacetate 3-monooxygenase	N-term del
13	STY1167		hypothetical protein	63
14	STY1204		putative membrane transporter	188
15	STY1260		putative ROK-family protein	113
16	STY1304	*oppA*	periplasmic oligopeptide-binding protein	559
17	STY1326	*trpC*	indole-3-glycerol phosphate synthase	15
18	STY1410	*dbpA*	ATP-dependent RNA helicase	30
19	STY1457		putative lipoprotein	185
20	STY1476		putative NADP-dependent oxidoreductase	168
C	STY1485	*narV*	respiratory nitrate reductase 2 gamma chain	N-term del
C	STY1486	*narW*	respiratory nitrate reductase 2 delta chain	C-term del
21	STY1488	*narZ*	respiratory nitrate reductase 2 alpha chain	870
D	STY1503	*glgX*	putative hydrolase	deleted
22	STY1507		putative aminotransferase	251
E	STY1507		putative aminotransferase	N-term del
E	STY1508		putative transport protein	deleted
E	STY1509		hypothetical protein	N-term del
G	STY1536		putative aldehyde-dehydrogenase	C-term del
23	STY1553		putative D-mannonate oxidoreductase	182
H	STY1568	*dmsC*	putative dimethyl sulphoxide reductase subunit	deleted
H	STY1569		hypothetical protein	deleted
H	STY1570		putative ABC transporter membrane protein	deleted
H	STY1571		putative ABC transporter periplasmic binding protein	deleted
H	STY1572		putative ABC transporter membrane protein	deleted
H	STY1573		putative ABC transporter ATP/GTP-binding protein	deleted
H	STY1574		putative voltage gated chloride channel protein	deleted
H	STY1575		putative dethiobiotin synthetase	N-term del
24	STY1587		putative membrane protein	304
I	STY1648		hypothetical protein	N-term del

ID	Gene ID	Strain(s) with the mutation	Expression
A	STY0068	O	98
A	STY0069	O	97
1	STY0089	M	no data
2	STY0212	D	94
3	STY0336	C	98
4	STY0349	F, H, O	100
5	STY0371	S	98
6	STY0428	J, G, D, M, E, F, H, O, S, B, H58/H63	94
7	STY0519	G	94
8	STY0519	I	94
9	STY0590	M	100
10	STY0682	C, A	98
11	STY0971	H	94
12	STY0992	J	90
B	STY1131	M, E, F, H, O, S, B, H58/H63	no data
13	STY1167	J	95
14	STY1204	*B, E	95
15	STY1260	E	98
16	STY1304	J	68
17	STY1326	H, O	89
18	STY1410	H58/H63	98
19	STY1457	H58/H63	97
20	STY1476	A	81
C	STY1485	S, B, H58/H63	94
C	STY1486	S, B, H58/H63	94
21	STY1488	E	100
D	STY1503	C	92
22	STY1507	H, O	98
E	STY1507	H58/H63	98
E	STY1508	H58/H63	no data
E	STY1509	H58/H63	no data
G	STY1536	*M, F	no data
23	STY1553	S, B, H58/H63	94
H	STY1568	L	no data
H	STY1569	L	90
H	STY1570	L	92
H	STY1571	L	89
H	STY1572	L	89
H	STY1573	L	92
H	STY1574	L	92
H	STY1575	L	98
24	STY1587	P	95
I	STY1648	*B, A, C	98

ID	Gene ID	Name	Product	Mutation
I	STY1649		outer membrane protein	deleted
I	STY1650		hypothetical protein	N-term del
25	STY1683		putative oxidoreductase	3
26	STY1693	*ydhB*	putative transcriptional regulator	2
27	STY1738	*ttrA*	tetrathionate reductase subunit A (SPI2)	622
28	STY1924	*treA*	periplasmic trehalase	92
29	STY1977	*proP*	proP effector	203
30	STY2005		conserved hypothetical protein	179
K	STY2238	*cbiC*	precorrin-8X methylmutase	N-term del
31	STY2398	*pbpG*	penicillin-binding protein	5
32	STY2501		putative transmembrane transport protein	85
33	STY2506	*nrdA*	ribonucleoside-diphosphate reductase 1 alpha chain	762
34	STY2526	*ais*	Ais protein	34
L	STY2717	*aegA*	putative oxidoreductase	(internal deletion)
M	STY2791		putative RNA methyltransferase	(internal deletion)
35	STY2838	*yfiK*	putative membrane protein	85
36	STY2877		putative type I secretion protein (SPI9)	719
37	STY2913	*gabP*	GabA permease (4-amino butyrate transport carrier)	244
38	STY3001	*stpA*	tyrosine phosphatase (translational regulation) (SPI1)	322
39	STY3001	*stpA*	tyrosine phosphatase (translational regulation) (SPI1)	185
40	STY3034		hypothetical protein	74
41	STY3049	*rpoS*	RNA polymerase sigma subunit RpoS (sigma-38)	52
42	STY3049	*rpoS*	RNA polymerase sigma subunit RpoS (sigma-38)	43
43	STY3138	*ppdA*	prepilin peptidase dependent protein A precursor	141
44	STY3352		possible AraC-family transcriptional regulator	119
45	STY3428	*tdcA*	TDC operon transcriptional activator	277
46	STY3508		hypothetical protein	228
N	STY3617		hypothetical protein	N-term del
N	STY3618		putative membrane protein	deleted
47	STY4008		putative inner membrane transport protein	326
IS	STY4123	*yiaO*	putative periplasmic protein	IS1 insertion
48	STY4134	*malS*	alpha-amylase	376
49	STY4162	*yhjW*	putative membrane protein	488
O	STY4162	*yhjW*	putative membrane protein	
50	STY4329	*yhfK*	hypothetical protein	685
R	STY4575		hypothetical protein (SPI7)	deleted
R	STY4576		hypothetical protein (SPI7)	deleted
R	STY4577		hypothetical protein (SPI7)	deleted
R	STY4578		putative membrane protein (SPI7)	deleted
R	STY4579		putative membrane protein (SPI7)	deleted
S	STY4580		putative membrane protein (SPI7)	deleted
S	STY4582		possible exported protein (SPI7)	deleted
51	STY4728	*yjfJ*	hypothetical protein	184

ID	Gene ID	Strain(s) with the mutation	Expression
I	STY1649	*B, A, C	100
I	STY1650	*B, A, C	87
25	STY1683	B	87
26	STY1693	E	90
27	STY1738	B	92
28	STY1924	J, G, D, M, E, F, H, O, S, B, H58/H63	98
29	STY1977	O	97
30	STY2005	J	98
K	STY2238	J, G, D, M, E, F, H, O, S, B, H58/H63	no data
31	STY2398	J, G, D, M, E, F, H, O, S, B, H58/H63	97
32	STY2501	J, G, D, M, E, F, H, O, S, B, H58/H63	97
33	STY2506	A	90
34	STY2526	C	90
L	STY2717	J, G, D, M, E, F, H, O, S, B, H58/H63	95
M	STY2791	A	98
35	STY2838	C	98
36	STY2877	D, M, E, F, H, O, S, B, H58/H63	94
37	STY2913	S, B, H58/H63	95
38	STY3001	J	100
39	STY3001	H58/H63	100
40	STY3034	H, O	no data
41	STY3049	L	98
42	STY3049	K	98
43	STY3138	C	89
44	STY3352	A	92
45	STY3428	C	94
46	STY3508	C	95
N	STY3617	N, P	92
N	STY3618	N, P	97
47	STY4008	G	94
IS	STY4123	M	95
48	STY4134	F, H, O	94
49	STY4162	C	95
O	STY4162	J, G, D, M, E, F, H, O, S, B, H58/H63	95
50	STY4329	C	90
R	STY4575	A	95**
R	STY4576	A	95**
R	STY4577	A	98**
R	STY4578	A	94**
R	STY4579	A	94**
S	STY4580	S, B	98**
S	STY4582	S, B	95**
51	STY4728	C, A	95

ID	Gene ID	Name	Product	Mutation
P	STY4728	*yjfJ*	hypothetical protein	deleted
P	STY4728a		hypothetical protein	deleted
P	STY4729	*yjfK*	hypothetical protein	N-term del
Q	STY4786		hypothetical protein	C-term del
Q	STY4787		putative BglB-family transcriptional antiterminator	N-term del
52	STY4811		putative exported protein	36
53	STY4849		helicase related protein (SPI10)	635
54	STY4849		helicase related protein (SPI10)	381
55	STY4853		hypothetical protein	180

ID	Gene ID	Strain(s) with the mutation	Expression
P	STY4728	J	95
P	STY4728a	J	no data
P	STY4729	J	98
Q	STY4786	A	73*
Q	STY4787	A	90
52	STY4811	G	97
53	STY4849	J	89
54	STY4849	C, A	89
55	STY4853	A	95

Appendix B

Paratyphi A isolates sequenced in pools

* = Isolate also sequenced individually. Kolkata isolates were provided by Shanta Dutta, National Institute of Cholera and Enteric Diseases, Kolkata; Delhi isolates by Rajni Gaind, Safdarjung Hospital, Delhi; and Karachi isolates by Rumina Hasan, Aga Khan University Hospital, Karachi. Isolates in pools MA1-18 are part of the *Salmonella* collection at the Pasteur Institute, Paris and DNA was provided by Francois-Xavier Weill.

Pool	Strains	Year	Country	Region
JW1	B1357	2004/5	India	Kolkata
	D2383	2004/5	India	Kolkata
	D1985	2004/5	India	Kolkata
	B943	2004/5	India	Kolkata
	C4672	2004/5	India	Kolkata
JW2	A1338	2004/5	India	Kolkata
	AKU_12601*	2002	India	Kolkata
	B4173	2004/5	India	Kolkata
	B418	2004/5	India	Kolkata
	B964	2004/5	India	Kolkata
	D441	2004/5	India	Kolkata

Pool	Strains	Year	Country	Region
JW3	A1345	2004/5	India	Kolkata
	A2248	2004/5	India	Kolkata
	B4986	2004/5	India	Kolkata
	C806	2004/5	India	Kolkata
	D4075	2004/5	India	Kolkata
	F846	2004/5	India	Kolkata
JW4	2129	2004/5	Kuwait	-
	BL8758*	2004/5	Pakistan	Karachi
	BL4595	2004/5	Pakistan	Karachi
	BL14275	2004/5	Pakistan	Karachi
	58/38	2004/5	India	Delhi
	138/69	2004/5	India	Delhi
JW5	2664	2004/5	India	-
	BL1344	2004/5	Pakistan	Karachi
	BL4579	2004/5	Pakistan	Karachi
	B7697	2004/5	UK	-
	6911*	2007	Kenya	-
	6912*	2007	Kenya	-
JW6	181	2005	India	Delhi
	11	2005	India	Delhi
	1	2005	India	Delhi
	331-32	2005	India	Delhi
	5	2005	India	Delhi
	40	2005	India	Delhi
JW7	4	2007	India	Delhi
	83	2007	India	Delhi
	56	2007	India	Delhi
	105	2007	India	Delhi
	123	2007	India	Delhi
	31	2007	India	Delhi

Pool	Strains	Year	Country	Region
JW8	38/1/06	2006	India	Delhi
	181/8/06	2006	India	Delhi
	56/8/05	2005	India	Delhi
	68/1/04	2004	India	Delhi
	92/5/03	2002/3	India	Delhi
JW9	BL28008	2002/3	Pakistan	Karachi
	BL27136	2002/3	Pakistan	Karachi
	BL23318	2002/3	Pakistan	Karachi
	BL1893	2002/3	Pakistan	Karachi
	2460	2005/6	Kuwait	-
	1540	2005/6	Kuwait	-
MA1	A68-37	1968	Cambodia	-
	Banker Type 1	1943	India	-
	99 3482	1999	India	-
	9910258	1999	Indonesia	-
	99 7863	1999	Sri Lanka	-
	00 6053	2000	India	-
MA2	99 5900	1999	Cambodia	-
	WS0783	1925	Palestine	-
	02 1960	2002	Cambodia	-
	05 3784	2005	Cambodia	-
	8-58	1958	Cambodia	-
	06-610	2006	Cambodia	-
MA3	Banker Type 3	1954	India	-
	01 5766	2001	India	-
	04 2589	2004	India	-
	04 6500	2004	India	-
	05 5304	2005	India	-
	05 0208	2005	India	-

Pool	Strains	Year	Country	Region
MA4	A65-4	1965	Cambodia	-
	Banker Type 4	1954	Egypt	-
	98 9652	1998	Morocco	-
	99 0915	1999	Nepal	-
	A52-409	1952	Turkey	-
	9-65	1965	Turkey	-
MA5	99 7158	1999	Thailand	-
	04 9176	2004	Indonesia	-
	01 7057	2001	Indonesia	-
	03 9604	2003	Burma	-
	05 6721	2005	Burma	-
	Banker Type 5	1955-1962	Indonesia	-
MA6	A63-3 (Banker Type 6)	1963	Vietnam	-
	WS0179	1946	Vietnam	-
	9-63	1963	Vietnam	-
	05 0473	2005	Vietnam	-
	06 7153	2006	India	-
	06-4418	2006	India	-
MA7	00 3513	2000	Cambodia	-
	00 6735	2000	Chad	-
	97 1822	1997	India	-
	97 7358	1997	India	-
	98 2812	1998	Pakistan	-
	03 7001	2003	Pakistan	-
MA8	97 0613	1997	Pakistan	-
	9710913	1997	Pakistan	-
	99 7252	1999	Turkey	-
	01 8877	2001	Pakistan	-
	05 6792	2005	Pakistan	-
	04 0406	2004	Turkey	-

Pool	Strains	Year	Country	Region
MA9	A62-5	1962	Algeria	-
	A61-79 (7-58)	1958	France	-
	A61-149 (A50-4)	1950	France	-
	A61-81 (9-58)	1958	France	-
	A61-145 (4-51)	1951	Vietnam	-
	WS0784	1967	France	-
MA10	A63-72	1963	Cambodia	-
	A 63-73	1963	France	-
	WS0785	1917	Albania	-
	WS0065	1971	US	-
	02 1383	2002	Guinea	-
	02 8292	2002	Guinea	-
MA11	J307	1949	Algeria	-
	J852	1949	Algeria	-
	6-60	1960	Brasil	-
	4-56	1956	Ethiopia	-
	A52-428	1952	Turkey	-
	A61-125 (3-52)	1952	Turkey	-
MA12	A103	1945	France	-
	A506	1949	France	-
	Reil 18	1949	France	-
	Reil 90	1949	France	-
	Brun12	1949	France	-
	Balt8	1949	Iran	-
MA13	1-57	1957	Morocco	-
	A61-139 (20-52)	1952	Senegal	-
	7-54	1954	Tunisia	-
	00 2046	2000	Guinea	-
	00 6712	2000	Morocco	-
	00 5851	2000	Nepal	-

Pool	Strains	Year	Country	Region
MA14	A80-82	1980	Brasil	-
	A80-2	1980	Togo	-
	A73-2	1973	Morocco	-
	A77-46	1977	Mali	-
	02 4282	2002	Mali	-
	06 6491	2006	Mali	-
MA15	06-1246	2006	India	-
	06-5568	2006	India	-
	5-66	1966	Senegal	-
	01 4749	2001	Nepal	-
	01 8552	2001	Peru	-
MA16	05 6761	2005	Bengladesh	-
	01 6979	2001	China	-
	04 3588	2004	Nepal	-
	05 7465	2005	Nepal	-
	WS0063, WS 0178	1899	US	-
	17-66	1966	Morocco	-
MA17	06-6204	2006	Turkey	-
	06-2861	2006	Pakistan	-
	06-2633	2006	Senegal	-
	01 1852	2001	Senegal	-
	04 6031	2004	Senegal	-
	06 0906	2006	Senegal	-
MA18	A80-26	1980	Congo	-
	02 7555	2002	Benin	-
	WS0782	1934	Jordania	-
	WS0064	pre 1963	-	-
	SARB42/ATCC9150*	pre 1993	-	-
	02 2076	2002	Indonesia	-

Appendix C

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

Gene IDs correspond to AKU_12601 annotation. Gene symbol is given where possible.

Gene ID	Symbol	Product
SSPA0020		Fimbrial chaperone
SSPA0061	*citG*	CitG protein
SSPA0101	*araB*	L-ribulokinase
SSPA0103		DedA family integral membrane protein
SSPA0107		Putative ABC transporter periplasmic solute binding protein
SSPA0157		Putative exported protein
SSPA0163		Putative transcriptional regulator
SSPA0193	*fhuC*	Ferrichrome transport ATP-binding protein FhuC
SSPA0196	*stfA*	Putative fimbrial subunit
SSPA0277		Putative oxidoreductase
SSPA0316		Putative lipoprotein
SSPA0319		Putative anaerobic reductase component
SSPA0331		Putative exported protein
SSPA0348		Putative arsenate reductase
SSPA0361		Succinyl-diaminopimelate desuccinylase
SSPA0457		Div protein

Gene ID	**Symbol**	**Product**
SSPA0477		Histidine transport ATP-binding protein
SSPA0483		Putative transcriptional regulator
SSPA0489		Putative membrane protein
SSPA0511		NADH dehydrogenase I chain M
SSPA0555		Hypothetical protein
SSPA0559		Regulator of capsule synthesis B component
SSPA0583		Putative sulphatase
SSPA0613		Colicin I receptor
SSPA0643		D-lactate dehydrogenase
SSPA0661		Putative outer membrane usher protein
SSPA0670		Fructose-bisphosphate aldolase class I
SSPA0696	*mdtC*	Putative RND-family transporter protein
SSPA0698		Putative uncharacterized protein
SSPA0719		Phosphomannomutase
SSPA0720		Putative transmembrane transport protein
SSPA0728		Glucose-1-phosphate cytidylyltransferase
SSPA0733	*rfbX*	Putative O-antigen transporter
SSPA0735		Putative glycosyltransferase
SSPA0739		Undecaprenyl-phosphate galactosephosphotransferase
SSPA0764	*pduT*	Putative propanediol utilization protein PduT
SSPA0791		Synthesis of vitamin B12 adenosyl cobalamide precursor
SSPA0827		Colanic acid capsullar biosynthesis activation protein A
SSPA0850	*fliC*	Flagellin
SSPA0856		Putative ABC transport ATP-binding protein
SSPA0860		Invasion response-regulator
SSPA0910		High-affinity zinc uptake system membrane protein
SSPA0931		Hypothetical protein
SSPA0953		Hypothetical protein
SSPA0956		Putative membrane protein
SSPA0957a	*proQ*	ProP effector
SSPA1014	*nifE*	Hydrogenase-1 large chain (NifE hydrogenase)
SSPA1032	*narL*	Nitrate/nitrite response regulator protein NarL
SSPA1072		Anthranilate phosphoribosyltransferase
SSPA1083		Aconitate hydratase 1 (Citrate hydro-lyase 1)
SSPA1095		Enoyl-[acyl-carrier-protein] reductase (NADH)
SSPA1120	*mppA*	Periplasmic murein peptide-binding protein MppA
SSPA1125a		Hypothetical protein
SSPA1136		Fumarate and nitrate reduction regulatory protein
SSPA1149		Putative periplasmic protein
SSPA1166		Hypothetical protein
SSPA1188		Putative regulatory protein

Gene ID	Symbol	Product
SSPA1196		Putative uncharacterized protein
SSPA1198		Respiratory nitrate reductase 2 beta chain
SSPA1201		Nitrite extrusion protein
SSPA1220		Putative aminotransferase
SSPA1271		Putative regulatory protein
SSPA1328	*mdtK*	Hypothetical integral membrane protein
SSPA1386		Phosphoenolpyruvate synthase
SSPA1434		Putative MutT-family protein
SSPA1445		Glyceraldehyde 3-phosphate dehydrogenase A
SSPA1451		Diguanylate cyclase/phosphodiesterase domain 1
SSPA1490		Putative toxin-like protein
SSPA1505	*phoQ*	Sensor protein PhoQ, regulator of virulence determinants
SSPA1519	*trcF*	Transcription-repair coupling factor (TrcF)
SSPA1531a	*fhuE*	FhuE receptor precursor
SSPA1533		Putative uncharacterized protein
SSPA1574		Putative cytochrome
SSPA1576		Hypothetical protein
SSPA1584		Putative uncharacterized protein
SSPA1594		Putative uncharacterized protein
SSPA1605		Proline dehydrogenase (Proline oxidase)
SSPA1613a	*scsB*	Membrane protein, suppressor for copper-sensitivity B precursor
SSPA1724	*clpA*	ATP-dependent Clp protease ATP-binding subunit ClpA
SSPA1731		Putative periplasmic protein
SSPA1742		Arginine-binding periplasmic protein 1
SSPA1754	*potG*	Putrescine transport ATP-binding protein PotG
SSPA1777		Hypothetical ABC transporter ATP-binding protein
SSPA1809		Hypothetical protein
SSPA1820a	*slrP*	Leucine-rich repeat protein SlrP
SSPA1838		Molybdate-binding periplasmic protein
SSPA1844		Galactose-1-phosphate uridylyltransferase
SSPA1899	*kdpD*	Sensor protein KdpD
SSPA1900		KDP operon transcriptional regulatory protein
SSPA1902		Putrescine-ornithine antiporter
SSPA1915		Putative outer membrane protein
SSPA1928a	gltJ	Glutamate/aspartate transport system permease protein GltJ
SSPA1968		Ribonuclease I
SSPA1984		Carbon starvation protein A
SSPA1993	*fepG*	Ferric enterobactin transport protein FepG
SSPA1999		Ferrienterobactin receptor
SSPA2034	*lpxH*	Putative uncharacterized protein
SSPA2045a	*ybbW*	Putative allantoin permease

Gene ID	Symbol	Product
SSPA2074		Inosine-guanosine kinase
SSPA2167		Branched chain amino acid transport system II carrier protein
SSPA2171	*sbcC*	Exonuclease SbcC
SSPA2191		Hypothetical protein
SSPA2212		Possible transcriptionl regulator
SSPA2221		Putative fimbrial protein
SSPA2277		Aminoacyl-histidine dipeptidase
SSPA2292		Outermembrane fimbrial usher protein
SSPA2318		Hypothetical protein
SSPA2460		Putative nickel transporter
SSPA2509		Glucitol operon repressor
SSPA2563		Surface presentation of antigens protein
SSPA2592	*rpoS*	RNA polymerase sigma subunit RpoS (Sigma-38)
SSPA2598	*ispD*	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase
SSPA2613a		Similar to putative metal-dependent hydrolases of the beta-lactamase superfamily II
SSPA2617	*pyrG*	CTP synthetase
SSPA2620		Outer membrane usher protein
SSPA2639		Putative serine transporter
SSPA2644		Fuculose-1-phosphate aldolase
SSPA2663		Protease III (Pitrilysin)
SSPA2664		Exonuclease V subunit
SSPA2745		Possible ABC-transport protein, ATP-binding component
SSPA2775		Nucleoside permease
SSPA2794		Putative oxidoreductase
SSPA2805		Glutathionylspermidine synthetase/amidase
SSPA2807		Hypothetical protein
SSPA2848		Putative uncharacterized protein
SSPA2862		Putative membrane protein
SSPA2873	*gcp*	Possible glycoprotease
SSPA2908		TDC operon transcriptional activator
SSPA2916		PTS system, sugar phosphotransferase enzyme IIBC component
SSPA3036	*prmA*	Ribosomal protein L11 methyltransferase
SSPA3043		RND family, multidrug transport protein,acriflavin resistance protein F
SSPA3087		Type III leader peptidase
SSPA3110		Cyclic AMP receptor protein,catabolite gene activator
SSPA3120		Putative nitrite transporter
SSPA3122		Tryptophanyl-tRNA synthetase
SSPA3133		Penicillin-binding protein 1A
SSPA3153		4-alpha-glucanotransferase
SSPA3155	*malT*	MalT regulatory protein
SSPA3162	*glgC*	Glucose-1-phosphate adenylyltransferase

Gene ID	Symbol	Product
SSPA3163	*glgX*	Glycogen operon protein
SSPA3164		1,4-alpha-glucan branching enzyme
SSPA3202		Putative lipoprotein
SSPA3209		Hypothetical ABC transporter ATP-binding protein
SSPA3215		Hypothetical protein
SSPA3234		Hypothetical luxR-family transcriptional regulator
SSPA3240	*dctA*	C4-dicarboxylate transport protein
SSPA3295		Putative transcriptional regulator
SSPA3307		L-lactate permease
SSPA3308		Putative L-lactate dehydrogenase operon regulator
SSPA3336		Lipopolysaccharide core biosynthesis protein
SSPA3354	*ligB*	Putative DNA ligase
SSPA3369		Hypothetical protein
SSPA3432		Two-component sensor protein histidine protein kinase
SSPA3448		Putative hydrolase
SSPA3475	*rbsD*	High affinity ribose transport protein RbsD
SSPA3480		Hypothetical 20.8 kDa protein in rbsr-rrsc intergenic region
SSPA3484		Putative magnesium chelatase, subunit ChlI
SSPA3499	*gppA*	Guanosine-5'-triphosphate,3'-diphosphate pyrophosphatase
SSPA3511	*wecF/rffT*	Putative uncharacterized protein
SSPA3531		Magnesium and cobalt transport protein
SSPA3555		Putative deoxyribonuclease
SSPA3629		Two-component sensor kinase protein
SSPA3646		Putative ABC transporter permease protein
SSPA3655		Putative glycerol metabolic protein
SSPA3725		Two-component system sensor protein
SSPA3824		Melibiose carrier protein
SSPA3864		Fumarate reductase, flavoprotein subunit
SSPA3866		Putative amino acid permease
SSPA3893	*aidB*	Probable acyl Co-A dehydrogenase
SSPA3921		2',3'-cyclic-nucleotide 2'-phosphodiesterase
SSPA3963		Carbamate kinase
SSPA4071	*lplA*	Lipoate-protein ligase A

Appendix D

Variable pseudogenes in the Paratyphi A population

N = number of nonsense SNPs, Freq = SNP frequency or ‘d’ for deletion.

Gene ID	N	Freq	Symbol	Product
SSPA0005a	-	d		Fimbrial chaperone
SSPA0020	1	1		Fimbrial chaperone
SSPA0021	1	15		Fimbrial usher
SSPA0068	1	2	*caiC*	Probable crotonobetaine/carnitine-CoA ligase
SSPA0078	1	2		Probable secreted protein
SSPA0103	1	5		DedA family integral membrane protein
SSPA0107	1	2		Putative ABC transporter periplasmic solute binding protein
SSPA0209	1	1		Hypothetical protein
SSPA0223	1	3	*yaeT*	Outer membrane protein
SSPA0304	1	1	*iscR*	Putative uncharacterized protein
SSPA0330	1	1		Putative exported protein
SSPA0337	1	1		Putative outer membrane lipoprotein
SSPA0367	1	1		Putative exported protein
SSPA0376	1	2		Putative phosphate acyltransferase
SSPA0381	1	2		Putative alchohol dehydrogenase
SSPA0387	1	1	*eutK*	Ethanolamine utilization protein EutK
SSPA0401	1	2		Putative exported protein
SSPA0422	1	1		Putative decarboxylase
SSPA0470	1	24		Hypothetical protein
SSPA0477	1	1		Histidine transport ATP-binding protein
SSPA0526	1	2		Melittin resistance protein PqaB

Gene ID	N	Freq	Symbol	Product
SSPA0546	1	1		Glycerol-3-phosphate transporter
SSPA0558	1	2	*rscC*	Sensor protein RcsC
SSPA0560	1	2		Putative two-component system sensor kinase
SSPA0579a	-	d	*ccmH*	Cytochrome c-type biogenesis protein H1
SSPA0604	1	1		Putative membrane protein
SSPA0615a	1	1	*sdaC*	Putative L-serine dehydratase
SSPA0620	2	1,3		Mgl repressor and galactose ultrainduction factor
SSPA0621	1	7		D-galactose-binding periplasmic protein precursor
SSPA0638	1	5		Putative lipoprotein
SSPA0662	1	2		Putative exported protein
SSPA0670	1	1		Fructose-bisphosphate aldolase class I
SSPA0689	1	2		Putative exported protein
SSPA0708	-	d	*wcaA*	Hypothetical protein
SSPA0719a	1	14	*wcaJ*	Putative extracellular polysaccharide biosynthesis protein
SSPA0720	-	d		Membrane transport protein
SSPA0728	1	3		Glucose-1-phosphate cytidylyltransferase
SSPA0735	1	30		Putative glycosyltransferase
SSPA0738	1	6		Phosphomannomutase
SSPA0768	1	15		Putative uncharacterized protein
SSPA0775	1	16	*pduG*	Propanediol dehydratase reactivation protein
SSPA0780a	-	d	*pduF*	Propanediol diffusion facilitator
SSPA0790	1	1		Synthesis of vitamin B12 adenosyl cobalamide
SSPA0791	1	1		Synthesis of vitamin B12 adenosyl cobalamide precursor
SSPA0795	1	4	*cbiQ*	Putative cobalt transport protein CbiQ
SSPA0802a	-	d	*yeeO*	Putative inner membrane protein
SSPA0816a	1	1		Putative membrane protein
SSPA0938	1	2		Putative hydrolase
SSPA0951	1	1		Serine/threonine protein phosphatase 1
SSPA0957a	2	1,1	*proQ*	ProP effector
SSPA0978	1	1	*nudL*	Putative uncharacterized protein
SSPA1000	1	1		Alanine racemase
SSPA1002	1	2		Hypothetical protein
SSPA1072	2	3,5		Anthranilate phosphoribosyltransferase
SSPA1083	1	1		Aconitate hydratase 1 (Citrate hydro-lyase 1)
SSPA1180	1	57		Putative inner membrane protein
SSPA1197a	-	d	*narW*	Respiratory nitrate reductase 2 delta chain
SSPA1204a	-	d	*nmpC*	Outer membrane porin
SSPA1217	1	4		Putative hydrolase
SSPA1227	1	1	*hyaE2*	Hydrogenase-1 operon protein HyaE2
SSPA1249	1	1		Putative oxidoreductase
SSPA1267	1	1		Putative ABC transporter membrane protein
SSPA1285	1	4		Two component sensor kinase
SSPA1311	1	10		Putative HlyD-family protein
SSPA1391	2	1,1		Hypothetical protein

Gene ID	N	Freq	Symbol	Product
SSPA1447	1	128		Putative oxidoreductase
SSPA1449a	-	d	*yeaG*	Putative uncharacterized protein
SSPA1578	1	1	*mdtG*	Putative membrane transport protein
SSPA1590	1	2		Putative regulatory protein
SSPA1706	1	1		Anaerobic dimethyl sulfoxide reductase chain A
SSPA1724	1	1	*clpA*	ATP-dependent Clp protease ATP-binding subunit ClpA
SSPA1773	1	3		Putative uncharacterized protein
SSPA1786	1	2		Putative exported protein
SSPA1797	1	2		Hypothetical Zinc-finger containing protein
SSPA1798	1	1		Hypothetical protein
SSPA1829a	1	1	*hutU*	Urocanate hydratase
SSPA1832	1	5	*hutI*	Imidazolonepropionase
SSPA1902	1	1		Putrescine-ornithine antiporter
SSPA1906	1	1	*ybfF*	Putative esterase/lipase YbfF
SSPA1915	1	2		Putative outer membrane protein
SSPA1921a	-	d	*asnB*	Asparagine synthetase B
SSPA1928a	3	1,1,8	*gltJ*	Glutamate/aspartate transport system permease protein GltJ
SSPA2005	1	1		Oxygen-insensitive NAD(P)H nitroreductase
SSPA2007	1	3		Putative membrane protein
SSPA2017a	3	4,5,5		Putative membrane protein
SSPA2090	1	3	*acrB*	Acriflavin resistance protein B
SSPA2100	-	d	*tesB*	Acyl-CoA thioesterase II
SSPA2118	1	2	*bolA*	Transcriptional regulator
SSPA2182	1	1	*psiF*	Phosphate starvation-inducible protein PsiF
SSPA2185a	2	2,2	*ddlA*	D-alanine:D-alanine ligase A
SSPA2334	1	2		Hypothetical protein
SSPA2338	1	4		Putative uncharacterized protein
SSPA2375	1	4		Putative type I secretion protein, ATP-binding protein
SSPA2376	1	1		Putative type I secretion protein
SSPA2592	3	2,2,2	*rpoS*	RNA polymerase sigma subunit RpoS (Sigma-38)
SSPA2598	1	1	*ispD*	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase
SSPA2606	1	2		Hypothetical protein
SSPA2630	1	2		Sensor protein
SSPA2643	-	d		Lactaldehyde reductase
SSPA2663	1	1		Protease III (Pitrilysin)
SSPA2760	1	1		Probable global regulatory protein homolog
SSPA2775	1	1		Nucleoside permease
SSPA2807a	1	1		Possible ABC-transport protein, periplasmic-binding component
SSPA2848	1	1		Putative uncharacterized protein
SSPA2877	1	14		G/U mismatch-specific DNA glycosylase
SSPA2900a	1	19		Hypothetical transport protein
SSPA2907	1	1		Catabolic threonine dehydratase
SSPA2916	1	2		PTS system, sugar phosphotransferase enzyme IIBC component
SSPA2939	1	1		Putative uncharacterized protein

Gene ID	N	Freq	Symbol	Product
SSPA2987	1	1		Hypothetical protein
SSPA3040	1	1		Diguanylate cyclase/phosphodiesterase domain 2
SSPA3087	1	2		Type III leader peptidase
SSPA3117	1	1	*tsgA*	Putative membrane protein
SSPA3146	1	1		Ferrous iron transport protein B
SSPA3155	1	1	*malT*	MalT regulatory protein
SSPA3202	-	d		Putative lipoprotein
SSPA3209	1	15		Hypothetical ABC transporter ATP-binding protein
SSPA3240	1	1	*dctA*	C4-dicarboxylate transport protein
SSPA3259a	1	18	*yhjW*	Putative membrane protein
SSPA3281	1	3		Putative exported amidase
SSPA3307	1	2		L-lactate permease
SSPA3308	1	1		Putative L-lactate dehydrogenase operon regulator
SSPA3329	1	1	*waaK*	Lipopolysaccharide 1,2-N-acetylglucosaminetransferase
SSPA3370	1	2		Putative autotransported protein
SSPA3432	1	1		Two-component sensor protein histidine protein kinase
SSPA3446	1	1		Hypothetical protein
SSPA3448	1	1		Putative hydrolase
SSPA3475	1	9	*rbsD*	High affinity ribose transport protein RbsD
SSPA3476a	-	d	*rbsC*	High affinity ribose transport protein RbsC
SSPA3478a	-	d	*rbsR*	Ribose operon repressor
SSPA3484	1	2	*chlI*	Putative magnesium chelatase, subunit ChlI
SSPA3485	1	1		Acetolactate synthase large
SSPA3499	1	1	*gppA*	Guanosine-5'-triphosphate,3'-diphosphate pyrophosphatase
SSPA3499a	-	d	*rhlB*	ATP-dependent RNA helicase
SSPA3565	1	2		Molybdopterin-guanine dinucleotide biosynthesis protein B
SSPA3578	1	2		Glutamine synthetase
SSPA3581	1	93		Hypothetical protein
SSPA3616	1	3	*rhaD*	Rhamnulose-1-phosphate aldolase
SSPA3629	1	1		Two-component sensor kinase protein
SSPA3676	1	4		Putative GntR-family regulatory protein
SSPA3720	1	3	*nfi*	Putative endonuclease
SSPA3784	1	15		Putative type-I secretion protein
SSPA3871	1	2	*artJ*	Probable arginine-binding periplasmic protein
SSPA3893	-	d	*aidB*	Probable acyl Co-A dehydrogenase
SSPA3921	1	3		2',3'-cyclic-nucleotide 2'-phosphodiesterase
SSPA3950	1	1		Anaerobic ribonucleoside-triphosphate reductase
SSPA3979	1	2		GntP family, L-idonate transport protein
SSPA4005	1	1		Putative uncharacterized protein
SSPA4022	1	1		Probable transcriptional activator
SSPA4028	1	1		Putative membrane protein
SSPA4030	1	3		Putative uncharacterized protein
SSPA4071	-	d		Lipoate-protein ligase A
SSPA4083	1	12		Putative two-component response regulator

Appendix E

Typhi isolates used for SNP typing

Typhi isolates were sourced from the following collections:
S - Sequenced isolates - DNA from Sanger Institute;
A - Pasteur Institute (isolates from travellers returning to France with typhoid fever) - Francois-Xavier Weill, Pasteur Institute, Paris, France;
B - Mekong Delta - Christiane Dolecek, Oxford University Clinical Research Unit, Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam;
C - Kathmandu - Andrew Pollard, Department of Pediatrics, Oxford University, Oxford, UK;
D - Kolkata - Shanta Dutta, National Institute for Cholera and Enteric Diseases, Kolkata, India
E - Kenya - Sam Kariuki, Kenya Medical Research Institute, Nairobi, Kenya.

Isolate	Year	Country	Haplotype	IncHI1	Study
B01787	2004	India	H16	-	D
A02292	2005	India	H16	-	D
D02273	2004	India	H42	-	D
D02104	2004	India	H42	-	D
C07620	2006	India	H42	-	D
C05146	2005	India	H42	-	D
C05029	2005	India	H42	-	D
C04892	2005	India	H42	-	D
C04881	2005	India	H42	-	D
C04809	2005	India	H42	-	D
C04190	2005	India	H42	-	D
C02132	2004	India	H42	-	D
C01667	2004	India	H42	-	D
C01071	2004	India	H42	-	D
C00951	2003	India	H42	-	D
B07476	2006	India	H42	-	D
B07459	2006	India	H42	-	D
B07437	2006	India	H42	-	D
B05142	2005	India	H42	-	D
B03080	2004	India	H42	-	D
B02933	2004	India	H42	-	D
B02555	2004	India	H42	-	D
B01794	2004	India	H42	-	D
B01741	2004	India	H42	-	D
B01714	2004	India	H42	-	D
B01664	2004	India	H42	-	D
A02904	2006	India	H42	-	D
A01481	2004	India	H42	-	D
A00929	2004	India	H42	-	D
D02422	2004	India	H42/H85	-	D
C01057	2004	India	H42/H85	-	D
C07566	2006	India	H42	-	D
C07548	2006	India	H42	-	D
B03159	2004	India	H42	-	D
D02348	2004	India	H50	-	D
D00763	2003	India	H50	-	D
D00205	2003	India	H50	-	D
C05572	2005	India	H50	-	D
C01662	2004	India	H50	-	D
C01606	2004	India	H50	-	D
B00679	2003	India	H50	-	D
B00181	2003	India	H50	-	D
A01910	2005	India	H50	-	D

Isolate	Year	Country	Haplotype	IncHI1	Study
D01946	2004	India	H52	-	D
D03021	2004	India	H58-A	-	D
D02274	2004	India	H58-A	-	D
D02190	2004	India	H58-A	-	D
D01988	2004	India	H58-A	-	D
D00405	2003	India	H58-A	-	D
C05685	2005	India	H58-A	ST6	D
C02633	2004	India	H58-A	ST6	D
C02066	2004	India	H58-A	-	D
C01802	2004	India	H58-A	-	D
D00193	2003	India	H58-B	-	D
C07539	2006	India	H58-B	-	D
C07343	2006	India	H58-B	-	D
C07121	2006	India	H58-B	-	D
C06953	2006	India	H58-B	-	D
C06502	2006	India	H58-B	-	D
C06427	2006	India	H58-B	-	D
C05732	2005	India	H58-B	-	D
C05573	2005	India	H58-B	-	D
C05557	2005	India	H58-B	-	D
C05518	2005	India	H58-B	-	D
C05475	2005	India	H58-B	-	D
C05446	2005	India	H58-B	-	D
C05443	2005	India	H58-B	-	D
C05440	2005	India	H58-B	-	D
C05423	2005	India	H58-B	-	D
C05279	2005	India	H58-B	-	D
C04862	2005	India	H58-B	-	D
C04529	2005	India	H58-B	-	D
C04404	2005	India	H58-B	-	D
C04401	2005	India	H58-B	-	D
C02780	2004	India	H58-B	-	D
C02670	2004	India	H58-B	-	D
C01038	2004	India	H58-B	-	D
C00145	2003	India	H58-B	-	D
B07521	2006	India	H58-B	-	D
B06864	2006	India	H58-B	-	D
B06399	2006	India	H58-B	-	D
B06375	2006	India	H58-B	-	D
B06297	2006	India	H58-B	-	D
B06295	2006	India	H58-B	-	D
B06266	2006	India	H58-B	-	D
B06198	2006	India	H58-B	-	D

Isolate	Year	Country	Haplotype	IncHI1	Study
B06078	2006	India	H58-B	-	D
B05961	2006	India	H58-B	-	D
B05787	2006	India	H58-B	-	D
B05658	2005	India	H58-B	-	D
B05600	2005	India	H58-B	-	D
B05563	2005	India	H58-B	-	D
B05561	2005	India	H58-B	-	D
B05517	2005	India	H58-B	-	D
B05505	2005	India	H58-B	-	D
B05476	2005	India	H58-B	-	D
B05467	2005	India	H58-B	-	D
B04751	2005	India	H58-B	-	D
B04716	2005	India	H58-B	-	D
B04487	2005	India	H58-B	-	D
B04421	2005	India	H58-B	-	D
B03274	2004	India	H58-B	-	D
B03166	2004	India	H58-B	-	D
B02181	2004	India	H58-B	-	D
B02150	2004	India	H58-B	-	D
B02034	2004	India	H58-B	-	D
B02026	2004	India	H58-B	-	D
B01961	2004	India	H58-B	-	D
B01911	2004	India	H58-B	-	D
B01861	2004	India	H58-B	-	D
B01813	2004	India	H58-B	-	D
B01667	2004	India	H58-B	-	D
B01661	2004	India	H58-B	-	D
B01535	2004	India	H58-B	-	D
B01342	2004	India	H58-B	-	D
B01240	2004	India	H58-B	-	D
B01074	2004	India	H58-B	-	D
B01037	2004	India	H58-B	-	D
B01021	2004	India	H58-B	-	D
B00965	2003	India	H58-B	-	D
B00756	2003	India	H58-B	-	D
B00279	2003	India	H58-B	-	D
B00116	2003	India	H58-B	-	D
B00045	2003	India	H58-B	-	D
A02964	2006	India	H58-B	-	D
A02783	2005	India	H58-B	-	D
A00981	2004	India	H58-B	-	D
A00118	2003	India	H58-B	-	D
D02950	2004	India	H58-G	-	D

Isolate	Year	Country	Haplotype	IncHI1	Study
D01545	2004	India	H58-G	-	D
D00964	2003	India	H58-G	-	D
D00043	2003	India	H58-G	-	D
D00030	2003	India	H58-G	-	D
C07610	2006	India	H58-G	-	D
C07179	2006	India	H58-G	-	D
C07087	2006	India	H58-G	-	D
C07079	2006	India	H58-G	-	D
C07052	2006	India	H58-G	-	D
C06932	2006	India	H58-G	-	D
C06876	2006	India	H58-G	-	D
C06875	2006	India	H58-G	-	D
C06855	2006	India	H58-G	-	D
C06806	2006	India	H58-G	-	D
C06567	2006	India	H58-G	-	D
C05501	2005	India	H58-G	-	D
C04932	2005	India	H58-G	ST6	D
C04903	2005	India	H58-G	ST6	D
C04365	2005	India	H58-G	-	D
C04334	2005	India	H58-G	-	D
C04062	2005	India	H58-G	-	D
C03891	2005	India	H58-G	ST6	D
C03112	2004	India	H58-G	-	D
C01818	2004	India	H58-G	-	D
C01777	2004	India	H58-G	-	D
C00777	2003	India	H58-G	-	D
B06681	2006	India	H58-G	-	D
B06390	2006	India	H58-G	-	D
B05529	2005	India	H58-G	-	D
B03103	2004	India	H58-G	-	D
B03000	2004	India	H58-G	-	D
B02377	2004	India	H58-G	-	D
B02176	2004	India	H58-G	-	D
B02095	2004	India	H58-G	-	D
B02071	2004	India	H58-G	-	D
B02020	2004	India	H58-G	-	D
B01501	2004	India	H58-G	-	D
B00031	2003	India	H58-G	ST6	D
B00025	2003	India	H58-G	-	D
A03175	2006	India	H58-G	-	D
A00832	2004	India	H58-G	-	D
A00763	2004	India	H58-G	-	D
D01604	2004	India	H64	-	D

Isolate	Year	Country	Haplotype	IncHI1	Study
C05035	2005	India	H64	-	D
C03495	2005	India	H64	-	D
A02467	2005	India	H64	-	D
A01014	2004	India	H64	-	D
A00102	2003	India	H64	-	D
B02997	2004	India	H58-K	-	D
B01772	2004	India	H8	-	D
C03646	2005	India	pre-H58	-	D
B03046	2004	India	pre-H58	-	D
C03656	2005	India	H45	-	D
NPL1871	2006	Nepal	H42	-	C
NPL882	2005	Nepal	H42	-	C
NPL726	2005	Nepal	H42	-	C
NPL1922	2006	Nepal	H42	-	C
NPL1708	2006	Nepal	H42	-	C
NPL1591	2006	Nepal	H42	-	C
NPL1421	2006	Nepal	H42	-	C
NPL1402	2006	Nepal	H42	-	C
NPL1121	2006	Nepal	H42	-	C
NPL1077	2006	Nepal	H42	-	C
NPL1265	2006	Nepal	H42	-	C
NPL728	2005	Nepal	H50	-	C
NPL1382	2006	Nepal	H50	-	C
NPL107	2005	Nepal	H50	-	C
NPL959	2006	Nepal	H58-B	-	C
NPL239	2005	Nepal	H58-G	-	C
NPL1493	2006	Nepal	H58-G	-	C
NPL972	2006	Nepal	H58-G	-	C
NPL95	2005	Nepal	H58-G	-	C
NPL830	2005	Nepal	H58-G	-	C
NPL809	2005	Nepal	H58-G	-	C
NPL764	2005	Nepal	H58-G	-	C
NPL73	2005	Nepal	H58-G	-	C
NPL716	2005	Nepal	H58-G	-	C
NPL699	2005	Nepal	H58-G	-	C
NPL64	2005	Nepal	H58-G	-	C
NPL587	2005	Nepal	H58-G	-	C
NPL537	2005	Nepal	H58-G	-	C
NPL528	2005	Nepal	H58-G	-	C
NPL527	2005	Nepal	H58-G	-	C
NPL462	2005	Nepal	H58-G	-	C
NPL453	2005	Nepal	H58-G	-	C
NPL1921	2006	Nepal	H58-G	-	C

Isolate	Year	Country	Haplotype	IncHI1	Study
NPL1838	2006	Nepal	H58-G	-	C
NPL17	2005	Nepal	H58-G	-	C
NPL1505	2006	Nepal	H58-G	-	C
NPL1487	2006	Nepal	H58-G	-	C
NPL1354	2006	Nepal	H58-G	-	C
NPL1305	2006	Nepal	H58-G	-	C
NPL1287	2006	Nepal	H58-G	-	C
NPL1255	2006	Nepal	H58-G	-	C
NPL1238	2006	Nepal	H58-G	-	C
NPL117	2005	Nepal	H58-G	-	C
NPL1143	2006	Nepal	H58-G	-	C
NPL1048	2006	Nepal	H58-G	-	C
NPL872	2005	Nepal	H58	-	C
FEB6075	2004	Kenya	H73	ST6-like	E
FEB7273	2008	Kenya	H42	-	E
Sam6	1998	Kenya	H52	-	E
LIV915	1989	Kenya	H7	-	E
LIV901	1989	Kenya	H7	-	E
LIV822	1989	Kenya	H7	-	E
LIV586	1988	Kenya	H7	-	E
LIV5223	1992	Kenya	H7	-	E
LIV516	1988	Kenya	H7	-	E
LIV279	1988	Kenya	H7	-	E
FEB6319	2005	Kenya	H7	-	E
Sam8	1998	Kenya	H55	-	E
Sam11	2001	Kenya	H55	-	E
FEB7271	2008	Kenya	H58-B	ST6	E
FEB7263	2008	Kenya	H58-B	ST6	E
FEB7231	2008	Kenya	H58-B	ST6	E
FEB7223	2008	Kenya	H58-B	-	E
FEB7212	2008	Kenya	H58-B	-	E
FEB7195	2008	Kenya	H58-B	ST6	E
FEB6747	2006	Kenya	H58-B	ST6	E
FEB6465	2006	Kenya	H58-B	ST6	E
FEB6384	2005	Kenya	H58-B	ST6	E
FEB6329	2005	Kenya	H58-B	ST6	E
FEB6323	2005	Kenya	H58-B	ST6	E
FEB6318	2005	Kenya	H58-B	ST6	E
FEB6157	2005	Kenya	H58-B	ST6	E
FEB6154	2005	Kenya	H58-B	ST6	E
FEB6124	2005	Kenya	H58-B	ST6	E
FEB6123	2005	Kenya	H58-B	ST6	E
FEB6094	2004	Kenya	H58-B	ST6	E

Isolate	Year	Country	Haplotype	IncHI1	Study
FEB6079	2004	Kenya	H58-B	ST6	E
FEB6078	2004	Kenya	H58-B	ST6	E
FEB6077	2004	Kenya	H58-B	ST6	E
FEB6489	2006	Kenya	H58-G	-	E
Sam9	2001	Kenya	H58-J1	-	E
Sam7	1998	Kenya	H58-J1	ST6	E
Sam5	1994	Kenya	H58-J1	ST6	E
Sam4	1995	Kenya	H58-J1	ST6	E
Sam3	1994	Kenya	H58-J1	ST6	E
Sam2	2005	Kenya	H58-J1	ST6	E
Sam10	2001	Kenya	H58-J1	ST6	E
Sam1	2003	Kenya	H58-J1	ST6	E
KEN980	1992	Kenya	H58-J1	ST6	E
KEN738	1991	Kenya	H58-J1	ST6	E
KEN678	1991	Kenya	H58-J1	ST6	E
KEN417	1990	Kenya	H58-J1	ST6	E
KEN297	1989	Kenya	H58-J1	ST6	E
KEN294	1989	Kenya	H58-J1	ST6	E
KEN239	1988	Kenya	H58-J1	ST6	E
FEB7108	2008	Kenya	H58-J1	-	E
FEB6466	2006	Kenya	H58-J1	ST6	E
FEB6156	2005	Kenya	H58-J1	ST6	E
FEB6125	2005	Kenya	H58-J1	ST6	E
FEB6074	2004	Kenya	H58-J1	ST6	E
FEB4106	2003	Kenya	H58-J1	ST6	E
FEB1441	2001	Kenya	H58-J1	ST6	E
FEB1408	2001	Kenya	H58-J1	ST6	E
FEB1361	2001	Kenya	H58-J1	ST6	E
FEB1303	2001	Kenya	H58-J1	ST6	E
FEB1300	2001	Kenya	H58-J1	ST6	E
FEB1294	2001	Kenya	H58-J1	ST6	E
FEB1227	2001	Kenya	H58-J2	ST6	E
FEB1226	2001	Kenya	H58-J2	ST6	E
LIV846	1989	Kenya	H16	-	E
KEN858	1991	Kenya	H16	-	E
KEN741	1991	Kenya	H16	-	E
FEB6653	2006	Kenya	H16	-	E
FEB6645	2006	Kenya	H16	-	E
FEB6784	2006	Kenya	H45	-	E
FEB6126	2005	Kenya	H45	-	E
FEB6472	2006	Kenya	H14	-	E
FEB6377	2005	Kenya	H14	-	E
FEB6118	2005	Kenya	H14	-	E

Isolate	Year	Country	Haplotype	IncHI1	Study
BJ365	2004	Vietnam	contam	contam	B
BJ382	2004	Vietnam	contam	contam	B
BJ64	2004	Vietnam	H1	-	B
BJ63	2004	Vietnam	H1	-	B
BJ105	2004	Vietnam	H1	-	B
BJ264	2004	Vietnam	H45	-	B
BJ9	2004	Vietnam	H50	-	B
BJ3	2004	Vietnam	H52	-	B
BJ359	2004	Vietnam	H58-B	ST6	B
BJ99	2005	Vietnam	H58-C	ST6	B
BJ95	2005	Vietnam	H58-C	-	B
BJ94	2005	Vietnam	H58-C	-	B
BJ91	2005	Vietnam	H58-C	ST6	B
BJ89	2005	Vietnam	H58-C	ST6	B
BJ83	2004	Vietnam	H58-C	ST6	B
BJ76	2004	Vietnam	H58-C	ST6	B
BJ75	2004	Vietnam	H58-C	ST6	B
BJ71	2004	Vietnam	H58-C	ST6	B
BJ70	2004	Vietnam	H58-C	ST6	B
BJ69	2004	Vietnam	H58-C	ST6	B
BJ67	2004	Vietnam	H58-C	-	B
BJ66	2004	Vietnam	H58-C	-	B
BJ60	2004	Vietnam	H58-C	-	B
BJ6	2004	Vietnam	H58-C	ST6	B
BJ57	2004	Vietnam	H58-C	ST6	B
BJ525	2005	Vietnam	H58-C	ST6	B
BJ524	2005	Vietnam	H58-C	ST6	B
BJ521	2005	Vietnam	H58-C	ST6	B
BJ520	2005	Vietnam	H58-C	-	B
BJ52	2005	Vietnam	H58-C	-	B
BJ518	2005	Vietnam	H58-C	-	B
BJ517	2005	Vietnam	H58-C	ST6	B
BJ515	2005	Vietnam	H58-C	ST6	B
BJ514	2005	Vietnam	H58-C	ST6	B
BJ512	2005	Vietnam	H58-C	ST6	B
BJ511	2005	Vietnam	H58-C	ST6	B
BJ510	2005	Vietnam	H58-C	ST6	B
BJ507	2005	Vietnam	H58-C	ST6	B
BJ506	2005	Vietnam	H58-C	ST6	B
BJ505	2005	Vietnam	H58-C	ST6	B
BJ504	2005	Vietnam	H58-C	ST6	B
BJ503	2005	Vietnam	H58-C	ST6	B
BJ5	2004	Vietnam	H58-C	-	B

Isolate	Year	Country	Haplotype	IncHI1	Study
BJ402	2005	Vietnam	H58-C	ST6	B
BJ401	2005	Vietnam	H58-C	ST6	B
BJ400	2005	Vietnam	H58-C	ST6	B
BJ398	2005	Vietnam	H58-C	ST6	B
BJ397	2005	Vietnam	H58-C	ST6	B
BJ396	2005	Vietnam	H58-C	ST6	B
BJ395	2005	Vietnam	H58-C	ST6	B
BJ394	2005	Vietnam	H58-C	ST6	B
BJ393	2005	Vietnam	H58-C	ST6	B
BJ392	2005	Vietnam	H58-C	-	B
BJ391	2005	Vietnam	H58-C	ST6	B
BJ388	2005	Vietnam	H58-C	ST6	B
BJ387	2005	Vietnam	H58-C	ST6	B
BJ386	2005	Vietnam	H58-C	ST6	B
BJ384	2005	Vietnam	H58-C	-	B
BJ380	2005	Vietnam	H58-C	ST6	B
BJ379	2005	Vietnam	H58-C	ST6	B
BJ378	2005	Vietnam	H58-C	ST6	B
BJ377	2005	Vietnam	H58-C	-	B
BJ375	2005	Vietnam	H58-C	ST6	B
BJ373	2005	Vietnam	H58-C	ST6	B
BJ372	2005	Vietnam	H58-C	ST6	B
BJ370	2005	Vietnam	H58-C	ST6	B
BJ367	2005	Vietnam	H58-C	ST6	B
BJ366	2005	Vietnam	H58-C	ST6	B
BJ364	2005	Vietnam	H58-C	ST6	B
BJ362	2005	Vietnam	H58-C	ST6	B
BJ361	2005	Vietnam	H58-C	ST6	B
BJ360	2004	Vietnam	H58-C	ST6	B
BJ358	2005	Vietnam	H58-C	ST6	B
BJ356	2005	Vietnam	H58-C	ST6	B
BJ353	2005	Vietnam	H58-C	ST6	B
BJ351	2004	Vietnam	H58-C	ST6	B
BJ345	2004	Vietnam	H58-C	-	B
BJ336	2004	Vietnam	H58-C	ST6	B
BJ324	2004	Vietnam	H58-C	ST6	B
BJ322	2004	Vietnam	H58-C	-	B
BJ321	2004	Vietnam	H58-C	ST6	B
BJ318	2004	Vietnam	H58-C	ST6	B
BJ315	2004	Vietnam	H58-C	ST6	B
BJ311	2004	Vietnam	H58-C	ST6	B
BJ297	2004	Vietnam	H58-C	ST6	B
BJ296	2004	Vietnam	H58-C	ST6	B

Isolate	Year	Country	Haplotype	IncHI1	Study
BJ290	2004	Vietnam	H58-C	ST6	B
BJ288	2004	Vietnam	H58-C	ST6	B
BJ287	2004	Vietnam	H58-C	ST6	B
BJ286	2004	Vietnam	H58-C	ST6	B
BJ283	2004	Vietnam	H58-C	ST6	B
BJ279	2004	Vietnam	H58-C	ST6	B
BJ275	2004	Vietnam	H58-C	-	B
BJ271	2004	Vietnam	H58-C	-	B
BJ260	2004	Vietnam	H58-C	-	B
BJ26	2004	Vietnam	H58-C	-	B
BJ258	2004	Vietnam	H58-C	-	B
BJ256	2004	Vietnam	H58-C	-	B
BJ251	2004	Vietnam	H58-C	-	B
BJ249	2004	Vietnam	H58-C	ST6	B
BJ248	2004	Vietnam	H58-C	ST6	B
BJ240	2004	Vietnam	H58-C	ST6	B
BJ230	2004	Vietnam	H58-C	ST6	B
BJ223	2004	Vietnam	H58-C	ST6	B
BJ220	2004	Vietnam	H58-C	ST6	B
BJ217	2004	Vietnam	H58-C	ST6	B
BJ2	2004	Vietnam	H58-C	ST6	B
BJ196	2004	Vietnam	H58-C	ST6	B
BJ193	2004	Vietnam	H58-C	-	B
BJ192	2004	Vietnam	H58-C	ST6	B
BJ190	2004	Vietnam	H58-C	ST6	B
BJ187	2004	Vietnam	H58-C	ST6	B
BJ185	2004	Vietnam	H58-C	ST6	B
BJ174	2004	Vietnam	H58-C	ST6	B
BJ168	2004	Vietnam	H58-C	-	B
BJ164	2004	Vietnam	H58-C	ST6	B
BJ163	2004	Vietnam	H58-C	ST6	B
BJ162	2004	Vietnam	H58-C	ST6	B
BJ161	2004	Vietnam	H58-C	-	B
BJ160	2004	Vietnam	H58-C	ST6	B
BJ159	2004	Vietnam	H58-C	-	B
BJ154	2004	Vietnam	H58-C	ST6	B
BJ151	2005	Vietnam	H58-C	ST6	B
BJ104	2005	Vietnam	H58-C	ST6	B
BJ1	2004	Vietnam	H58-C	ST6	B
BJ7	2004	Vietnam	H58-C	-	B
BJ88	2005	Vietnam	H58-D3	ST6	B
BJ516	2005	Vietnam	H58-D3	ST6	B
BJ502	2005	Vietnam	H58-D3	-	B

Isolate	Year	Country	Haplotype	IncHI1	Study
BJ228	2004	Vietnam	H58-D3	ST6	B
BJ59	2004	Vietnam	H58-E1	ST6	B
BJ509	2005	Vietnam	H58-E1	ST6	B
BJ508	2005	Vietnam	H58-E1	ST6	B
BJ4	2004	Vietnam	H58-E1	-	B
BJ399	2005	Vietnam	H58-E1	ST6	B
BJ383	2005	Vietnam	H58-E1	ST6	B
BJ376	2005	Vietnam	H58-E1	ST6	B
BJ374	2005	Vietnam	H58-E1	ST6	B
BJ368	2005	Vietnam	H58-E1	ST6	B
BJ298	2004	Vietnam	H58-E1	ST6	B
BJ201	2004	Vietnam	H58-E1	ST6	B
BJ191	2004	Vietnam	H58-E1	ST6	B
BJ188	2004	Vietnam	H58-E1	ST6	B
BJ165	2004	Vietnam	H58-E1	ST6	B
BJ102	2005	Vietnam	H58-E1	ST6	B
BJ90	2005	Vietnam	H58-E2	ST6	B
BJ8	2004	Vietnam	H58-E2	-	B
BJ519	2005	Vietnam	H58-E2	ST6	B
BJ357	2005	Vietnam	H58-E2	-	B
BJ355	2005	Vietnam	H58-E2	-	B
BJ352	2004	Vietnam	H58-E2	-	B
BJ350	2004	Vietnam	H58-E2	-	B
BJ349	2004	Vietnam	H58-E2	-	B
BJ348	2004	Vietnam	H58-E2	-	B
BJ347	2004	Vietnam	H58-E2	-	B
BJ346	2004	Vietnam	H58-E2	-	B
BJ344	2004	Vietnam	H58-E2	-	B
BJ343	2004	Vietnam	H58-E2	-	B
BJ340	2004	Vietnam	H58-E2	-	B
BJ339	2004	Vietnam	H58-E2	-	B
BJ338	2004	Vietnam	H58-E2	-	B
BJ334	2004	Vietnam	H58-E2	-	B
BJ333	2004	Vietnam	H58-E2	-	B
BJ332	2004	Vietnam	H58-E2	-	B
BJ331	2004	Vietnam	H58-E2	-	B
BJ330	2004	Vietnam	H58-E2	-	B
BJ329	2004	Vietnam	H58-E2	-	B
BJ328	2004	Vietnam	H58-E2	-	B
BJ327	2004	Vietnam	H58-E2	-	B
BJ326	2004	Vietnam	H58-E2	-	B
BJ325	2004	Vietnam	H58-E2	-	B
BJ323	2004	Vietnam	H58-E2	ST6	B

Isolate	Year	Country	Haplotype	IncHI1	Study
BJ320	2004	Vietnam	H58-E2	-	B
BJ319	2004	Vietnam	H58-E2	-	B
BJ314	2004	Vietnam	H58-E2	-	B
BJ313	2004	Vietnam	H58-E2	-	B
BJ310	2004	Vietnam	H58-E2	-	B
BJ309	2004	Vietnam	H58-E2	-	B
BJ308	2004	Vietnam	H58-E2	-	B
BJ307	2004	Vietnam	H58-E2	-	B
BJ305	2004	Vietnam	H58-E2	-	B
BJ304	2004	Vietnam	H58-E2	-	B
BJ302	2004	Vietnam	H58-E2	-	B
BJ301	2004	Vietnam	H58-E2	-	B
BJ299	2004	Vietnam	H58-E2	-	B
BJ295	2004	Vietnam	H58-E2	-	B
BJ294	2004	Vietnam	H58-E2	-	B
BJ293	2004	Vietnam	H58-E2	-	B
BJ292	2004	Vietnam	H58-E2	ST6	B
BJ289	2004	Vietnam	H58-E2	ST6	B
BJ285	2004	Vietnam	H58-E2	-	B
BJ284	2004	Vietnam	H58-E2	-	B
BJ282	2004	Vietnam	H58-E2	-	B
BJ281	2004	Vietnam	H58-E2	-	B
BJ280	2004	Vietnam	H58-E2	-	B
BJ278	2004	Vietnam	H58-E2	-	B
BJ276	2004	Vietnam	H58-E2	-	B
BJ274	2004	Vietnam	H58-E2	-	B
BJ273	2004	Vietnam	H58-E2	-	B
BJ269	2004	Vietnam	H58-E2	-	B
BJ267	2004	Vietnam	H58-E2	-	B
BJ266	2004	Vietnam	H58-E2	-	B
BJ265	2004	Vietnam	H58-E2	-	B
BJ263	2004	Vietnam	H58-E2	-	B
BJ262	2004	Vietnam	H58-E2	-	B
BJ261	2004	Vietnam	H58-E2	-	B
BJ259	2004	Vietnam	H58-E2	-	B
BJ257	2004	Vietnam	H58-E2	-	B
BJ255	2004	Vietnam	H58-E2	-	B
BJ254	2004	Vietnam	H58-E2	-	B
BJ252	2004	Vietnam	H58-E2	-	B
BJ247	2004	Vietnam	H58-E2	-	B
BJ245	2004	Vietnam	H58-E2	-	B
BJ244	2004	Vietnam	H58-E2	-	B
BJ243	2004	Vietnam	H58-E2	-	B

Isolate	Year	Country	Haplotype	IncHI1	Study
BJ239	2004	Vietnam	H58-E2	-	B
BJ238	2004	Vietnam	H58-E2	-	B
BJ234	2004	Vietnam	H58-E2	-	B
BJ233	2004	Vietnam	H58-E2	ST6	B
BJ216	2004	Vietnam	H58-E2	-	B
BJ215	2004	Vietnam	H58-E2	-	B
BJ212	2004	Vietnam	H58-E2	ST6	B
BJ206	2004	Vietnam	H58-E2	-	B
BJ205	2004	Vietnam	H58-E2	ST6	B
BJ200	2004	Vietnam	H58-E2	-	B
BJ198	2004	Vietnam	H58-E2	-	B
BJ197	2004	Vietnam	H58-E2	-	B
BJ195	2004	Vietnam	H58-E2	-	B
BJ194	2004	Vietnam	H58-E2	ST6	B
BJ189	2004	Vietnam	H58-E2	-	B
BJ186	2004	Vietnam	H58-E2	-	B
BJ184	2004	Vietnam	H58-E2	ST6	B
BJ183	2004	Vietnam	H58-E2	ST6	B
BJ182	2004	Vietnam	H58-E2	-	B
BJ180	2004	Vietnam	H58-E2	-	B
BJ179	2004	Vietnam	H58-E2	ST6	B
BJ178	2004	Vietnam	H58-E2	-	B
BJ177	2004	Vietnam	H58-E2	-	B
BJ176	2004	Vietnam	H58-E2	-	B
BJ175	2004	Vietnam	H58-E2	-	B
BJ173	2004	Vietnam	H58-E2	-	B
BJ172	2004	Vietnam	H58-E2	-	B
BJ171	2004	Vietnam	H58-E2	ST6	B
BJ170	2004	Vietnam	H58-E2	-	B
BJ169	2004	Vietnam	H58-E2	-	B
BJ166	2004	Vietnam	H58-E2	ST6	B
BJ158	2004	Vietnam	H58-E2	-	B
BJ157	2004	Vietnam	H58-E2	-	B
BJ156	2004	Vietnam	H58-E2	ST6	B
BJ155	2004	Vietnam	H58-E2	-	B
BJ153	2004	Vietnam	H58-E2	-	B
BJ152	2004	Vietnam	H58-E2	ST6	B
BJ523	2005	Vietnam	H58-F2	ST6	B
BJ389	2005	Vietnam	H58-F2	ST6	B
BJ385	2005	Vietnam	H58-F2	ST6	B
BJ381	2005	Vietnam	H58-F2	ST6	B
BJ218	2004	Vietnam	H58-F2	-	B
BJ87	2005	Vietnam	H58-F3	ST6	B

Isolate	Year	Country	Haplotype	IncHI1	Study
BJ522	2005	Vietnam	H58-F3	ST6	B
BJ312	2004	Vietnam	H58-F3	ST6	B
BJ246	2004	Vietnam	H58-F3	ST6	B
BJ335	2004	Vietnam	H58	-	B
BJ291	2004	Vietnam	H58	ST6	B
BJ167	2004	Vietnam	H58	ST6	B
1(04)C	2004	Vietnam	H1	-	A
69-67	1967	Vietnam	H1	-	A
67-67	1967	Vietnam	H1	-	A
66-67	1967	Vietnam	H1	-	A
72-1258	1972	Mexico	H11	ST3	A
49-65	1965	Madagascar	H15	-	A
48-65	1965	Madagascar	H15	-	A
12-66	1966	Madagascar	H15	-	A
E97-3246	1997	Madagascar	H17	-	A
E97-9141	1997	Turkey	H18	-	A
E98-2107	1998	Senegal	H19	-	A
E98-6926	1998	Mauritania	H21	-	A
E98-8119	1998	Peru	H22	-	A
E98-8120	1998	Cameroon	H23	-	A
E99-1028	1999	Senegal	H24	-	A
E99-4879	1999	Morocco	H25	-	A
E99-5920	1999	Tunisia	H26	-	A
E99-6359	1999	Mali	H27	-	A
E99-6478	1999	Guinea	H28	-	A
E02-5919	2002	China	H28	-	A
E99-6646	1999	Algeria	H29	-	A
E99-6785	1999	Morocco	H30	-	A
E99-7012	1999	Morocco	H31	-	A
E99-8013	1999	Morocco	H32	-	A
E99-8095	1999	Algeria	H33	-	A
E99-9794	1999	Comoros	H35	-	A
E00-6924	2000	Algeria	H36	-	A
31-66	1966	Algeria	H36	-	A
E00-2756	2000	India	H37	-	A
E00-3201	2000	Mali	H38	-	A
E01-7923	2001	Ivory Coast	H39	-	A
E01-7101	2001	Togo	H39	-	A
134-67	1967	Senegal	H39	-	A
131-67	1967	Ivory Coast	H39	-	A
106-67	1967	Ivory Coast	H39	-	A
102-66	1966	Senegal	H39	-	A
39-67	1967	Ivory Coast	H39	-	A

Isolate	Year	Country	Haplotype	IncHI1	Study
E02-0530	2002	Nigeria	H4	-	A
E00-5869	2000	Bangladesh	H40	-	A
E00-6599	2000	CapeVerde	H41	-	A
76-1406	1976	Indonesia	H42	ST2	A
04 6845	2004	Benin	H42	ST2	A
03-4747	2003	Togo	H42	ST2	A
75-67	1967	Morocco	H42	-	A
50-67	1967	Congo	H42	-	A
31-67	1967	Congo	H42	-	A
13-62	1962	Senegal	H42	-	A
E00-6999	2000	Peru	H43	-	A
E00-7463	2000	Morocco	H44	-	A
E03-0658	2003	Philippines	H45	-	A
E00-9821	2000	Congo	H46	-	A
133-67	1967	Congo	H46	-	A
129-66	1966	Congo	H46	-	A
12-58	1958	Cameroon	H46	-	A
E01-1747	2001	Cameroon	H47	-	A
E99-8067	1999	Algeria	H48	-	A
E01-1811	2001	Mali	H49	-	A
E98-4364	1998	Mexico	H50	-	A
E03-6643	2003	India	H50	-	A
80-2002	1980	Madagascar	H50	-	A
76-1261	1976	Zaire	H50	-	A
81-863	1981	Peru	H50	ST8	A
73-114	1973	Vietnam	H50	-	A
162-66	1966	Algeria	H50	-	A
104-67	1967	Ivory Coast	H50	-	A
76-54	1976	Chile	H50	central	A
73-99	1973	Vietnam	H50	-	A
68-63	1963	Chad	H50	-	A
67-63	1963	Chad	H50	-	A
66-61	1961	Tunisia	H50	-	A
49-67	1967	Vietnam	H50	-	A
49-66	1966	Algeria	H50	-	A
40-67	1967	Madagascar	H50	-	A
37-66	1966	Cameroon	H50	-	A
32-66	1966	Cameroon	H50	-	A
28-62	1962	Chad	H50	-	A
06-62	1962	Senegal	H50	-	A
05-59	1959	Vietnam	H50	-	A
CIS9661/06	unk	unk	H50	-	A
E01-7006	2001	Lebanon	H51	-	A

Isolate	Year	Country	Haplotype	IncHI1	Study
E00-6172	2000	Indonesia	H52	-	A
171-66	1966	Morocco	H52	-	A
84-66	1966	Tunisia	H52	-	A
73-43	1973	France	H52	-	A
68-61	1961	Tunisia	H52	-	A
64-63	1963	Chad	H52	-	A
63-63	1963	Chad	H52	-	A
41-63	1963	Chad	H52	-	A
29-66	1966	Algeria	H52	-	A
27-67	1967	Senegal	H52	-	A
27-64	1964	Chad	H52	-	A
19-66	1966	Congo	H52	-	A
12-62	1962	Senegal	H52	-	A
10-64	1964	Chad	H52	-	A
08-64	1964	Chad	H52	-	A
07-62	1962	Senegal	H52	-	A
06-64	1964	Chad	H52	-	A
05-67	1967	Congo	H52	-	A
CIS9662/06	unk	unk	H52	-	A
E01-8716	2001	Srilanka	H53	-	A
E02-0232	2002	French Guiana	H54	-	A
75-2507	1975	India	H55	ST2	A
77-303	1977	India	H55	ST2	A
77-302	1977	India	H55	ST2	A
69-61	1961	Tunisia	H56	-	A
E99-8635	1999	Nepal	H58-A	-	A
E03-6418	2003	Bangladesh	H58-A	-	A
19(02)S	2002	Vietnam	H58-B	ST6	A
14/96	1996	Vietnam	H58-C	ST6	A
4(02)N	2002	Vietnam	H58-C	ST6	A
318(98)N	1998	Vietnam	H58-C	-	A
209(97)S	1997	Vietnam	H58-C	ST6	A
43(97)S	1997	Vietnam	H58-C	ST6	A
E03-5712	2003	Cambodia	H58-C	ST6	A
8(04)S	2004	Vietnam	H58-C	-	A
8(02)S	2002	Vietnam	H58-C	-	A
8(02)C	2002	Vietnam	H58-C	ST6	A
7(02)N	2002	Vietnam	H58-C	-	A
49(98)S	1998	Vietnam	H58-C	ST6	A
43(98)S	1998	Vietnam	H58-C	ST6	A
4(04)C	2004	Vietnam	H58-C	-	A
4(02)S	2002	Vietnam	H58-C	-	A
4(02)C	2002	Vietnam	H58-C	ST6	A

Isolate	Year	Country	Haplotype	IncHI1	Study
39(98)S	1998	Vietnam	H58-C	-	A
38(98)S	1998	Vietnam	H58-C	ST6	A
30(98)S	1998	Vietnam	H58-C	-	A
3(04)C	2004	Vietnam	H58-C	ST6	A
3(02)C	2002	Vietnam	H58-C	-	A
219(99)S	1999	Vietnam	H58-C	ST6	A
21(04)S	2004	Vietnam	H58-C	-	A
205(97)S	1997	Vietnam	H58-C	ST6	A
2(02)S	2002	Vietnam	H58-C	-	A
2(02)N	2002	Vietnam	H58-C	ST6	A
20(02)N	2002	Vietnam	H58-C	-	A
17(02)S	2002	Vietnam	H58-C	ST6	A
16(04)S	2004	Vietnam	H58-C	ST6	A
12(02)S	2002	Vietnam	H58-C	-	A
11(02)S	2002	Vietnam	H58-C	ST6	A
1(02)C	2002	Vietnam	H58-C	-	A
358/98	1998	Vietnam	H58-C	ST6	A
339/98	1998	Vietnam	H58-C	ST6	A
192(99)S	1999	Vietnam	H58-D2	-	A
2(04)S	2004	Vietnam	H58-D3	-	A
197(99)S	1999	Vietnam	H58-D3	-	A
43-64	1964	Chad	H58-E2	-	A
E00-9345	2000	India	H58-G	-	A
E02-2159	2002	Srilanka	H58-G	ST6	A
14(02)S	2002	Vietnam	H60	-	A
226(97)S	1997	Vietnam	H61	ST6	A
04-2176	2004	India	H58-I1	-	A
E00-6111	2000	India	H58-I2	-	A
E02-1963	2002	Laos	H58	-	A
31(98)S	1998	Vietnam	H58	ST6	A
76-1292	1976	Zaire	H6	-	A
05-3275	2005	Morocco	H6	-	A
E01-5741	2001	Angola	H6	-	A
SonLa-1	2002	Vietnam	H68	-	A
72-1907	1972	Vietnam	H68	ST2	A
27-58	1958	Morocco	H69	-	A
162/95	1995	Vietnam	H75	-	A
81-918	1981	Peru	H77	ST8	A
81-424	1981	Peru	H77	ST8	A
14-58	1958	Cameroon	H77	-	A
E02-1687	2002	Thailand	H79	-	A
E01-5612	2001	Indonesia	H8	-	A
E00-4624	2000	China	H8	-	A