Chapter 4

Convergent evolution of Typhi and Paratyphi A

4.1 Introduction

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

be differences at the molecular level, it is assumed that the mechanisms of infection and transmission in Typhi and Paratyphi A are highly similar. It is proposed that this convergence in pathogenic phenotype should be reflected to a significant degree by convergence at the genetic level, which stands in contrast to the features that each serovar shares with their host-generalist relatives.

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

While host restriction and host adaptation are well known in *S. enterica*, the mechanisms underlying these phenomena are far from clear. The concept of "host adaptation" can be understood as the ability to circulate within a specific host population, transmitted from one member of the population to another (629). For example, Typhi

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

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

transmissibility as carriers remain infected and therefore infectious for a long time. The mechanisms by which these traits evolve are unclear, but presumably involve selection for acquisition, deletion and mutation of specific genes.

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

There are no known virulence genes unique to Typhi and Paratyphi A. Until recently SPI8 was thought to be unique to these serovars, but it is present in the genome sequence of serovar Agona (EMBL: NC_011149) and was recently detected in other serovars using a PCR screen (156). The Paratyphi A genome does not carry SPI7 and does not produce Vi, but it does carry a SopE-prophage similar to that present within Typhi SPI7 and other serovars. The Typhi and Paratyphi A genomes harbour a large number of pseudogenes (>4% of coding sequences in each genome) (46, 47, 49) compared to many host-generalist relatives such as *S. enterica* serovar Typhimurium (0.9%) or *E. coli* K-12 (1.2%). As discussed above, loss of gene function through pseudogene formation and gene deletion appears to be a hallmark of host restricted pathogenic bacteria compared to their host-generalist relatives (46, 49, 263, 264, 265, 266). The reason is uncertain, but is likely to be a combination of (a) adaptation, whereby the loss of certain proteins has a selective advantage in the new host, and (b) genetic drift, due to

population bottlenecks following host restriction and/or the absence of selective pressure to maintain certain functions that are no longer required in the new host. It has been reported that Paratyphi A and Typhi share some of their pseudogenes (49), resulting in convergent loss of protein functions which may be associated with adaptation to their shared niche. The first Paratyphi C genome sequence (strain RKS4594/SARB49) was recently published and showed similarly high levels of pseudogenes (3.3%), a few of which were shared with Paratyphi A and Typhi (93). A single Paratyphi B genome sequence is available, however the annotation is incomplete and does not include many pseudogenes, which are likely to be missed by automated annotation (EMBL:CP000886 as of June 1, 2009). Furthermore it is unclear whether the sequenced strain, SPB7, is of the systemic pathotype (negative for tartrate fermentation, but also *sopE*-negative (225)).

4.1.1 Aims

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

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

4.2 Methods

4.2.1 Whole genome comparisons

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

4.2.2 Phylogenetic network analysis

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

4.2.3 Bayesian analysis of recombined and non-recombined genes

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

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

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

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

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

Thus the final analysis was performed using the coalescent prior with constant population size, and relaxed clock with uncorrelated log-normal rate distribution. An additional 10 million iterations were run using this combination of settings for both

recombined and non-recombined gene sets, and results from 20 million iterations combined for each data set. Dates were calibrated using two reported ages of the split between *E. coli* and *Salmonella*: 140 million years and 70 million years. The 140 million year estimate was based on comparison of DNA encoding 16S RNA between *E. coli* and *S. enterica* serovar Typhimurium (42). Later studies, based on protein alignments for glutamine synthetase (44) and other proteins (45) across the Bacterial and even other kingdoms, have yielded much lower estimates in the range 70-114 years.

4.2.4 Time estimation using dS

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

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

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

The numbers of synonymous SNP sites (S) included in the analyses of Paratyphi A (Chapter 3) and Typhi (Chapter 2) were 1.27 million and 1.35 million respectively. The mean number of synonymous SNPs accumulated in each genome since the most recent common ancestor ($\frac{1}{n}\sum_{i=1}^{n} s_i$) were 31 for Paratyphi A (s_i range 21-37) and 89 for Typhi (s_i range 71-102). The mutation rate for synonymous sites was estimated previously at $3.4\text{x}10^{-9}$, based on a divergence date for *Salmonella* and *E. coli* of 140 million years ago (2, 41, 42, 43). Using the lower estimate of 70 million years, the

upper bound on this rate would be 6.8x10^{-9}. The alignments of recombined and non-recombined genes described above were used to generate a novel rate estimate and to provide an alternative measure of divergence times. Pairwise synonymous site divergence ($dS_{i,j}$ for genomes i and j) was calculated using the method of Yang and Nielsen (650) implemented in the `yn00` algorithm of the software package PAML (651). Since pairwise dS incorporates evolution on both branches since the mrca ($dS_{i,j} = dS_i + dS_j$), estimates were divided by 2 before use in equation 4.1. The mean $dS_{i,j}$ between *E. coli* and *Salmonella* sequences was 0.8, providing a novel molecular clock rate estimate of 0.8/2/140,000,000 = 2.8x10^{-9} per site per year, or 5.6x10^{-9} using the alternative calibration time of 70 million years.

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

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

4.2.5 Comparison and annotation of pseudogenes

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

4.2.6 Data simulation

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

4.3 Results

4.3.1 Evolution of Typhi and Paratyphi A

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

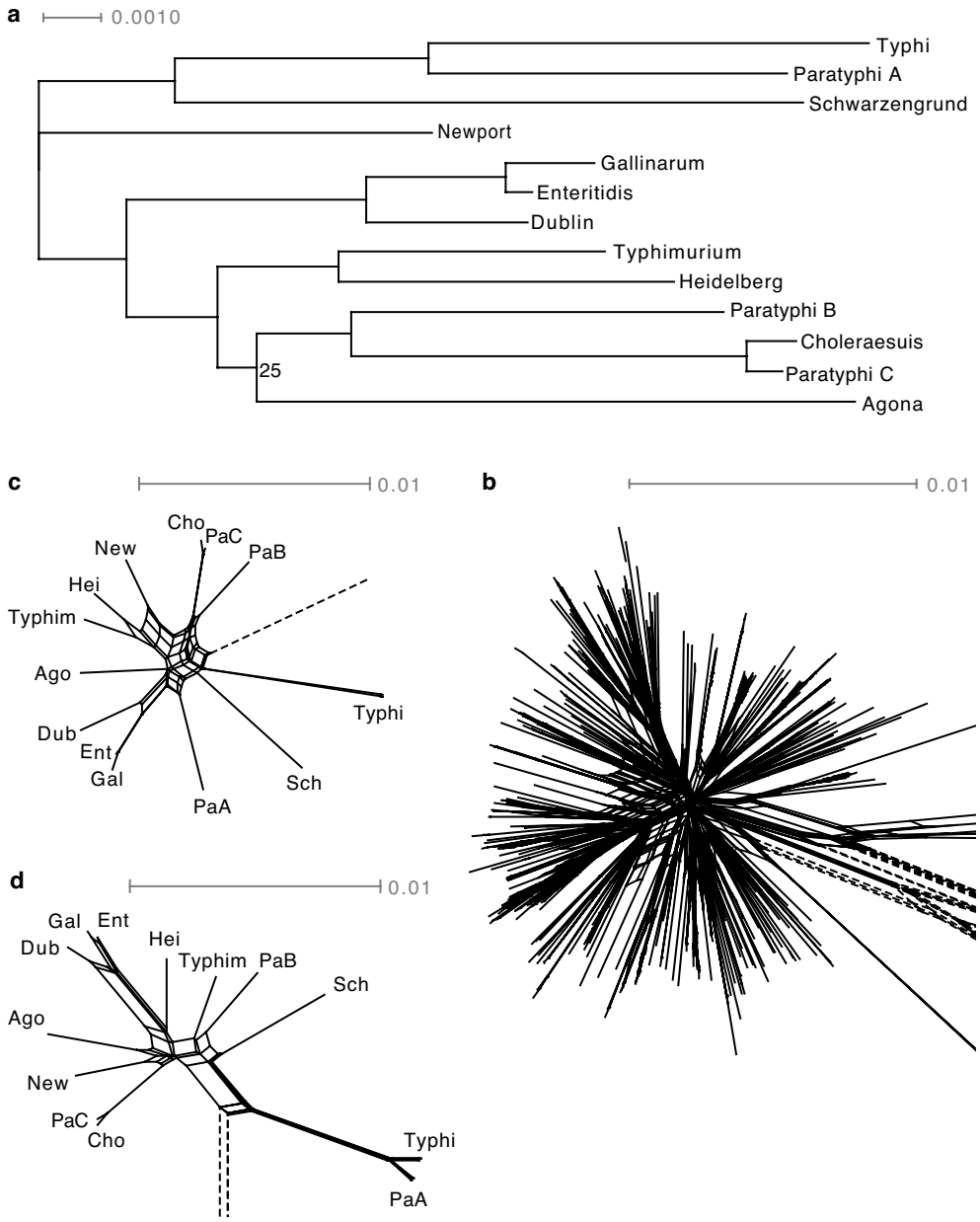

Figure 4.1: Phylogenetic trees for *Salmonella enterica* - Scale bars show substitutions per nucleotide. Dashed lines indicate where other *Salmonella* species join the networks. (a) Maximum likelihood phylogenetic tree of *S. enterica* based on whole genome alignment. The tree shown is the best fit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

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

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

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

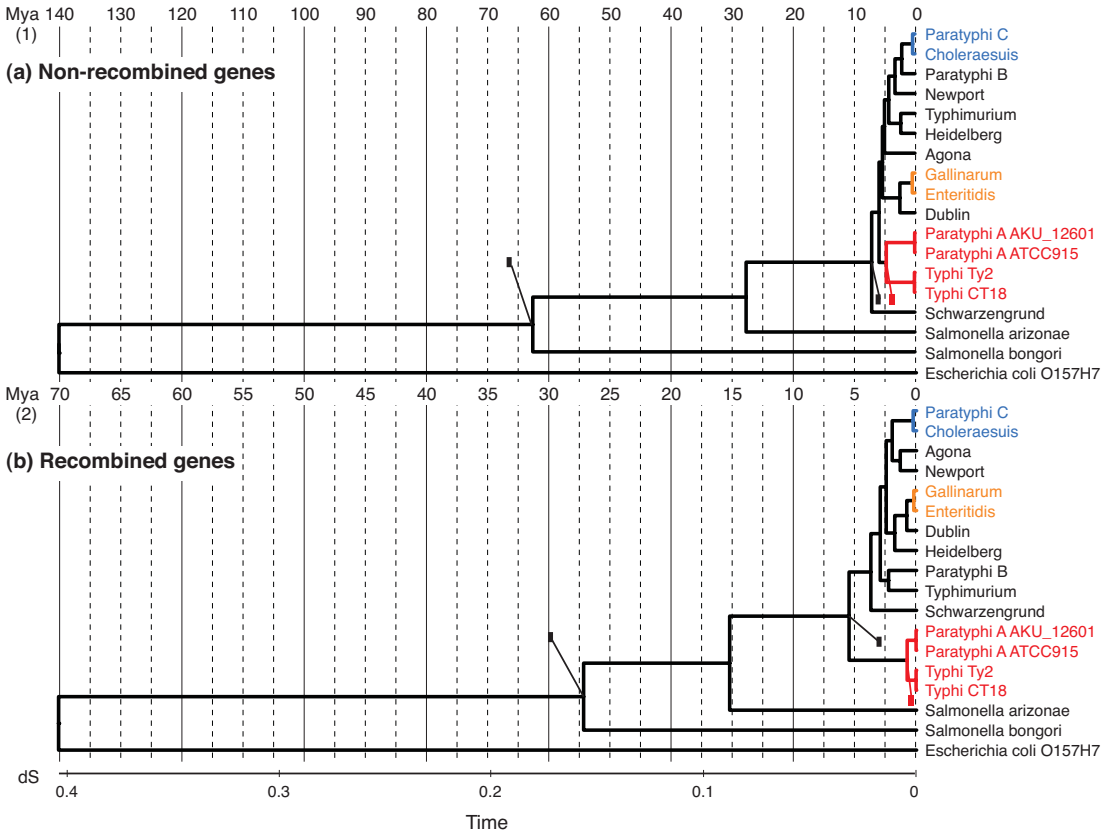

Figure 4.2: Phylogenetic trees for *Salmonella* and *E. coli* with divergence time estimates - Bayesian analysis was used to construct phylogenetic trees using (a) seven genes that were not recombined between Typhi and Paratyphi A and (b) seven genes that were recombined between them. Bayesian phylogenetic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

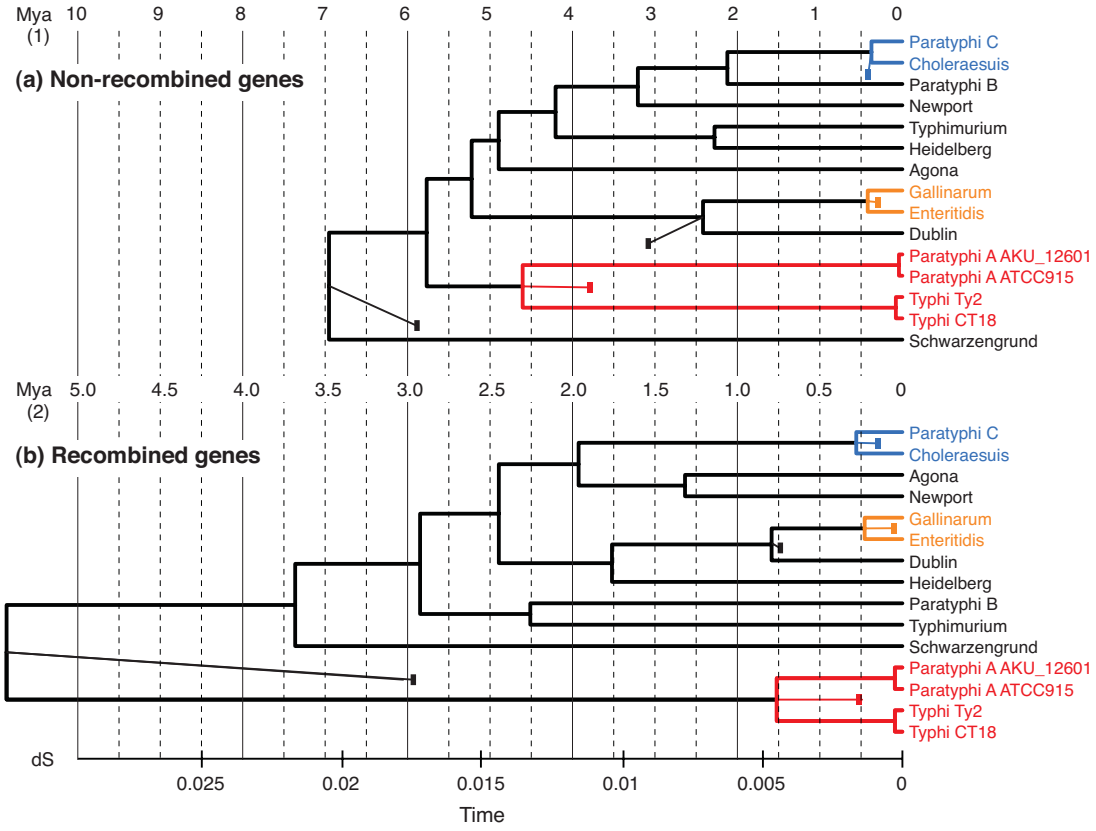

Figure 4.3: Phylogenetic trees for *S. enterica* with divergence time estimates - Zoom in on *Salmonella enterica* from the trees shown in Figure 4.2. Bayesian analysis was used to construct phylogenetic trees using (a) seven genes that were not recombined between Typhi and Paratyphi A and (b) seven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

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

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

4.3.2 Convergent features of the Typhi and Paratyphi A genomes

4.3.2.1 Shared genes

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

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

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

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

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

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

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

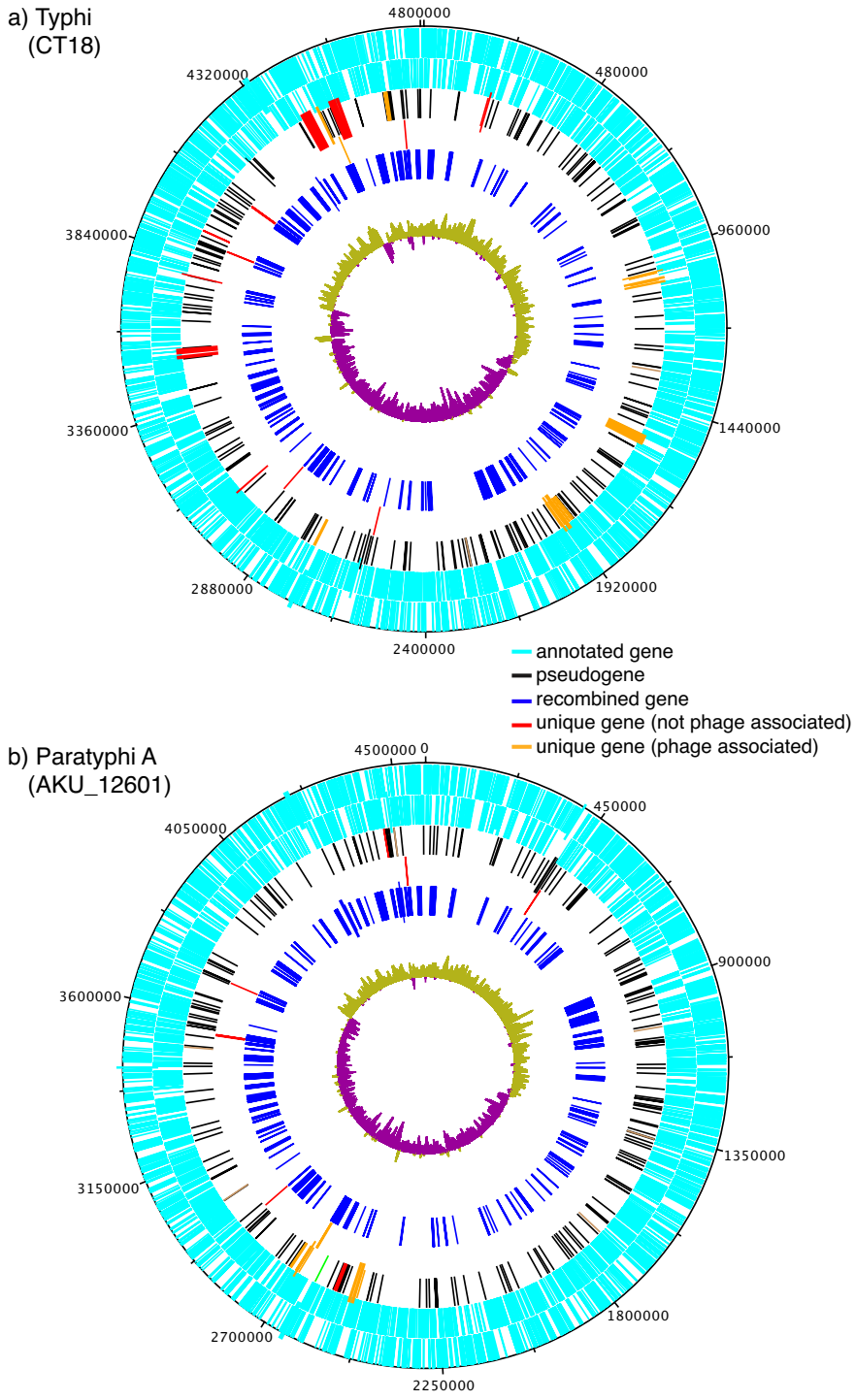

Figure 4.4: Pseudogenes, recombined genes, and unique genes in the Typhi and Paratyphi A genomes - Rings from outside: 1, 2 = coding sequences on forward, reverse strands; 3 = pseudogenes and genes unique to the serovar; 4 = genes present in both Typhi and Paratyphi A but absent from 11 other sequenced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

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

4.3.2.2 Comparison of pseudogenes in Typhi and Paratyphi A

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

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

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

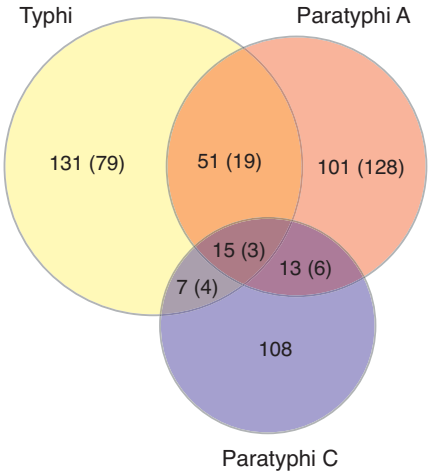

Figure 4.5: Overlap of pseudogenes in Typhi, Paratyphi A and Paratyphi C - Figures indicate the number of pseudogenes that are present in each serovar or combination of serovars; strain-specific pseudogenes are shown in brackets.

4.3.2.3 Genes missing from Typhi and Paratyphi A

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

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

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

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

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

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

4.3.2.4 Features shared with Paratyphi C

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

4.3.3 The role of recombination

4.3.3.1 Sharing of unique genes and deletions by recombination

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

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

recombination contributed to the sharing of unique genes between Typhi and Paratyphi A, most of the genes deleted from both genomes were the result of independent events.

4.3.3.2 Sharing of pseudogenes by recombination

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

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

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

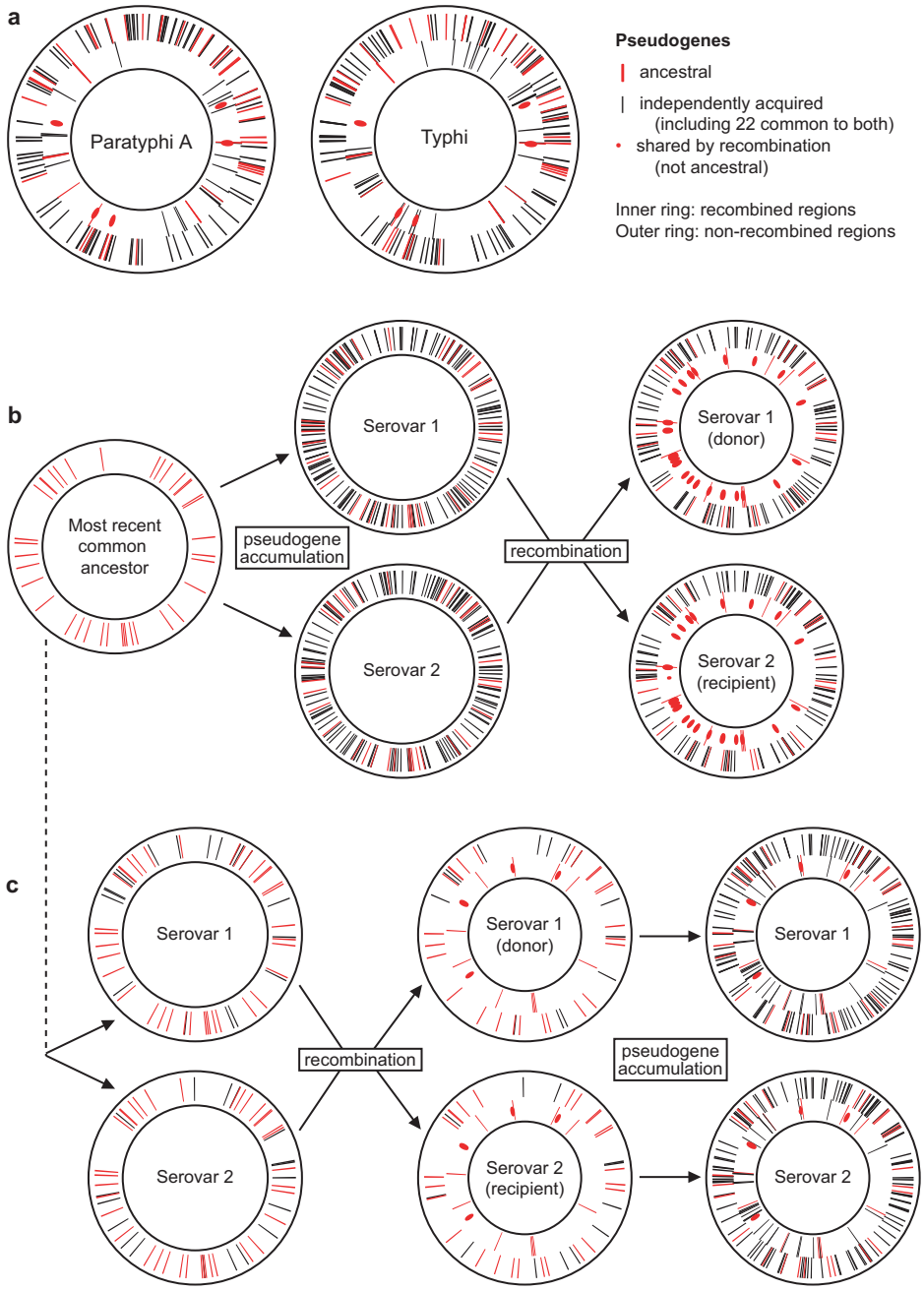

Figure 4.6: Scenarios of recombination and pseudogene formation in Paratyphi A and Typhi - (a) True distribution of pseudogenes in the Paratyphi A AKU_12601 and Typhi CT18 genomes (gene order based on gene coordinates in Typhi CT18). (b-c) Distribution of pseudogenes resulting from data simulated under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

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

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

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

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

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

4.3.4.1 Pseudogene formation over time

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

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

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

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

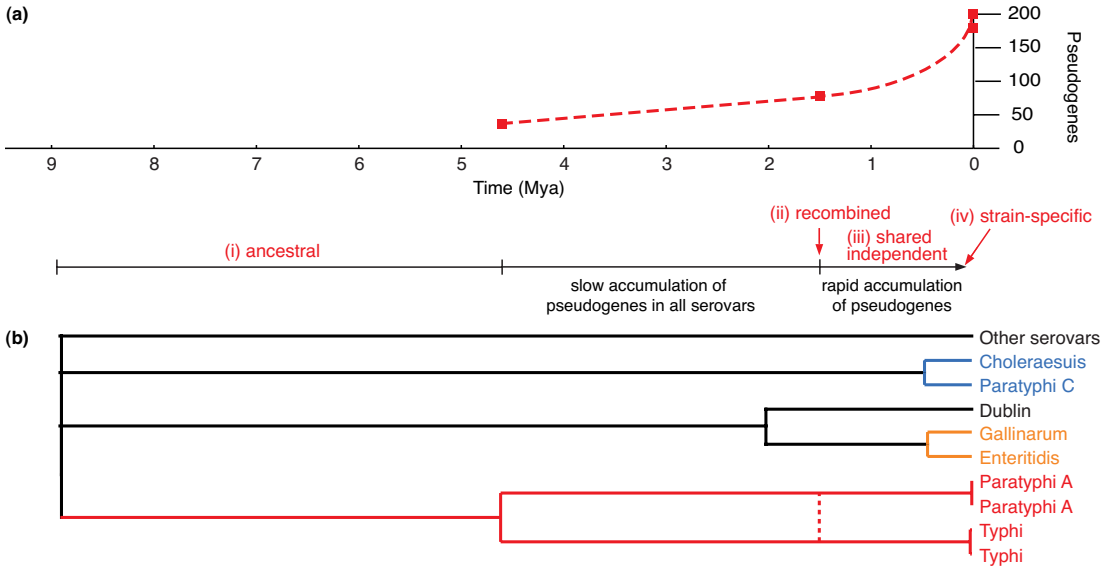

Figure 4.7: Pseudogene accumulation in Typhi and Paratyphi A over time - (a) Estimated number of pseudogenes in Typhi and Paratyphi A over time, with points corresponding to nodes in the phylogenetic tree. The number at the root is unknown; the number at the point of divergence, 39, is based on the analysis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

4.3.4.2 Pseudogenes potentially involved in host adaptation

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

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

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

Three known secreted effector proteins were inactivated in both Typhi and Paratyphi A: *sopD2*, *sopA* and *slrP* (Table 4.7). *SopD2* is assumed to have been inherited in inactive form by both Typhi and Paratyphi A, as the same mutation (a 2 bp insertion) is present in both serovars yet the gene sequence was most likely not shared by recombination (divergence 1.7%). The inactive form of *sopA* was most likely shared via recombination, as the two gene sequences are only 0.2% divergent and contain the same nonsense SNP. *SlrP* on the other hand was inactivated by multiple independent mutations in Paratyphi A and Typhi, and is unlikely to have been recombined (divergence 1.9%). Only two

other secreted effector proteins were inactivated in either Typhi (*sopE2*) or Paratyphi A (*sifB*). Interestingly, *sopA* and *slrP* were also inactivated in the Paratyphi C genome, by independent mutations to those found in Typhi or Paratyphi A. No other secreted effectors were identified as pseudogenes in Paratyphi C. Thus the inactivation of *sopA* and *slrP* may be a prequisite for systemic infection of the human host, whereas intact products of most other secreted effector genes may be essential for this kind of infection.

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

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

have been needed to respond to particular environmental conditions or stimuli that are not encountered in the human restricted niche.

4.4 Discussion

4.4.1 Strengths and limitations of the study

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

Interpretation of the functional impact of genetic convergence in the form of shared pseudogenes, deletions and unique genes is difficult for a number of reasons. It would be optimal to study shared features of all human adapted serovars relative to other genomes, and to correlate features that are shared within subsets of human adapted serovars with particular pathogenic features. However, the availability of genome sequences is still lacking, as is characterisation of clinical and molecular features of the various serovars. There are currently no genome sequences available for systemic infection-associated Paratyphi B or Sendai, and neither of these have been extensively studied in terms of disease progression and clinical outcomes. Although a Paratyphi C genome was recently published (93), the disease syndrome caused by this serovar is not well understood. This is partly due to the difficulty of differentiating Paratyphi C from other serotype O6,7:c:1,5 *S. enterica*, including Choleraesuis, which can only be done definitively using a range of biochemical tests (622) or a combination of *IS*200

typing and ribotyping (658). For example, isolation of Paratyphi C from animals has been reported (659) and several isolates from pigs, typed as Paratyphi C, have been submitted to the *S. enterica* MLST database (464). However biochemical typing of Paratyphi C isolates in the database found that the pig isolates were not Paratyphi C, and indeed all 'real' Paratyphi C isolates in the database came from human infections and formed a single clonal group by MLST (Satheesh Nair, Sanger Institute/Health Protection Agency, personal communication, May 2009). The difficulty of accurately typing Paratyphi C has probably had a limiting effect on the reporting of infection with this serovar. There are currently very few reports of Paratyphi C infection in the literature, with most studies dating back to the 1930s-1980s in British Guyana (624) or Africa (619, 660). A 1933 study reported that Paratyphi C infection in humans was milder than infection with Typhi and Paratyphi A, and caused severe disease only in patients also infected with malaria (624). However confirmation of this, and studies of the clinical features of Paratyphi C infection, is lacking. Functional interpretations from comparative genomic analyses are also hampered by the lack of information on gene functions and pathways in *Salmonella* and in particular during human infection, which is experimentally intractable. However, genomic studies should help to develop hypotheses regarding gene function which can be experimentally tested in animals and human tissue.

This study benefits from historical clues about the relationship between Typhi and Paratyphi A, in particular the rapid burst of recombination 0.25-1.5 million years ago revealed by comparative genome analysis (56). This event serves as a historical marker (see Figures 4.2, 4.3 and 4.7), affording an additional insight into the temporal dynamics of gene degradation in Typhi and Paratyphi A. The population variation data, first presented in Chapters 2 and 3 of this thesis, add an additional historic marker in the form of the most recent common ancestor of each serovar. This allows the distinction to be made between gene degradation events that occurred prior to the last population bottleneck of each serovar and are therefore fixed in the populations, and those that have occurred much more recently. Together, these historical markers reveal that the rate of pseudogene accumulation in both Typhi and Paratyphi A increased dramatically following recombination between them (Figure 4.7). The majority of these became fixed in the last <25,000 years, although pseudogenes continue to accumulate in both

serovars. Note though that the rate of this recent, strain-specific accumulation is likely to be underestimated by available variation data, which does not include small indel mutations (frameshifts) which are a common cause of gene inactivation.

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

In the absence of external calibration, bacterial divergence estimates usually rely on the idea of the 'molecular clock', which assumes that DNA substitutions become fixed at a constant or clock-like rate. However it is clear that different genes and different lineages evolve at different rates (666, 667, 668, 669). Furthermore there are numerous studies

pointing to time-dependency of substitution rates (526, 670, 671, 672), suggesting that the rate over short time scales is dramatically higher (at least an order of magnitude) than that over long time scales. It has been suggested that this phenomenon is associated with a number of factors including purifying selection, genetic drift and effective population size (526, 670, 671, 672). The Bayesian analysis presented here incorporates a relaxed clock model, which allows substitution rates to vary on different branches of the phylogenetic tree. However, in the absence of independent information for dating coalescent events within the phylogenetic tree of *Salmonella*, this is unlikely to result in adequate correction for time-dependent rate variation. A rough *post hoc* adjustment for the effect of time dependency might be to compress the most recent events into a shorter period, which would make the accumulation of pseudogenes in Typhi and Paratyphi A more, rather than less, dramatic. It would also result in a downwards revision of the age of Typhi and Paratyphi A, so that the estimates given here should be considered upper bounds.

4.4.2 Implications for host restriction and adaptation

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

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

4.4.2.1 Ancestral pseudogenes

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

4.4.2.2 Pseudogenes and novel genes shared by recombination

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

4.4.2.3 Recent pseudogenes: convergence after recombination

In addition to >100 pseudogenes specific to each serovar, group (iii) includes 22 shared pseudogenes containing different inactivating mutations in Paratyphi A and Typhi (Table 4.3). While it is possible that some of those lying outside recombined regions may

have been present prior to recombination, it is likely that most of these mutations arose in the period of rapid pseudogene accumulation after recombination. These pseudogenes are examples of convergent gene loss through independent mutation, and are therefore good candidates for involvement in adaptation to the human host. They include only one transposase gene, the remainder being genes of known or putative function, many of which have been implicated in host interactions in serovar Typhimurium (e.g. *fhuA, fhuE, shdA, ratB, sivH, slrP*) (49, 683).

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

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

4.4.2.4 Ongoing accumulation of strain-specific pseudogenes

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