1 Introduction

Some bacterial pathogens are defined by the host organism they infect. This thesis describes the infectious capacity of such organisms, the human-restricted *Salmonella enterica* serovars Typhi and Paratyphi A, and the chicken-restricted serovar Gallinarum. How this capacity can be related to loss of gene function and bacterial metabolism is demonstrated by comparison with the non-host-adapted serovar Typhimurium.

1.1 Salmonella

1.1.1 Classification

1.1.1.1 The genus Salmonella

The Gram negative genus *Salmonella* is currently divided into two species, *S. bongori* and *S. enterica*. A third species, *S. subterranea* was proposed in 2004 (Shelobolina et al. 2004), but this was later shown not to belong to the genus (Grimont and Weill 2007). *S. enterica* is further divided into six subspecies, *enterica*, *salamae*, *arizonae*, *diarizonae*, *houtenae* and *indica*, based upon DNA-DNA hybridisation, 16S RNA analysis and multi-locus enzyme electrophoresis (Crosa et al. 1973; Reeves et al. 1989). The vast majority of *S. enterica* serovars are found in subspecies *enterica* and account for >99.9% of known human and animal infection (Figure 1-1) (Selander et al. 1996).

A member of the Enterobacteriaceae, *Salmonella* is most closely related to *Shigella*, *Escherichia coli* and *Citrobacter* (Figure 1-2).

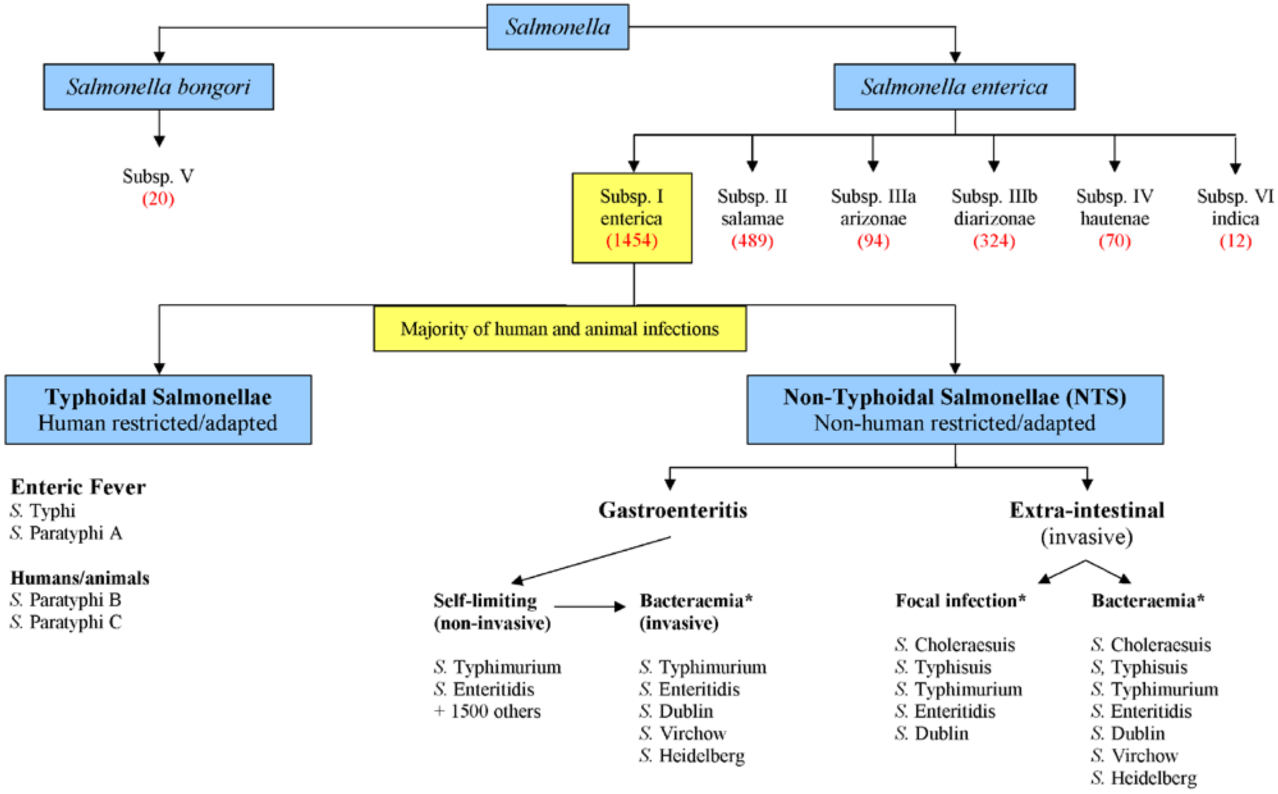


Figure 1-1 Classification of the genus *Salmonella*

Figure adapted from (Langridge et al. 2008). *Salmonella* subspecies have been defined by biotyping, DNA-DNA hybridisation, 16S RNA analysis and multi-locus enzyme electrophoresis. Serotyping is used for differentiation beyond the level of subspecies. Serovar numbers are given in red below each subspecies (Guibourdenche et al. 2009). * Common serotypes are listed but other serotypes may cause bacteraemia or focal infection; subsp., subspecies.

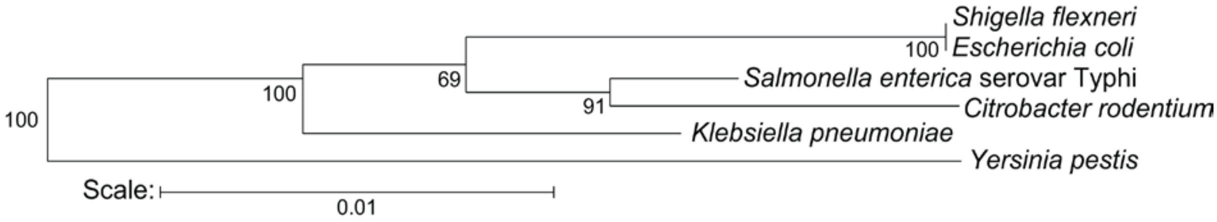


Figure 1-2 Relationships between the Enterobacteriaceae

Weighted neighbour joining tree based upon 16S rRNA sequences retrieved from the Ribosomal Database Project and built using Tree Builder (Cole et al. 2009). *S. enterica* serovar Typhi used as a representative of *S. enterica*. Numbers indicate bootstrap values for each branch and distance is based upon the Jukes-Cantor correction.

1.1.1.2 Serotyping and biochemical testing

As a genus, *Salmonella* is subdivided serologically into 2,610 serovars by the White-Kauffmann-Le Minor scheme (Grimont and Weill 2007; Guibourdenche et al. 2009). This scheme is based upon ‘O’ surface antigens and the expression of flagellar ‘H’ antigens. The polysaccharide O antigen is encoded by the *wba* (*rfb*) gene cluster and forms part of the lipopolysaccharide (LPS) found in the outer membrane of Gram negative bacteria. It consists of a variable number of oligosaccharide repeats, and sixty seven O antigen variants are currently known (Grimont and Weill 2007). The H antigens relate to flagellar phases 1 (motile) and 2 (non-motile) and are the respective products of *fliC* and *fljB*. Together, these antigens give rise to an antigenic formula that (in most cases) differentiates each serovar. This formula is written in the style O antigen(s): H antigen phase 1: H antigen phase 2: other, where other refers to rare ‘R’ or third H antigen phases (Grimont and Weill 2007). The Vi capsular antigen, found in few serovars, is given in square brackets after the O antigen designation. Some serovars have identical antigenic formulas but are distinguished by biochemical properties, pathogenicity or environmental niche.

Salmonella strains are serotyped by slide or tube agglutination with antisera. One H antigen is expressed predominantly; strains can be grown in soft agar containing specific antibody against the expressed antigen to induce a switch to the other H antigen (if present). Agglutination results can be difficult to interpret, leading to difficulties in standardisation even between reference laboratories. In a clinical setting, prior to serotyping, suspected Enterobacteriaceae are often subjected to biochemical testing. Methods such as the API20E (BioMérieux, France) test the ability of an isolate to grow upon certain substrates, produce various metabolic products or alter the pH. Again, results of such tests are open to interpretation and require experience and a high level of technical skill to generate reproducible results.

1.1.1.3 Multi Locus Enzyme Electrophoresis and Sequence Typing

Molecular methods have also been used to distinguish *Salmonella* serovars. Multi locus enzyme electrophoresis (MLEE) distinguishes strains based upon the electrophoretic mobilities of various intracellular enzymes (Selander and Levin 1980; Whittam et al. 1983). Mutations in the genes encoding these enzymes may result in amino acid substitutions that affect the electrostatic charge of the enzyme and hence its electrophoretic mobility. However, the gel-based nature of this typing scheme means that this too is subject to interpretation and difficult to standardise across laboratories.

With the advent of fast, low-cost nucleotide sequencing, Maiden and colleagues published their multi locus sequence typing (MLST) scheme in *Neisseria meningitidis*, which directly captured nucleotide sequence variation in 6 housekeeping genes (i.e. genes under stabilising selection) (Maiden et al. 1998). This scheme was subsequently adapted to *Salmonella* using seven housekeeping genes (Kidgell et al. 2002). Defined fragments from each gene are PCR-amplified and sequenced to determine any nucleotide differences. Any base difference in the sequence is considered a new allele and is given a new number. The numbers from each of the seven alleles are put together to form a 'barcode' and each new barcode is designated a sequence type (ST). The primers used for both the PCR amplification and the sequencing are published as standards so results are reproducible between laboratories. A database for *Salmonella enterica* was set up in 2005 (http://pubmlst.org/databases.shtml) and currently has over 3,300 entries encompassing >800 STs (accessed 22 July 2010).

Even finer resolution can be gained by multi locus sequence analysis (MLSA). This approach discriminates between isolates based on the concatenated nucleotide sequence of each typing gene, thus making use of the number, position and type of sequence changes causing allelic variation (Lorenzon et al. 2003; Maidhof et al. 2002; Naser et al. 2005; Thompson et al. 2005). A phylogenetic tree based upon multi locus sequence analysis of multiple *Salmonella* isolates indicates the evolutionary distances between the *S. enterica* subspecies and *S. bongori* (Figure 1-3).

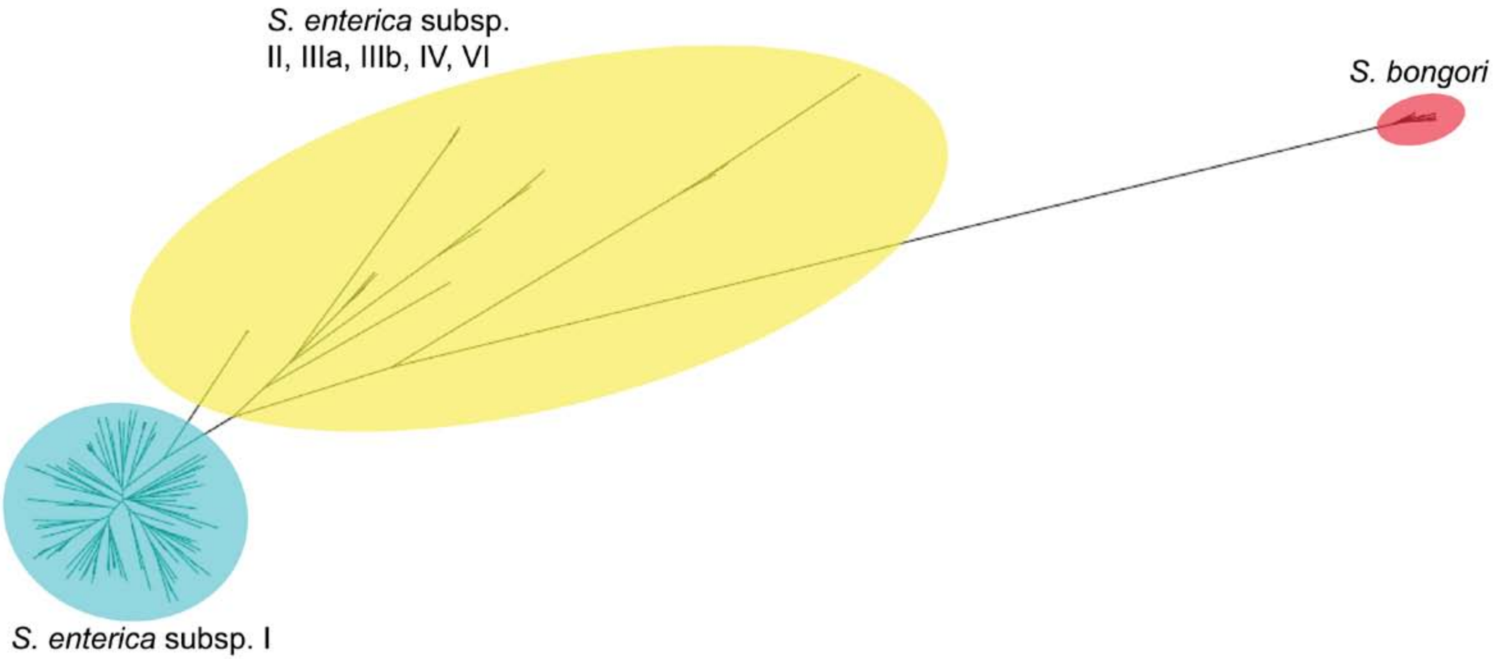


Figure 1-3 MLSA tree for *Salmonella*

Phylogenetic tree based upon concatenated sequence from seven housekeeping genes. Light blue ellipse, *S. enterica* subspecies I; yellow ellipse, *S. enterica* subspecies II, IIIa, IIIb, IV and VI; red ellipse, *S. bongori* (also known as subspecies V). Tree reproduced with permission from Maria Fookes.

1.1.2 *S. bongori*

In 1989, Reeves and colleagues demonstrated by MLEE that subgroup V (as it was then known) was so divergent from the remaining subgroups of *Salmonella* that they believed it should be elevated to separate species status (Reeves et al. 1989). *Salmonella bongori* is thus the second species of the genus *Salmonella*. Analyses using fluorescent amplified-fragment length polymorphism and MLST have both confirmed the earlier findings that *S. bongori* are the most divergent from the other subgroups and represent a binary split from a common ancestor with *S. enterica* (Falush et al. 2006; Scott et al. 2002).

S. bongori is composed of 23 serovars, which exhibit very little sequence variation (Guibourdenche et al. 2009) (Nick Thomson, personal communication). As a species, it is associated with infection mainly in cold-blooded vertebrates; human infections occur rarely and usually as a result of direct contact with reptiles (e.g. pets) (Woodward et al. 1997). However, one serovar, designated 48:z35;- has been documented in southern Italy as the cause of enteritis in young children (Pignato et al. 1998). A molecular analysis in the same region of 31 strains isolated over 15 years from humans, animals and the environment indicated that the isolates may be persisting in the environment and habitually causing human infection (Giammanco et al. 2002).

The whole genome sequencing of multiple *S. bongori* strains in progress at the Sanger Institute will provide an extremely useful resource for *Salmonella* genomics. While this species is not the ancestor of *S. enterica*, extensive genetic information regarding *S. bongori*, a pathogen of mainly cold-blooded animals, addresses the imbalance of knowledge currently held about *Salmonella* that largely comes from the subspecies of *S. enterica* that cause disease in birds and mammals.

Based upon a core set of genes shared between 15 *Salmonella*, *E. coli* and *Shigella*, *S. bongori* occupies an intermediate taxonomic position between *S. enterica* subspecies I and *E. coli* (N. Thomson, personal communication). It also has the smallest genome of any *Salmonella* sequenced (< 4.5 Mbp) and the lowest average G+C content.

1.1.3 *S. enterica* subspecies *enterica*

Currently, 1,547 serovars make up subspecies I, by far the largest in *S. enterica* (Guibourdenche et al. 2009). These serovars differ widely in the range of hosts they can infect and the diseases caused. The serovars of interest to this project are introduced below.

1.1.3.1 *S. enterica* serovar Typhi

S. enterica serovar Typhi (Typhi) is the causative agent of typhoid fever, a febrile systemic illness that is a form of enteric fever. Typhi was first described by Eberth and Koch in 1880 and first cultured by Gaffki in 1884. It is identified by the antigenic formula 9[Vi]:d:-, and specific biochemical reactions, including an inability to ferment ornithine and rhamnose. It causes disease solely in humans and current estimates indicate that Typhi causes over 21 million cases of typhoid annually (Bhutta and Threlfall 2009; Crump et al. 2004; Kothari et al. 2008). Research efforts have been focused upon the pathogenicity of Typhi and its host restriction, with the aim of understanding more about the host-pathogen interaction, and of finding and developing new drug targets and vaccines. Such studies typically concentrated upon the pathogenicity of one strain, with limited knowledge of population genetics.

Analysis by MLST addressed this issue and revealed that Typhi forms a monophyletic group that is approximately 50,000 years old (Kidgell et al. 2002). The 4.8 Mbp genome of Typhi CT18 is predicted to contain more than 4,500 protein coding sequences and 204 pseudogenes (Parkhill et al. 2001a). These pseudogenes are largely conserved (>90%) in Typhi Ty2 (Deng et al. 2003) and in the group of 19 Typhi isolates sequenced more recently (Holt et al. 2008), lending support to the concept that genome degradation has contributed to the host restriction of serovar Typhi. DNA acquisition has also played an important role in the evolution of salmonellae (Kingsley and Baumler 2000). In particular, Typhi contains *Salmonella* Pathogenicity Island (SPI)-7 (Liu and Sanderson 1995b; Wain et al. 2002) which encodes the *viaB* region: genes responsible for the synthesis and transport of the Vi capsular polysaccharide (Hashimoto et al. 1993; Virlogeux et al. 1995). SPI-7 also harbours the SPI-1 type III secretion system (TTSS) effector protein *sopE* (Hardt et al. 1998), and the *pil* genes which encode a type IVB pilus implicated in bacterial self-association (Morris et al. 2003; Tsui et al. 2003) and interaction with epithelial cells (Zhang et al. 2000). Culture collections of Typhi have been shown to contain SPI-7-negative strains (Nair et al. 2004), but the island is almost always present in fresh clinical isolates (Wain et al. 2005). This suggests that SPI-7 plays an important role during the infection process. Whilst all *S. enterica* serovar Paratyphi C and some *S. enterica* serovar Dublin isolates harbour very similar forms of SPI-7

(Pickard et al. 2003), the subtle differences may affect the properties encoded by specific island regions.

1.1.3.2 ***S. enterica*** **serovar Paratyphi A**

Infection with *S. enterica* serovar Paratyphi A (Paratyphi A) also causes enteric fever, being the causative agent of paratyphoid fever. It is identified by the antigenic formula 1,2,12:a:[1,5] which indicates that it is most commonly isolated as monophasic (H:a) but rare isolates are diphasic with phase 2 H:1,5 (Grimont and Weill 2007). Similarly to Typhi, Paratyphi A causes disease solely in humans. Whilst typhoid fever cases have long been known to outnumber paratyphoid, the prevalence of paratyphoid cases in several Asian countries is on the increase (Jin 2008; Ochiai et al. 2005; Palit et al. 2006; Woods et al. 2006). However, the risk factors for acquiring Paratyphi A may differ from those for Typhi (Vollaard et al. 2004), suggesting potential differences in transmission routes.

There are two full 4.6 Mbp genome sequences now available for Paratyphi A, which show the presence of ~ 4,200 protein coding sequences and over 170 pseudogenes per genome (Holt et al. 2009b; McClelland et al. 2004). The close genetic relationship believed to be shared between Paratyphi A and Typhi is the result of multiple recombination events occurring over a quarter of their genomes (Didelot et al. 2007). The remaining three-quarters are as diverse as any pair of *S. enterica* serovars. In addition, Paratyphi A does not harbour SPI-7, an intriguing observation that raises questions concerning how this serovar causes a disease clinically indistinguishable from typhoid and about the role of SPI-7 in typhoid fever (Maskey et al. 2006; Woods et al. 2006).

1.1.3.3 S. enterica serovar Paratyphi C

S. enterica serovar Paratyphi C (Paratyphi C) is capable of causing invasive disease in humans and is probably restricted to this host (MLST study in progress, Satheesh Nair, personal communication). Paratyphi C shares its antigenic formula (6,7:c:1,5) with four other serovars that are recognised as variants based on biochemical properties: Typhisuis, Choleraesuis, Choleraesuis var. Kunzendorf, and Choleraesuis var. Decatur (Grimont and Weill 2007). In most clinical settings, defining a *Salmonella* isolate by its antigenic formula is the furthest point to which typing is taken. Hence, distinguishing serovars beyond this has only been undertaken in a research environment or in reference labotatories. Besides biochemical testing, two other techniques, ribotyping and IS*200* fingerprinting have been used to separate these five serovars (Uzzau et al. 1999)(S. Nair, personal communication). More recently, a global collection of Paratyphi C was characterised by MLST, revealing that all strains formed a complex of three STs, separated only by single allele differences. The complex was dominated by ST146 (34/47 strains), with ST90 and ST114 occurring much more rarely (10/47 and 3/47 respectively, Table 1-1). The strains were also investigated for the presence of pathogenicity islands, revealing that all carried SPI-7, though not in identical form to the Typhi SPI-7 (S. Nair, personal communication). While the *viaB* locus was present, only one strain out of 47 was found to express the Vi antigen.

One of the rare Paratyphi C STs (ST 114) has been sequenced, revealing a 4.8 Mbp genome encoding ~ 4,600 genes and 152 pseudogenes (Liu et al. 2009). This gives insight into the evolution of a potentially human host-restricted serovar separate from Typhi and Paratyphi A, but may not be representative of Paratyphi C as a whole. Current sequencing efforts are focused upon a strain from ST146, the commonest form currently in circulation (S. Nair, personal communication).

Table 1-1 Host specificities and multi-locus sequence types for 6,7:c:1,5 *Salmonella*

MLST complex*	**ST**	**Serovar**	**Source**		
			Human	**Pig**	**Unknown**
145	66	CK	28	25	3
	68	C	2	4	5
	133	CK	1	0	0
	139	C	0	0	4
	145	CK	6	7	20
	246	†	0	0	1
	363	CK	0	0	1
146	90	P	5	0	5
	114	P	0	0	3
	146	P	18	0	16
147	147	T	0	1	3

Adapted with permission from S. Nair and based upon a global collection of >150 strains. * Sequence Types (ST)s within a complex share 6 alleles; C, Choleraesuis; CK, Choleraesuis var. Kunzendorf; P, Paratyphi C; T, Typhisuis; † unknown serovar.

1.1.3.4 *S. enterica* serovar Choleraesuis

S. enterica serovar Choleraesuis (Choleraesuis) shares the antigenic formula of Paratyphi C. It is partly defined by its host range; Choleraesuis is adapted to pigs, where it causes swine paratyphoid fever, but it is also capable of causing disease in humans (Table 1-1). Human infection is often associated with underlying diseases in the patient (Chiu et al.

2004a; Chiu et al. 2004b; Wang et al. 2006) and is highly invasive. A measure of the invasiveness of a serovar can be obtained by dividing the number of isolates derived from blood by the total number of isolates to give an ‘invasive index’. The higher the percentage, the more invasive the serovar. There are reports from Taiwan, the USA and England & Wales of invasive indexes ranging from 52% to 74%, indicating that infection with Choleraesuis often leads to systemic disease in humans (Chiu et al. 2006; Langridge et al. 2009a; Lauderdale et al. 2006; Threlfall et al. 1992; Vugia et al. 2004).

MLST analysis of Choleraesuis, Choleraesuis var. Kunzendorf and Choleraesuis var. Decatur showed that the first two form a complex of 7 STs, dominated by ST66 and ST145. The latter did not form part of this complex; it was separated into three different STs, each different by 4 or more alleles, a strong indicator of polyphyly. The genome of Choleraesuis SC-B67 (ST 66) is available and has been shown to contain 151 pseudogenes (Chiu et al. 2005), a level of genome degradation perhaps reflecting its strong adaptation to swine.

1.1.3.5 *S. enterica* serovar Typhisuis

S. enterica serovar Typhisuis (Typhisuis) is a host-restricted serovar that is not isolated from humans. Infection in the natural host, pigs, ranges from enterocolitis to chronic paratyphoid fever (Rodriguez-Buenfil et al. 2004; Uzzau et al. 2000). Typhisuis is a cystine auxotroph and forms a monophyletic group by MLEE and by MLST (from the few isolates tested) (Boyd et al. 1993; Selander et al. 1990; Uzzau et al. 2000; S. Nair, personal communication). Differentiation from Paratyphi C is based on the prototrophic nature of the latter and that all Typhisuis strains typed by MLST thus far are ST147 (S. Nair, personal communication). The use of arabinose and trehalose fermentation tests has also been described, since Typhisuis utilises both but Paratyphi C only utilises arabinose (Table 1-2) (Uzzau et al. 1999). Differentiation from Choleraesuis is based upon tartrate utilisation, and for var. Kunzendorf, hydrogen sulphide (H_2S) production.

Table 1-2 Metabolic properties that distinguish 6,7:c:1,5 serovars

Serovar of *Salmonella*	Tartrate	Dulcitol	H_2S	Trehalose	Arabinose
Paratyphi C	+	+	+	-	+
Choleraesuis	+	-	-	-	-
Choleraesuis var. Kunzendorf	+	-	+	-	-
Typhisuis	-	-	-	+	+

+ indicates production of substrate (H_2S) or a positive growth phenotype upon substrate, - indicates the opposite. Taken from (Le Minor et al.) and (Uzzau et al. 1999).

1.1.3.6 S. enterica serovar Gallinarum

S. enterica serovar Gallinarum (Gallinarum) was described as the causative agent of an invasive typhoid-like disease in chickens nearly a century ago (Smith and TenBroeck 1915; St John-Brooks and Rhodes 1923). This serovar is non-motile and has the antigenic formula 1,9,12:-:-. Gallinarum is highly adapted to the chicken host and hence presents little public health threat (Shivaprasad 2000). However, in common with human-restricted Typhi, the genome sequence of this serovar has revealed a large amount of genome degradation (~7% of the genome is represented by pseudogenes) (Thomson et al. 2008). Gallinarum forms a related strain cluster with both host-adapted (*S. enterica* serovars Dublin and Pullorum) and host-generalist (*S. enterica* serovar Enteritidis) serovars (Thomson et al. 2008).

1.1.3.7 S. enterica serovar Typhimurium

S. enterica serovar Typhimurium (Typhimurium) is one of the leading causes of foodborne gastroenteritis (Zhang et al. 2003) and is the focus of national surveillance systems across the globe. The antigenic formula for this serovar is 1,4,[5],12:i:1,2 but many subtypes have been described, mainly defined by phage typing (Anderson et al. 1977; Callow 1959) and less frequently by MLST. These are often referred to as 'variants' or 'pathovariants' of Typhimurium as they differ in both host range and level of host adaptation (Rabsch et al. 2002). For example, in both developed and less developed countries, Typhimurium is one of the most commonly reported causes of extra-intestinal non-typhoidal salmonellosis (Brown and Eykyn 2000; Kariuki et al. 2005). Such systemic infections are clinically distinct from those caused by Typhi and Paratyphi A (Gordon et al. 2002; Graham et al. 2000). Recently, MLST has been used to analyse a set of Typhimurium strains that caused invasive disease in humans in sub-Saharan Africa; ST313 was identified as the dominant type (Kingsley et al. 2009). At the other end of the spectrum, ST19 is a prototypical Typhimurium variant, having been recorded in the MLST database as being isolated from cattle, pigs, humans, poultry and horses. This ST, which includes the mouse-virulent strain SL1344, is often the host-generalist comparator used in studies comparing host-adapted and non-adapted strains of *Salmonella enterica.*

1.2 Host adaptation and restriction

1.2.1 Definitions

The host range of a bacterium is defined by which higher organisms (hosts) it is capable of naturally infecting. Bacteria which can infect multiple hosts are known as host-promiscuous or host generalists. When bacteria are highly associated with infection in one organism but remain capable of natural infection in others, they are referred to as host adapted. This is the case for a number of *Salmonella* serovars, including Dublin (adapted to cattle), Choleraesuis (to pigs), and Obortusovis (to sheep). However, when bacteria are only capable of infecting a single host, this is termed host restriction or specialism and may be due to both gene acquisition and loss of gene function. Often, a spectrum of host specificities exist within the same genus, e.g. *Bordetella*, *Escherichia* and *Mycobacterium*, as well as *Salmonella*.

1.2.2 Process of adaptation and restriction

The genomes of pathogenic bacteria are continually changing through various evolutionary processes. Acquisition of new virulence factors or even pathogenicity islands can occur via horizontal gene transfer and subsequently affect the host range and virulence of the recipient. In *Francisella tularensis*, two of the four subspecies are human pathogenic strains; the remaining two are rarely implicated in disease. Genome comparisons between 3 of the subspecies revealed 41 genes present in the pathogenic strains that were absent from the non-pathogen (Rohmer et al. 2007). Similarly in *Listeria*, the two pathogenic species, *L. monocytogenes* and *L. ivanovii* both contain the

'*Listeria* virulence locus', which encodes *hlyA*, a toxin required for virulence. While the locus is present in one of the non-pathogenic *L. seeligeri*, it is apparently inactivated by an insertion. The same locus is absent from all other non-pathogenic *Listeria* species (Vázquez-Boland et al. 2001).

Acquired traits only form part of the picture regarding host adaptation and restriction. Entirely host restricted organisms are often characterised by extensive genome decay, through insertion sequence (IS) element proliferation, genomic rearrangement and pseudogene formation.

1.2.2.1 IS elements

Increase in IS element copy number has been linked with the early stages of genome reduction (Moran and Plague 2004). *Shigella flexneri* and *Bordetella pertussis* are prime examples of genomes containing high densities of IS elements (Parkhill et al. 2003) (Wei et al. 2003). This is believed to be associated with relatively recent host adaptation (Moran and Plague 2004). The more IS elements that are present in a genome, the greater the likelihood of recombination between these repeats. Indeed, a genome comparison of *B. pertussis* and *B. bronchispetica* revealed that almost 90% of the 150 recombination events in *B. pertussis* were delimited by IS elements (Parkhill et al. 2003). Conversely, Typhi has relatively few IS elements; 26 copies of IS200F and 3 copies of other ISs make up the entire complement in strain Ty2 (Deng et al. 2003).

1.2.2.2 Genome rearrangement

While rearrangements due to recombination of IS elements are minimal, large rearrangements have occurred around the seven *rrn* operons present in the *Salmonella* genome (Liu and Sanderson 1995c). While the order of genomic fragments (delimited by the *rrn* operons) is stable and conserved in Typhimurium, 21 different arrangements were observed in a 127 strain collection of Typhi (Liu and Sanderson 1996). Gene order within these fragments is conserved, but the mechanism(s) governing the genomic balance between origin and terminus of replication and any gene dosage effects appear to have a reduced role in Typhi (Liu and Sanderson 1996). Out of 10 Paratyphi A strains tested, only one genomic arrangement was found, although this represents an inversion of half the genome with respect to Typhimurium (Liu and Sanderson 1995a). It is possible that more rearrangements would be found if a larger strain collection was tested.

1.2.2.3 Pseudogenes

Genes whose functions have been lost over evolutionary time (pseudogenes) were initially thought to be rare in bacterial genomes, due to their compact genome size and general lack of non-coding DNA (Lawrence et al. 2001). However, as more bacterial genome sequences became available, especially those of recently emerged pathogens, pseudogenes were recognised as particularly common among obligate symbionts (Cole et al. 2001; Parkhill et al. 2003; Wei et al. 2003). Identification of pseudogenes remains largely due to comparative analyses between closely related genomes, looking for stop codons, small indels and truncations that may affect gene function (Dagan et al. 2006; Lerat and Ochman 2004; Lerat and Ochman 2005).

Pseudogene formation is a hallmark of many host restricted organisms. One extreme case is *Mycobacterium leprae*, where 41% of the potential protein-coding genes are non-functional (Cole et al. 2001). Smaller but nonetheless significant pseudogene complements have been discovered in *Shigella flexneri*, and in *Yersinia pestis*, which was extended upon comparison to the less pathogenic *Y. pseudotuberculosis* (Chain et al. 2004; Parkhill et al. 2001b; Wei et al. 2003). In all these cases, pseudogenes were predicted by comparison with intact orthologues. Hence, genes appear to be inactivated more quickly than they are removed by deletion (Cole et al. 2001). As expected of a host generalist, the sequence of Typhimurium strain LT2 (ST19) contains very few pseudogenes, making up only ~ 0.9% of its gene complement (McClelland et al. 2001). Host specialists Typhi and Paratyphi A however, have between 4.5 and 5% of their genes as pseudogenes (McClelland et al. 2004; Parkhill et al. 2001a). Comparisons of individual pseudogenes have shown that at least a third of those shared (estimates range from 28-66 depending upon genome annotation) between both serovars contain different inactivating mutations (Holt et al. 2009b; McClelland et al. 2004). Thus, these two pathogens, if adapting to the niche of the human host by loss of gene function, seem to be doing so along a path of convergent evolution. The degree of host specificity displayed by particular *Salmonella* serovars generally correlates positively with the amount of gene degradation revealed upon whole genome sequencing. Chicken-restricted Gallinarum remains the most extreme example, with over 300 pseudogenes (7.2%). Choleraesuis and Enteritidis, which display strong host associations with swine and chickens respectively, have 151 (3.3%) and 113 (2.6%) pseudogenes each. As mentioned above, a host promiscuous variant of Typhimurium contains only 39 (0.9%) pseudogenes. One serovar that does not quite fit this pattern is Paratyphi C. MLST analysis suggests this serovar is human restricted, but the pseudogene complement of 149 (3.3%) places it with Choleraesuis, a serovar capable of infecting both pigs and humans (Liu et al. 2009). However, this estimate may be revised upwards when the sequence from the more representative ST146 isolate becomes available.

1.3 Disease and infection of the host

Typhoid and paratyphoid fever (collectively enteric fever) occur primarily in regions of the world where clean water supplies and sanitation are inadequate. Over 21 million cases occur annually (Bhutta and Threlfall 2009; Crump et al. 2004; Kothari et al. 2008).

1.3.1 Clinical features of enteric fever

With enteric fever, the fever itself is the only consistent symptom (Parry 2004; Parry et al. 2002), although abdominal pain or discomfort, muscle and/or joint pain, and headache are frequently observed. Recent studies in Nepal suggest that Typhi and Paratyphi A cause clinically indistinguishable syndromes (Maskey et al. 2006; Woods et al. 2006). Once an appropriate antibiotic course has been administered, these symptoms usually resolve quickly. However, persistence of several weeks can occur in an untreated patient (Parry 2004; Parry et al. 2002). Complications can develop from Typhi infection (Deng et al. 2003), the most severe of which is gastrointestinal haemorrhage and perforation (Chanh et al. 2004; Everest et al. 2001). This condition requires both surgical and antimicrobial intervention and carries a high risk of mortality (Butler et al. 1985). An estimated 5% of typhoid patients become chronically infected by retention of Typhi in the gall bladder and continue to excrete Typhi for many years (Parry et al. 2002). These typhoid carriers not only pose a significant health risk to others, but also have a higher risk of developing cancer of the gallbladder, pancreas and large bowel (Caygill et al. 1995; Dutta et al. 2000; Shukla et al. 2000). Carriers also provide a reservoir for Typhi, contributing significantly to the persistence of typhoid in endemic regions (Roumagnac et al. 2006). Carriage has recently been demonstrated for Paratyphi A, but not to the same level as Typhi (Khatri et al. 2009). What effect Paratyphi A carriage has upon excretion rates and persistence in endemic areas has yet to be determined.

1.3.2 Infection of the host

1.3.2.1 Typhi

Modern publications on typhoid fever in humans are few, and much of the information on typhoid pathogenesis has been gained through studies on Typhimurium in mice. The infection of mice by Typhimurium is often cited as a model for typhoid fever, since this serovar causes a typhoid-like disease in the mouse, but translating these findings to typhoid fever in humans requires careful interpretation (Sabbagh et al. 2010). The current view of typhoid pathogenesis is shown in Figure 1-4. Typhi is transmitted via the faecal-oral route, with sufferers ingesting an infectious dose of between 10^5 and 10^9 bacterial cells in contaminated food or water (Wain et al. 2002). After passing through the stomach, Typhi invades the gut epithelium of the terminal ileum, possibly using the cystic fibrosis transmembrane conductance regulator (CFTR) for entry (Pier et al. 1998).

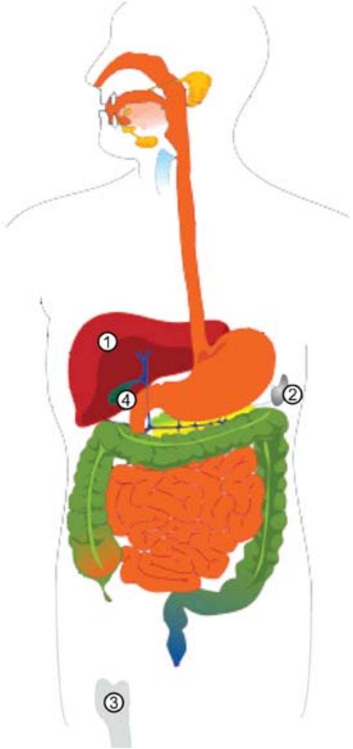


Figure 1-4 Pathogenesis of typhoid fever in humans

Adapted from (Wain et al. 2002). Transmission occurs via the faecal-oral route, leading to the stomach. After penetrating the intestinal epithelium, Typhi disseminates through the body in the bloodstream and seeds the liver (1), spleen (2), and bone marrow (3). Some patients make a full recovery, but a small percentage progress to an asymptomatic chronic infection (carriage) in the gall bladder (4). Human image from http://commons.wikimedia.org.

Synthesis of the Vi capsular polysaccharide (encoded on SPI-7) is down-regulated under the low osmotic conditions found at the intestinal epithelial barrier. These conditions also promote the secretion of effector proteins via the SPI-1 encoded type III secretion system and hence an adhesive and invasive phenotype (Sukhan 2000; Wehland and Bernhard 2000). This invasion triggers the secretion of interleukin(IL)-6 from host cells, and the bacteria are taken up by or invade macrophages. In addition, CD18+ host cells that have migrated into the gut lumen may also take up bacteria, which they transfer across the gut epithelium when they migrate back into gut tissue (Wain et al. 2002). High expression of

SPI-1 at this point results in caspase-1-mediated death of macrophages, release of more cytokines and Typhi is subsequently disseminated throughout the body via the bloodstream (House et al. 2001). Symptoms of typhoid fever are not yet apparent at this stage of the infection; it is a secondary bacteraemia resulting from bacterial replication within the liver, spleen and bone marrow that causes the onset of clinical symptoms, when Typhi can usually be cultured from the blood or bone marrow, albeit at very low levels (<1 CFU/mL) (Wain et al. 2002).

1.3.2.2 Differences in Typhimurium

Typhimurium is ingested in the same manner as Typhi, although gastroenteritis rather than systemic disease is the usual outcome. Typhimurium also invades the gut epithelium, but is associated with host cell release of IL-8 and recruitment of neutrophils to the site of infection, causing localised inflammation (House et al. 2001). Bacteria do not disseminate through the body, and the diarrhoeal symptoms caused by the infection are usually self-limiting.

1.3.3 Molecular basis of infection

1.3.3.1 Invading the macrophage

Wildtype Typhi is classified as a hazard group 3 organism, which is one reason why much of the literature on the interaction of *Salmonella* with macrophages is derived from studies performed with Typhimurium (hazard group 2) in murine cells.

The importance of infection and survival within the macrophage was noted when it was shown that Typhimurium mutants unable to survive inside these cells were avirulent in the mouse (Fields et al. 1986). A screen of almost 10,000 transposon mutants for survival in macrophages identified a number of auxotrophies that were also associated with decreased virulence in the murine model. These included requirements for purines, pyrimidines and histidine, which indicates the importance of bacterial metabolism in survival inside the macrophage (Fields et al. 1986). A proteomic study from the same group showed that over thirty Typhimurium proteins were synthesised during murine macrophage infection, with the heat-shock proteins DnaK and GroEL the most abundant (Buchmeier and Heffron 1990).

It has been postulated that Typhi is unable to infect other hosts because it lacks genes present in Typhimurium that allow it to colonise a broad host range (Morrow et al. 1999). Prior to the availability of full genome sequences, genomic subtractive hybridisation was used to isolate gene sequences in Typhimurium not present in Typhi. Using a technique called ‘selective capture of transcribed sequences’ (SCOTS), RNA from Typhimurium inside macrophages was labelled and hybridised against the subtracted sequences to identify those which were expressed inside the macrophage (Morrow et al. 1999). The

SCOTS technique identified a putative transcriptional regulator and a novel fimbrial operon expressed by Typhimurium inside macrophages, neither of which are present in the Typhi genome. However, when tested, mutations in the regulator did not affect virulence in the mouse or survival within the macrophage (Morrow et al. 1999). As such then, there may not be individual genes that can be linked with broad host avirulence in Typhi, but both the absence of genes present in broad-host range *Salmonella*, and the presence of pseudogenes in the Typhi genome likely contribute to the restriction of this serovar to its human host.

The SCOTS technique was also used to look (for the first time) at which Typhi genes are expressed inside human macrophages. In the study, over twenty Typhi genes were identified, including eight of unknown function (Daigle et al. 2001). Other predicted functions included Vi capsule biosynthesis and the stress response. In a follow-up study, effectively a complementation of the Typhimurium subtractive hybridisation experiment was performed to look for Typhi-specific genes expressed inside the human macrophage (Faucher et al. 2005). Thirty-six Typhi genes were found to be expressed intracellularly, of which twenty-five were encoded on pathogenicity islands including SPI-7, and prophage elements.

By combining SCOTS with microarray analysis, a more detailed picture began to emerge of the environmental conditions encountered by Typhi during macrophage infection. As a functional class, iron transport was repressed, as were motility and peroxide-induced functions, while antimicrobial peptide resistance was induced (Faucher et al. 2006). Typhi was therefore predicted to reside in an environment that is neither acidic nor oxidative.

Similar studies have been performed using Typhimurium, and have made other predictions about the intracellular environment. Microarray analysis showed that Typhimurium expression of amino acid, potassium and iron transport was not induced but that phosphate and magnesium may be limiting (Eriksson et al. 2003). Also, once inside murine macrophages, Typhimurium has been predicted to use gluconate and related carbohydrates for growth (Eriksson et al. 2003). However, Typhi does not encode the *dgo* genes required to break down these compounds, and there was no SCOTS evidence that the relevant transporters were being induced, suggesting that Typhi may use different carbon sources to Typhimurium for intracellular growth (Faucher et al. 2006).

1.3.3.2 The **Salmonella**-*containing vacuole*

In the early 1990s, microscopy techniques were used to investigate how *S. enterica* invade host epithelial cells and macrophages. It was initially discovered that *S. enterica* triggered membrane ruffling upon contact with the surface of host cells (Galan et al. 1992), and later shown that this was directly related to the formation of intracellular 'spacious phagosomes' containing *S. enterica* (Alpuche-Aranda et al. 1994; Garcia-del Portillo and Finlay 1994). A large body of work has been conducted regarding the generation and maintenance of what are now called '*Salmonella*-containing vacuoles' (SCVs), but this has, for the most part, been concentrated upon Typhimurium (see (Holden 2002) and (Knodler and Steele-Mortimer 2003) for reviews). However, one early study looked at both Typhi and Typhimurium and concluded that the mechanisms involved in invasion and intracellular trafficking are very similar for both serovars in epithelial cells, which are non-phagocytic (Mills and Finlay 1994).

Normally, eukaryotic cells, including epithelial cells, utilise the endocytic pathway for internalising, sorting, recycling and degrading molecules recovered from the cell surface or from extracellular space. Such molecules are endocytosed and progress through early endosomes, multivesicular bodies, late endosomes and finally lysosomes (Knodler and Steele-Mortimer 2003). Since invasion of epithelial cells requires expression of SPI-1 and SPI-1 related genes, the experimental conditions for achieving cell infection are relatively standardised and have produced a consensus view of SCV biogenesis and maturation in this cell type (Figure 1-5).

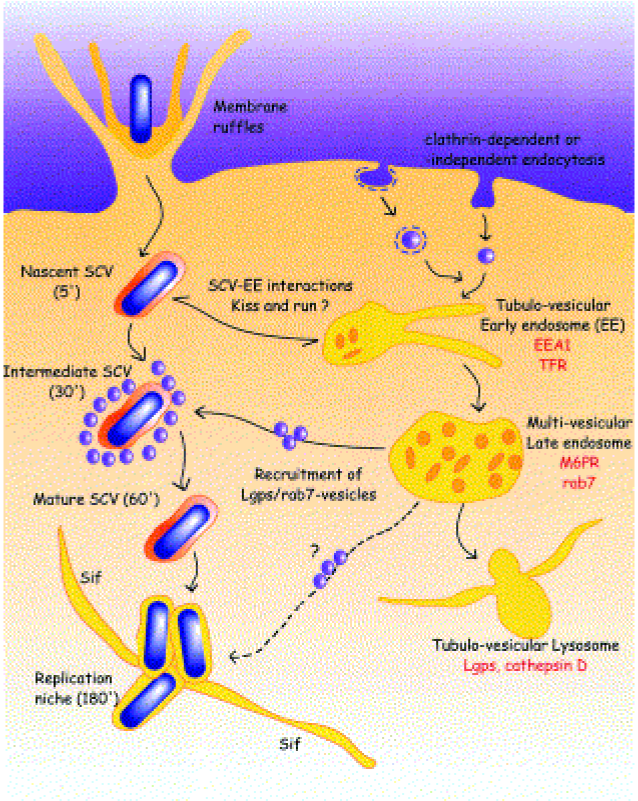

Membrane ruffles
clathrin-dependent or independent endocytosis
Nascent SCV (5')
SCV-EE interactions Kiss and run ?
Tubulo-vesicular Early endosome (EE)
EEA1
TFR
Intermediate SCV (30')
Multi-vesicular Late endosome
M6PR
rab7
Recruitment of Lgps/rab7-vesicles
Mature SCV (60')
?
Sif
Tubulo-vesicular Lysosome
Lgps, cathepsin D
Replication niche (180')
Sif

Figure 1-5 Intracellular pathway of *Salmonella* in epithelial cells

From (Gorvel and Méresse 2001). Membrane ruffles mark the beginning of invasion in non-phagocytic host cells. Once inside the host cell, *Salmonella* are found in nascent SCVs, which are modulated by *Salmonella* to achieve maturation and enable replication. The approximate time course is shown in minutes post infection.

In macrophages, molecules are phagocytosed and the subsequent phagosome interacts sequentially with the endocytic compartments and requires the activity of Rab GTPases. However, when *Salmonella* invade, while interactions do occur with the endocytic pathway, degradation is avoided by blocking the fusion of the SCV with terminal acidic lysosomes (Buchmeier and Heffron 1991; Ishibashi and Arai 1990). There has been greater debate over the process of SCV maturation in macrophages, for multiple reasons. Experimental conditions for infection of macrophages are much less consistent and since these cells are phagocytic, the mechanism of *Salmonella* entry varies. Conflicting evidence has also been presented regarding the interaction of SCVs with the late endocytic pathway, with some studies showing that fusion between the SCV and lysosomes is blocked (Buchmeier and Heffron 1991; Ishibashi and Arai 1990), but others demonstrating that this fusion does occur (Drecktrah et al. 2007; Oh et al. 1996). However, a recent study has documented an imbalance in the ratio of acidic lysosomes to SCVs caused by Typhimurium, such that the stock of lysosomes is exhausted before all the SCVs have been targeted (Eswarappa et al. 2010). This may provide a unifying explanation for the evidence for and against SCV/lysosome fusion, and serves to highlight the need for further investigation of the roles played by Typhi and Typhimurium inside macrophages.

1.4 *Genetic diversity in* Salmonella

The variety of diseases caused by different *S. enterica* serovars have helped to provide impetus for improving the methods by which serovars and strains can be distinguished. Initially, the main focus of such methods was for epidemiological and public health reasons, but as the resolution offered by new technologies has improved, such methods have become useful from evolutionary and population biology viewpoints as well.

1.4.1.1 Phage typing

In the late 1930s, early 1940s, a typing scheme for *Bacillus typhosus* (Typhi) was being used as an epidemiological tool to differentiate subtypes based upon resistance to Vi phage adapted to various *B. typhosus* strains (Anderson and Felix 1953; Felix 1943). A similar scheme was soon being applied to Typhimurium, which in the early stages used 29 phages to distinguish 34 types (Callow 1959). At this time, other schemes were also in use for Paratyphi A and Paratyphi B (Banker 1955; Felix and Callow 1943). By the late 1970s, phage typing was the method of choice for the epidemiological tracking of Typhimurium outbreaks, and the scheme had expanded to include 34 phages used to differentiate 207 types (Anderson et al. 1977). By the late 1980s, Enteritidis had become the top serovar causing human infection in the UK, so a phage typing scheme was developed that distinguished 27 types with 10 phages (Ward et al. 1987). Phage typing is still used today, particularly to differentiate the economically important subtypes of Enteritidis and Typhimurium such as PT4 and DT104 respectively. Enteritidis is a prominent cause of non-typhoidal salmonellosis, and phage type (PT) 4 caused 49% of

Enteritidis outbreaks in the USA in 1999 (Patrick et al. 2004). Similarly, Typhimurium definitive type (DT) 104 has caused smaller outbreaks of multidrug resistant salmonellosis since the early 1990s (Helms et al. 2005).

1.4.1.2 Plasmid profiling

Over 30 years ago, methods for extracting plasmid DNA were being optimised to aid in differentiating between bacterial isolates (Kado and Liu 1981; Schaberg et al. 1981). With a clean DNA sample, plasmids were run out on a gel by electrophoresis, resulting in a distinct ‘fingerprint’ pattern (Schaberg et al. 1981). The utility of this technique was shown even outside a clinical setting by Brunner and colleagues in Switzerland, where Typhimurium isolates were categorised in 6 plasmid patterns (PPs), with PP1 identified as the type responsible for a minor epidemic (Brunner et al. 1983). A later study looked at multiple outbreaks of Typhimurium in the USA and found that, in outbreaks where unrelated isolates were also analysed, the outbreak isolates were differentiated from the unrelated 8/9 times using plasmid profiling (Holmberg et al. 1984). Phage typing only differentiated 6/9 and antibiotic resistance screening 4/9. A Spanish study, looking mainly at Enteritidis, also compared plasmid profiling, phage typing and resistance screening to determine the relative merits of each (Borrego et al. 1992). They found plasmid profiling to be superior to phage typing and antibiotic resistance screening as an epidemiological marker although they also stated that none gave total discrimination. On a somewhat contradictory note, Threlfall and colleagues directly compared phage typing and plasmid profiling and found that a particular plasmid pattern was found in more than one phage type and that multiple plasmid patterns could be found in a single phage type (Threlfall et al. 1989).

Plasmid profiles therefore can be associated with a specific strain of *Salmonella*, but these may also be transferred. Fine typing of plasmids and the strains in which they reside is shedding more light on this relationship (Minh-Duy Phan, personal communication).

1.4.1.3 Signature tagged mutagenesis

In the mid-1990s, transposon-based mutant screening emerged. For the first time, multiple mutants could be screened simultaneously for attenuation in an animal model (Hensel et al. 1995). In the first published signature-tagged mutagenesis experiment, unique, identifying DNA tags were ligated onto a Tn5-derived transposon that was used to conjugate a mouse-virulent variant of Typhimurium (Hensel et al. 1995). Over 1100 successfully transformed exoconjugants were separated into 96-well microtitre plates and pools of 96 mutants were screened through mice. Bacteria recovered from the spleen were subjected to a tag-specific PCR incorporating a radiolabelled nucleoside triphosphate (dCTP). A probed colony blot of 'recovered' bacterial PCR products was then compared with the equivalent blot from the inoculum. This typified a 'negative' selection screen, where transposon mutants that were present in the inoculum but absent from the recovered pool represented those attenuated in the infection model. A total of 39 attenuated mutants were identified and the transposon-chromosome junction sequenced for 28 of them. Five of these mutants were in genes related to type III secretion and led the authors to investigate further (Shea et al. 1996). An entire virulence locus was then identified; yielding the first description of SPI-2, an island unique in *Salmonella* to the *enterica* species (Hensel et al. 1997a; Shea et al. 1996), that was later demonstrated to be required for the survival of *S. enterica* in macrophages (Cirillo et al. 1998; Hensel et al. 1998).

1.4.1.4 Microarrays

The advent of whole genome sequencing quickly led to the production of bacterial microarrays. These were initially the result of PCR products spotted onto glass slides that were probed with fluorescently labelled genomic DNA (Dziejman et al. 2002; McClelland et al. 2001; Smoot et al. 2002), but later oligonucleotide arrays were also produced. Microarrays were used as a tool for gene expression profiling and for phylogenetic typing (Lucchini et al. 2001). Using microarrays to type related bacteria meant that differences in gene content could be visualised and quantified without needing the full genome sequence of every strain tested. In 2003, Porwollik and colleagues produced a non-redundant microarray for Typhimurium and Typhi that contained probes for all the genes in Typhimurium and was supplemented with probes for any Typhi genes that were > 10% divergent from Typhimurium (Porwollik et al. 2003). This established the concept of using close homologues to characterise multiple serovars with the same underlying microarray (Porwollik et al. 2004). Characterisation of almost 80 strains showed that there may be hundreds of genes different between strains of the same serovar and that in some cases there was more intra-serovar variation than inter-serovar variation (Porwollik et al. 2004). DNA microarrays have also been used to investigate the role of prophage-like elements in generating diversity within *S. enterica* (Thomson et al. 2004).

Using a microarray based on Typhi, supplemented with novel Typhimurium probes, Thomson and colleagues were able to compare strains from 20 serovars against the phage complement present in Typhi. They found that the Typhi strains harboured a set of temperate bacteriophage unique from all the other serovars tested and were also able to detect more subtle intra-serovar variations (Thomson et al. 2004). Using the Sanger sequenced Enteritidis strain for microarray construction, a study found that phage type (PT)-8 strains harboured a particular set of phage genes believed to be the molecular basis of the distinction of PT8 from PT4, the two types responsible for the majority of infection caused by Enteritidis (Porwollik et al. 2005).

Transposon-based assays also took a step forward with the production of microarrays. Techniques such as TraSH (transposon site hybridisation) and TMDH (transposon-mediated differential hybridisation) relied upon the use of microarrays to determine the insertion sites of transposons recovered from mutants passed through a selective screen (Chaudhuri et al. 2009; Sassetti et al. 2001). In TMDH, RNA was generated by outwards *in vitro* transcription from the transposon into adjacent genomic DNA and hybridised to the microarray. Conditionally essential genes were identified when a strong hybridisation signal from the input sample was missing in the output. However, background levels of hybridisation made distinguishing the on/off signal difficult, and as with all microarrays, a key limitation was that information could only be gained about what was on the array – novel genes and intergenic regions could not be assessed.

1.4.1.5 High-throughput sequencing

As Typhi is genetically monomorphic, gaining deeper insight into global diversity and evolutionary history required a high resolution approach (Achtman 2008). Two hundred 500 bp DNA fragments were sequenced from a global collection of >100 strains to look for informative mutations. Nineteen mutations that marked the evolutionary history of Typhi were found, with a further 69 that helped to define 59 distinct haplotypes (Roumagnac et al. 2006). Unusually, the minimal spanning tree generated from this data was rooted in an extant haplotype, H45, from which several lineages have descended. What was also striking was that many haplotypes were found on multiple continents, including H45 which was isolated in Africa, Asia and North America. This sheds little light on the question of where Typhi evolved. An investigation of a further 161 Typhi isolates from Indonesia looked at 84 single nucleotide polymorphisms (SNPs) as markers of genome variation (Baker et al. 2008). These isolates were assigned to nine haplotypes, indicating that multiple haplotypes were in circulation and that such haplotypes persist, as some were isolated repeatedly over a 30 year period (Baker et al. 2008). SNP genotyping can be achieved with systems like the Illumina GoldenGate and Sequenom massARRAY. However, this does not allow for novel SNP detection. To achieve this, Holt and colleagues sequenced multiple strains of Typhi using both 454 and Illumina (formerly Solexa) sequencing (Holt et al. 2008) where the isolates were chosen to represent major nodes from the phylogenetic tree (Roumagnac et al. 2006). By comparing generated sequence reads to the reference sequence of Typhi CT18, almost 2000 SNPs were detected, 10-fold more than previously, which allowed an even higher-scale resolution of the tree with improved branch length estimates. While the cost of sequencing isolates individually was prohibitive to genome-wide SNP detection, a method was developed to reduce this cost. Six individual strains of Paratyphi A and a pooled DNA sample, containing an equivalent amount from each strain, were sequenced using the short read high-throughput Illumina system (Holt et al. 2009a). Of 550 SNP loci checked in each strain, over 400 had sufficient high quality sequence coverage to estimate reliable frequencies across the six strains. With the pooled sample, genome coverage of 40x was achieved, and the sequences obtained were compared with the loci and allele frequencies from the individual strains. The sensitivity of SNP detection in the pool was 100% for SNPs that occurred in 3 or more strains, although this declined to 37% if a SNP occurred in just one strain. Overall, this represented a cheaper method of sampling a bacterial population with the aim of unbiased detection of genetic variation (Holt et al. 2009a). More recently however, the ability to ‘tag’ samples within a pool has become possible, allowing multiple strains to be sequenced simultaneously on the Illumina Genome Analyzer II platform. This has been used to characterise a set of 63 *Staphylococcus aureus* strains, with 23 x coverage achieved on average per strain relative to the reference and revealing over 4000 SNPs (Harris et al.). Making use of whole genome sequences yields the optimum resolution, especially between bacterial isolates sharing the same multi-locus sequence type.

1.5 Metabolism in Salmonella

1.5.1 Phenotypic analysis and biochemistry

By the 1940s, typhoid and paratyphoid epidemics were largely a thing of the past in the USA, as serotyping these 'salmonelloses of human origin' had been the focus of a concerted public health effort (Borman et al. 1943). Instead, attention was turning towards more precise definitions of those *Salmonella* isolates that did not fall into the typhoid group. At this point, various metabolic tests were in place only to identify isolates to the genus level. For example, *Salmonella* were known to be negative for both indole and the Voges-Proskauer test, and usually salicin fermentation. Once an organism had been identified as a *Salmonella*, the only metabolic test that would differentiate below the species level was utilisation of tartrates, as a negative result for this test would suggest *S. schottmuelleri* (Paratyphi B) and was 'seldom wrong' (Borman et al. 1943). Paratyphi B (which causes paratyphoid fever) and Java (gastroenteritis) share the same antigenic formula, 1,4,(5),12:Hb:1,2 and are still distinguished based on the fermentation of *d*-tartrate. This metabolic difference is believed to be due to a putative cation transporter, which is inactivated by a single nucleotide polymorphism in Paratyphi B (Han et al. 2006).

In the 1970s, efforts were being made to collect together multiple biochemical tests to allow rapid and relatively high throughput identification of clinical bacterial isolates (Lindberg et al. 1974). Numerous test kits of varying accuracy became available for the identification of the Enterobacteriaceae. The API system of twenty tests in a sterile plastic strip was 'found to be the most reliable', having a 99% correlation with standard biochemical tests and a 94% identification rate (Nord et al. 1974). Over 60 bacterial species can currently be identified with the API20E, with identification extending to the serovar level for Typhi and Paratyphi A among others.

1.5.2 Metabolic pathways

Early interest in the metabolism of *Salmonella* and other bacteria was towards the identification of compounds that could act as sole carbon and nitrogen sources and the effects these had upon growth rate (Richmond and Maaloe 1962). A comprehensive screen of ~600 substrates identified 76 carbon sources and 26 nitrogen sources for Typhimurium LT2 (Gutnick et al. 1969). Interestingly, growth was only observed on some of these compounds for mutant derivatives of the parent strain. This information provided a baseline from which studies could be performed to elucidate the mechanisms by which these substrates were utilised. For example, the uptake and degradation of the pentose sugar xylose was investigated and found to mirror the transport system and two catabolic enzymes present in *E. coli* and *Aerobacter* (*Enterobacter*) *aerogenes* (Shamanna and Sanderson 1979). However, utilisation of ethanolamine as a sole carbon and nitrogen source was found to be complicated by the requirement for the cofactor vitamin B_{12} (Roof and Roth 1988). Mutants unable to grow in the presence of ethanolamine and vitamin B_{12} were located to a cluster of genes between *purC* and *cysA* on the chromosome, termed the *eut* (ethanolamine utilisation) region (Roof and Roth 1988).

The majority of experiments on bacterial metabolism have been performed on the genetically tractable model bacterium *E. coli* K12. As a close relative, pathways in

Salmonella have generally been studied in detail only where they differ from *E. coli*. A case in point is the *de novo* synthesis of coenzyme B_{12}, genes for which are present in *S. enterica* serovars but not in *E. coli* (Roth et al. 1996).

1.5.3 Pathway maps

During the late 1990's, numerous metabolic databases became available, either as a general resource or to depict the metabolism of a particular organism. By using computational symbolic theory, the aim of collating a vast quantity of metabolic data into a single database was to give scientists a way to analyse and understand the complexity of biochemical reactions and pathways (Karp 2001).

1.5.3.1 Metabolic databases

One of the best-known general purpose databases is the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000; Ogata et al. 1999). KEGG is based upon three linked databases, describing genes, higher order biological functions (e.g. metabolic pathways) and chemical compounds. In the higher order database, metabolism is represented as 'reference' pathways which are visualised as a network of enzyme names or Enzyme Commission (EC) numbers. Updated daily, the reference metabolic pathways are based upon gene information taken from completely sequenced genomes. Given suitable enzyme annotation, organism-specific pathways can be generated based upon the reference pathway. The success of this process depends upon whether the pathway is conserved across multiple organisms. Species-specific variations are also difficult to determine using KEGG.

Other general purpose databases were created, such as the WIT (What Is There) and MPW (Metabolic Pathways Database) systems which formed part of another set of linked databases (Overbeek et al. 2000; Selkov et al. 1998). Similar to KEGG, but on a smaller scale, they comprised a collection of metabolic pathway reconstructions from sequenced genomes (Selkov et al. 1998). However, many of these are now inactive, including PUMA2 (Maltsev et al. 2006)(successor to WIT) and aMAZE (van Helden et al. 2001).

A tool that is still widely used, Cytoscape was developed to model biological interaction networks, and is especially powerful when there are high quality protein-protein, protein-DNA and genetic interaction data available for the organism of interest (Shannon et al. 2003). To this end, model organisms often comprise the best datasets.

EcoCyc is an example of a single organism metabolic database that uses experimental data to generate metabolic pathways (Karp et al. 1997; Keseler et al. 2009). It describes the metabolic capability of *E. coli* (K12) MG1655 and has been curated from the vast literature on this model organism. By concentrating on a single organism, EcoCyc has captured metabolic variations specific to MG1655 and provides links to the experimental evidence that generated these predictions. While EcoCyc remains the most highly curated, a number of other databases are now available from the BioCyc collection (http://www.biocyc.org). However, at the commencement of this thesis project, no serovar-specific *Salmonella* databases were publically available.

1.5.3.2 Pathway prediction

Software for pathway prediction can be broadly divided into two categories based upon the input data required. The first category requires only sequence data while the second requires both sequence data and genome annotation.

metaSHARK and the GEM system represent the first category (Arakawa et al. 2006; Pinney et al. 2005). metaSHARK is a fully automated system for detecting genes encoding enzymes and the subsequent visualisation of these within metabolic networks (Pinney et al. 2005). The novel bioinformatics in this software is almost entirely concentrated upon the gene prediction aspect. The metabolic networks are simply imported from KEGG, with some corrections, and used as a general framework for locating enzymatic reactions. The GEM (Genome-based modelling) system is slightly more oriented towards metabolic pathway prediction, with pathways predicted based upon enzyme function, and then checked against references in KEGG and BioCyc (Arakawa et al. 2006).

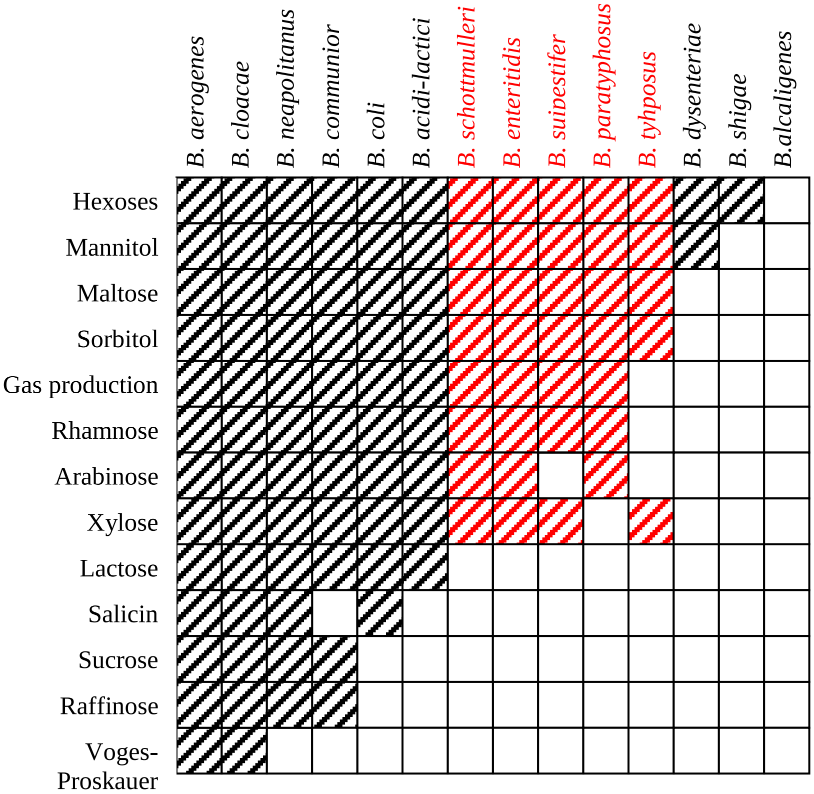
A member of the second category, Pathway Tools is the software developed by the bioinformaticians behind EcoCyc. This however requires genome annotation in order to predict the metabolic reactions and pathways encoded by the organism of interest. The reason stated by the authors is that manual efforts render a higher quality genome annotation than can be achieved by computation alone, and this should in turn produce a higher quality set of predicted pathways. Pathway Tools uses MetaCyc, “a database of non-redundant, experimentally elucidated metabolic pathways” (http://www.metacyc.org) as the reference for predicting pathways in the organism of interest. Like EcoCyc, MetaCyc is curated from the literature, but includes pathways from over 1,800 different organisms.

1.5.4 Network modelling

Once a set of metabolic pathways have been predicted, the next level of analysis comes in assessing how these interact with each other, as they must do in nature, to form a dynamic metabolic network. Often, metabolic networks are built *in silico* and tested and refined by analysing the biomass composition of the organism grown in batch culture (Novak and Loubiere 2000; Reed and Palsson 2003). Without the benefit of *Salmonella*-specific metabolic pathways, the only network analysis published thus far is based upon a large scale comparison with *E. coli* (AbuOun et al. 2009). As a step up from batch culture, the authors were able to make use of a metabolic phenotyping microarray that allowed the model to be tested for its ability to predict growth on over 250 substrates. Once this level of detail can be validated with *Salmonella*-specific information, this will provide a powerful tool for analysing *Salmonella* under different growth conditions.

1.6 Project focus

Despite their close genetic relatedness, it has been known for almost 100 years that *Salmonella* serovars display different metabolic phenotypes, varying for example in their fermentative ability. Figure 1-6 is reproduced from a paper published in 1919 showing metabolic phenotypes associated with organisms across a range of bacteria (Winslow et al. 1919). Of the five *Salmonella* depicted (in red), only two show the same metabolic phenotype across all 13 ‘fermentative relationships’, and Typhi appears to have a relatively restricted profile.



	B. aerogenes	*B. cloacae*	*B. neapolitanus*	*B. communior*	*B. coli*	*B. acidi-lactici*	*B. schottmulleri*	*B. enteritidis*	*B. suipestifer*	*B. paratyphosus*	*B. tyhposus*	*B. dysenteriae*	*B. shigae*	*B.alcaligenes*
Hexoses	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	
Mannitol	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨		
Maltose	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨			
Sorbitol	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨			
Gas production	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				
Rhamnose	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				
Arabinose	▨	▨	▨	▨	▨	▨	▨	▨		▨				
Xylose	▨	▨	▨	▨	▨	▨	▨	▨	▨		▨			
Lactose	▨	▨	▨	▨	▨	▨								
Salicin	▨	▨	▨		▨									
Sucrose	▨	▨	▨	▨										
Raffinose	▨	▨	▨	▨										
Voges-Proskauer	▨	▨												

Figure 1-6 Metabolic capabilities of *Salmonella* serovars

Figure adapted from (Winslow et al. 1919). Filled boxes represent bacterial ability to metabolise/ produce/react to substrate indicated. White boxes indicate absence of ability. *Salmonella* serovars are highlighted in red. Key: *B. schottmulleri*, Paratyphi B; *B. Enteritidis*, Enteritidis; *B. suipestifer*, Choleraesuis; *B. paratyphosus*, Paratyphi A; *B. typhosus*, Typhi.

While some metabolic phenotypes are absolute, others are more subtle. For example, hydrogen sulphide (H_2S) production is often used to identify particular salmonellae but variation in ability to reduce sulphate to sulphite between isolates of the same serovar and the test method used can alter the result. Thus, the tests used to differentiate salmonellae are not always stable and it is common to incorrectly identify a serovar. However, biochemical testing is still used globally and it performs well, to an extent. This suggests that there must be enough reproducible metabolic phenotypes to define some serovars. These phenotypes are of particular interest as they are likely to identify a large proportion of true serotype-specific metabolic capability.

The level of genome degradation in Typhi and Paratyphi A, coupled with evidence that these serovars have more restricted metabolic profiles than other serovars, has lead to the hypothesis of this thesis: that there is a causal link between host restriction and metabolic ability. However, at the beginning of this project there was little, if any, evidence in the literature. Whilst various complete *Salmonella* genome sequences were available, the lack of evidence was likely due in part to the paucity of information describing gene function and an inability to assign individual pseudogenes to metabolic pathways.

1.7 Aims of the project

Different *Salmonella* serovars adapted to different mammalian hosts contain unique but overlapping sets of pseudogenes. An understanding of the influence of these pseudogenes on the pathogenic and metabolic capability of the host-adapted serovars will help to explain the basic process of host-pathogen adaptation. This project explored the biology of host-restricted serovars of *Salmonella enterica* in comparison to the non-adapted Typhimurium, with the aims of:

- Creating metabolic pathway databases for Typhi and Typhimurium
- Assigning pseudogenes in host-restricted *Salmonella* to metabolic pathways
- Defining the essential gene lists of Typhi and Typhimurium
- Investigating the effect of macrophage invasion upon the metabolic networks of Typhi and Typhimurium