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 ∼4,500 genes (46, 47, 48, 49, 50, 51) and are genetically very similar, sharing ∼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 ∼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).

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

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.

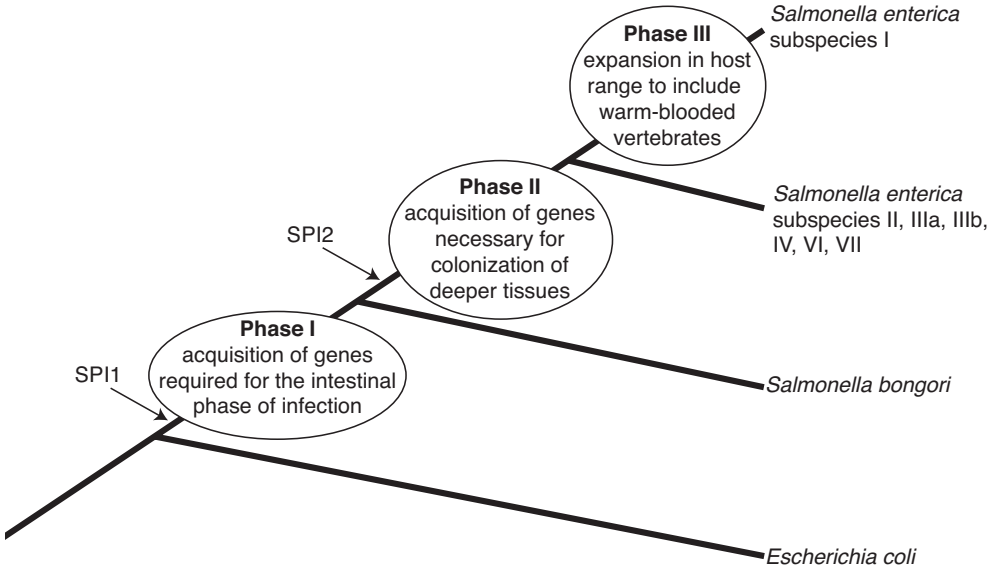

Figure 1.1: Model for the evolution of virulence in the genus *Salmonella* - Reproduced from (53), with SPI1 and SPI2 insertion points added. Original legend: "The three phases in which virulence evolved in the genus *Salmonella* since its divergence from the *E. coli* lineage have been proposed previously (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).

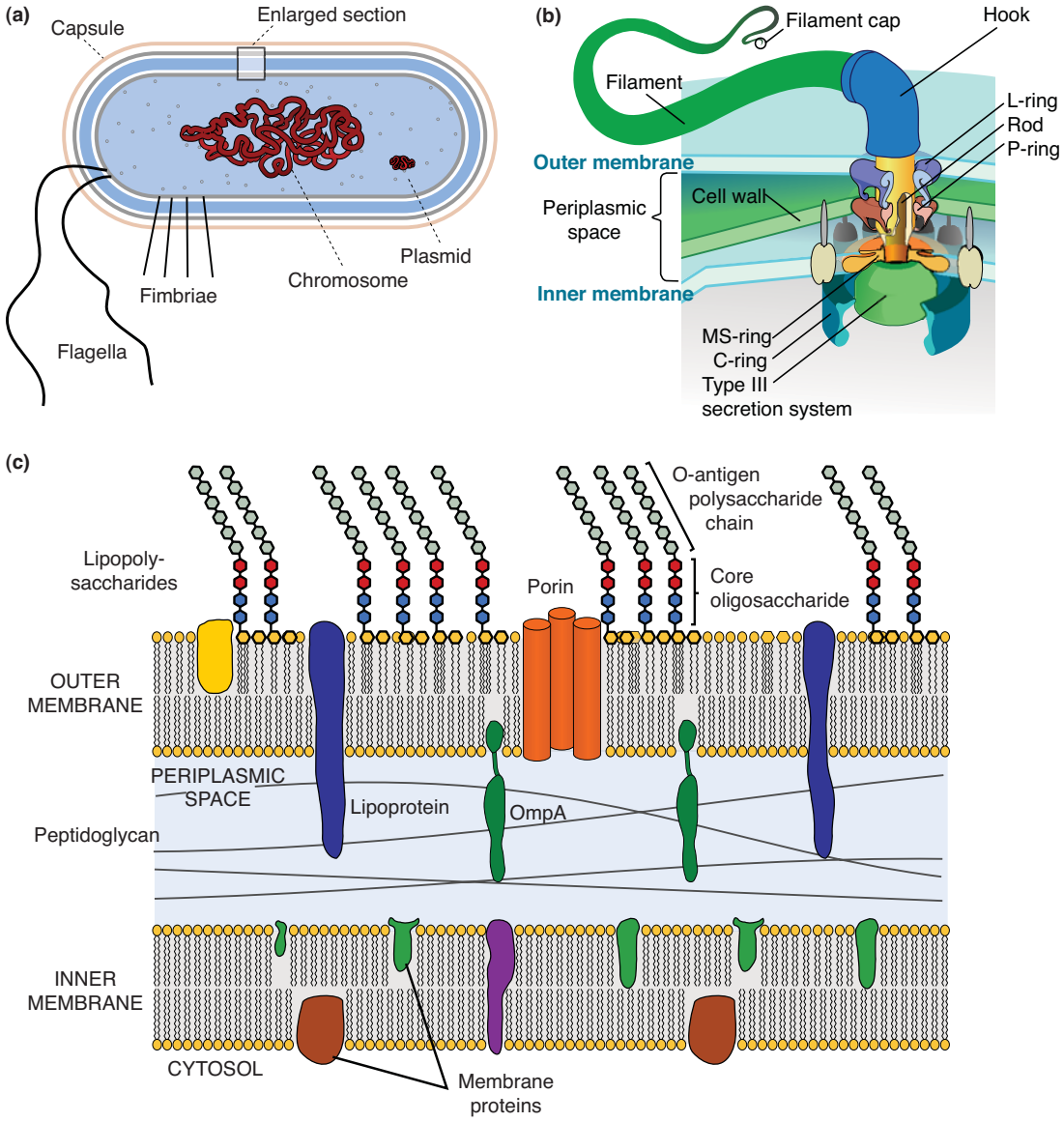

Figure 1.2: Structure of a *Salmonella* cell, flagellum and cell wall. - (a) Structure of a cell, section is enlarged in (c). (b) Structure of a flagellum. (c) Structure of the *Salmonella* cell wall. Reproduced from drawings by Jeff Dahl (a,c) and Mariana Ruiz Villarreal (b).

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

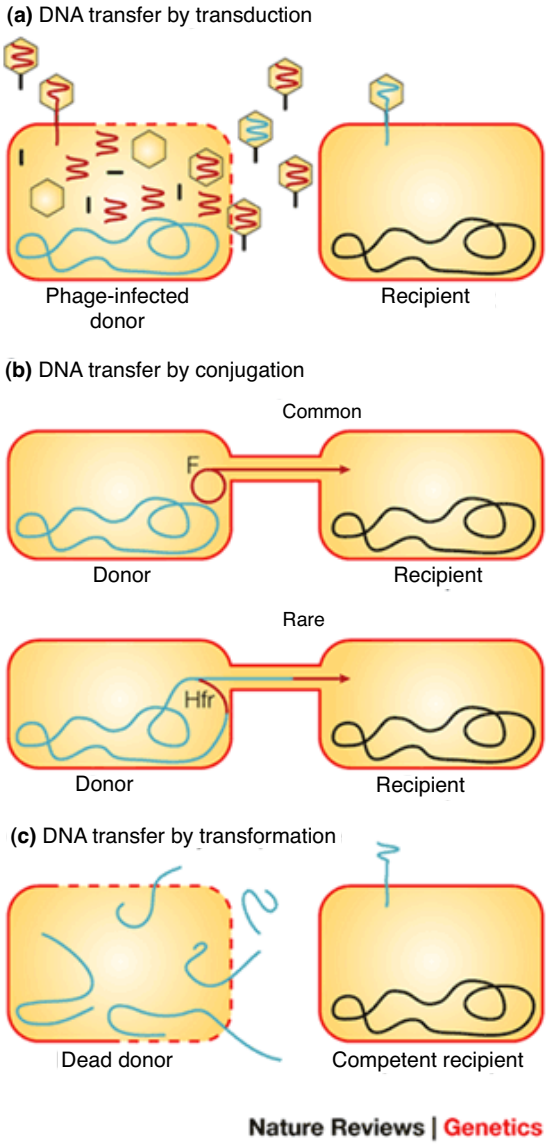

Figure 1.3: Methods of DNA transfer - Reproduced from (165). Original legend: "(a) Transduction is the phage-mediated transfer of host genetic information. In a phage-infected bacterial cell, fragments of the host DNA are occasionally packaged into phage particles and can then be transferred to a recipient 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

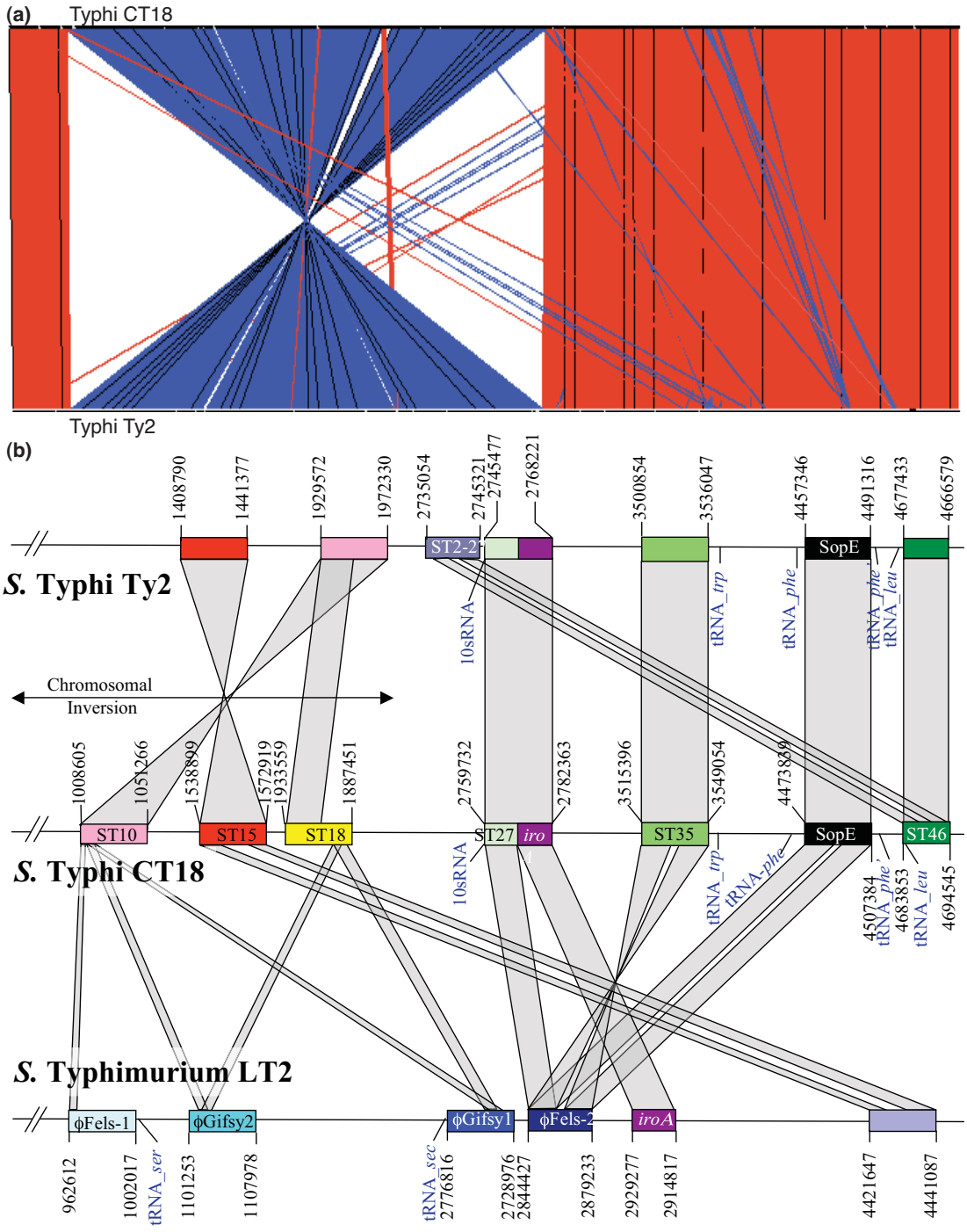

Figure 1.4: Genome rearrangements and phage differences between Typhi CT18 and Ty2 - (a) Linear comparison of Typhi CT18 and Ty2. (b) Prophage in Typhi and Typhimurium genome sequences. Reproduced from (224), original legend: "Illustration of the relative alignments of the prophage regions within the 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).

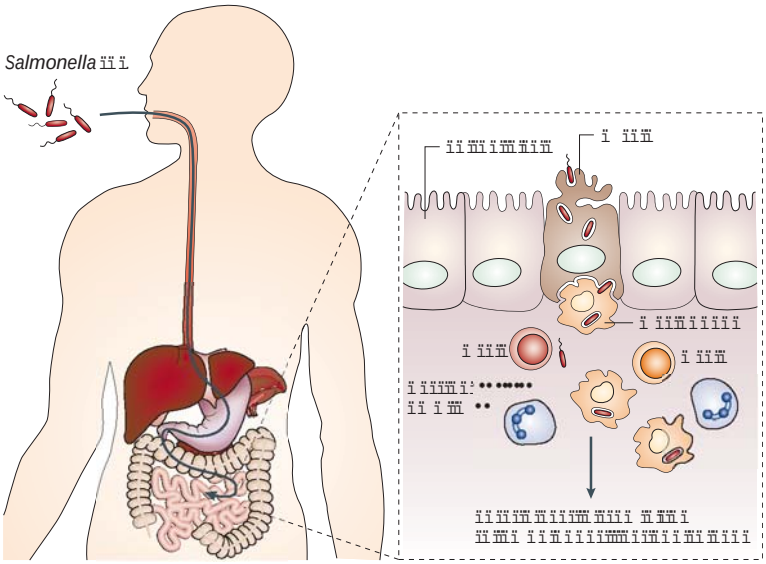

Figure 1.5: Biology of *Salmonella* infection - Reproduced from (296). Original legend: "Orally ingested salmonellae survive at the low pH of the stomach and evade the multiple defences of the small intestine in order to gain access to the epithelium. Salmonellae preferentially enter M cells, which transport 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

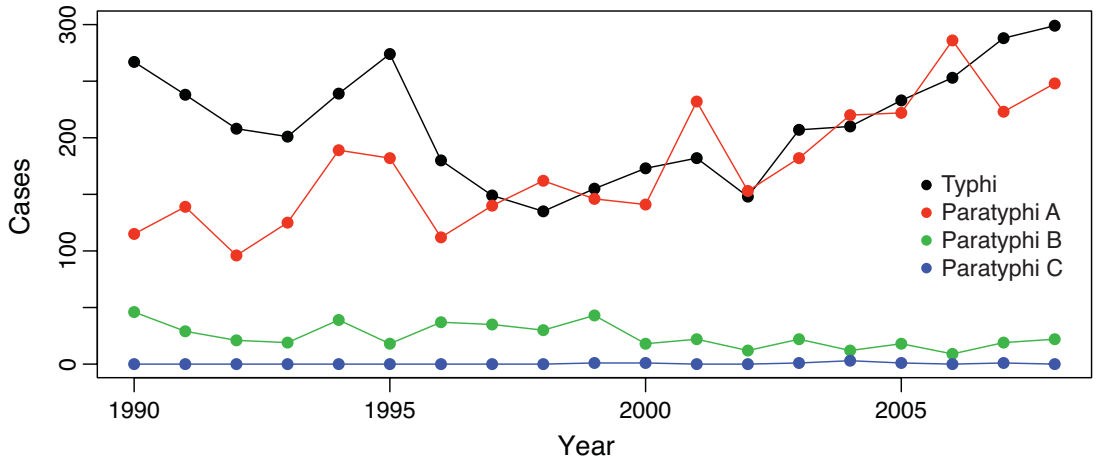

Figure 1.6: Trends in enteric fever incidence in the UK, 1990-2008 - Annual enteric fever cases per year for England, Wales and Northern Ireland, split by serotype. Sourced from publicly available data published online by the Health Protection Agency, London, UK (330).

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).

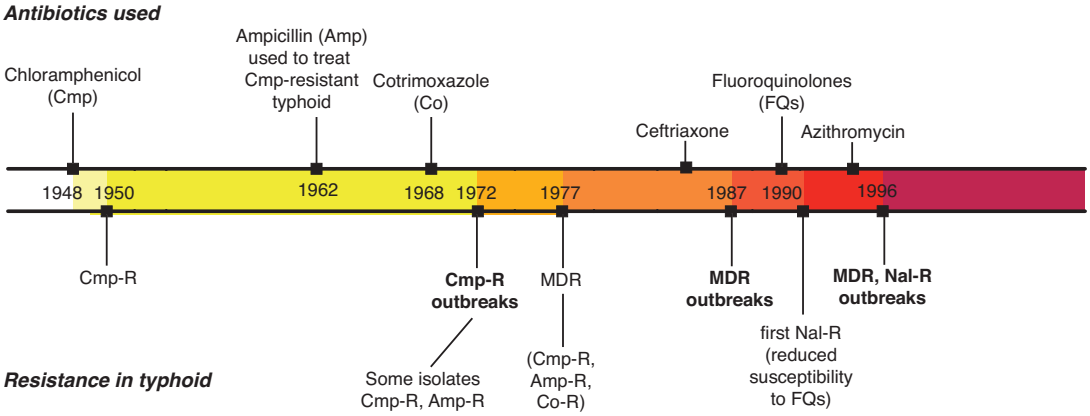

Figure 1.7: Timeline of the use of, and development of resistance to, antibiotics in enteric fever - R = resistant; MDR = multiple drug resistance; Nal = nalidixic acid, the prototype quinolone antibiotic. Details and citations for all events are given in the text.

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 al-

most 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

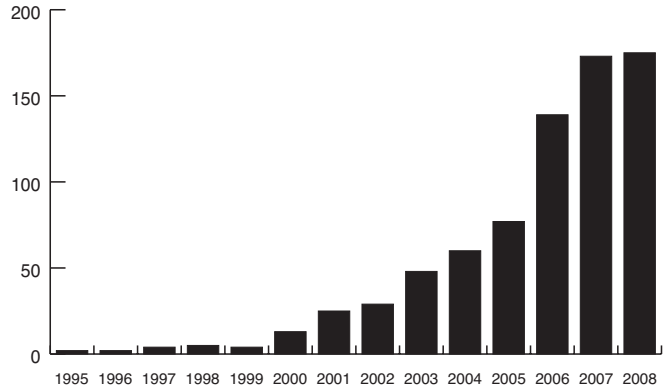

Figure 1.8: Complete bacterial genome sequences deposited in public databases - Data sourced from Genomes OnLine Database (488), July 2009. Note the data shown is the number of genome sequences deposited during each year, not the cumulative number of genome sequences in the databases.

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-

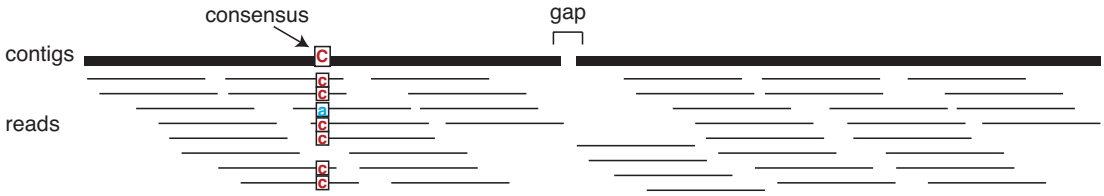

Figure 1.9: Sequence assembly - Reads are assembled into contigs (contiguous sequences) based on overlapping sequences. Gaps between contigs can be closed using additional PCR and sequencing experiments.

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

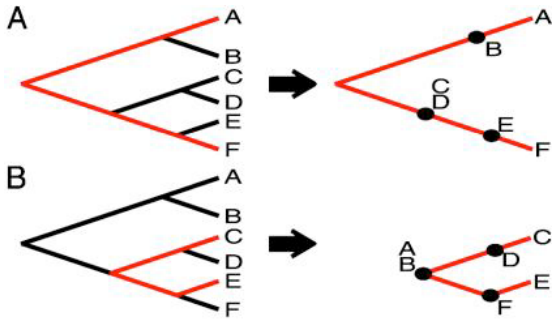

Figure 1.10: Phylogenetic discovery bias - Reproduced from (520). Original legend: "Evolutionary model showing the consequences of biased character discovery for nonhomoplastic molecular markers. (Left) The 'true' path structure of OTUs AF is shown. (A) When OTUs A and F are used for comparative character (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 be-

tween 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.