Chapter 2

Genomic sequence variation in Typhi

2.1 Introduction

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

tions/deletions (259), although it remains unclear whether the rearrangements result in significant functional changes, for example in gene expression.

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

While it provided many novel insights, the study by Roumagnac *et al.* (2) was limited to 88,739 bp, just 1.85% of the Typhi genome, and despite including 105 isolates yielded only 82 SNPs. It was clear therefore that to discover the true extent of SNPs and other variations in Typhi, and discover novel variations providing increased resolution in typing, would require comparative analysis at the whole-genome level. Recent advances in sequencing technology make this feasible using 454 and Solexa sequencing

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

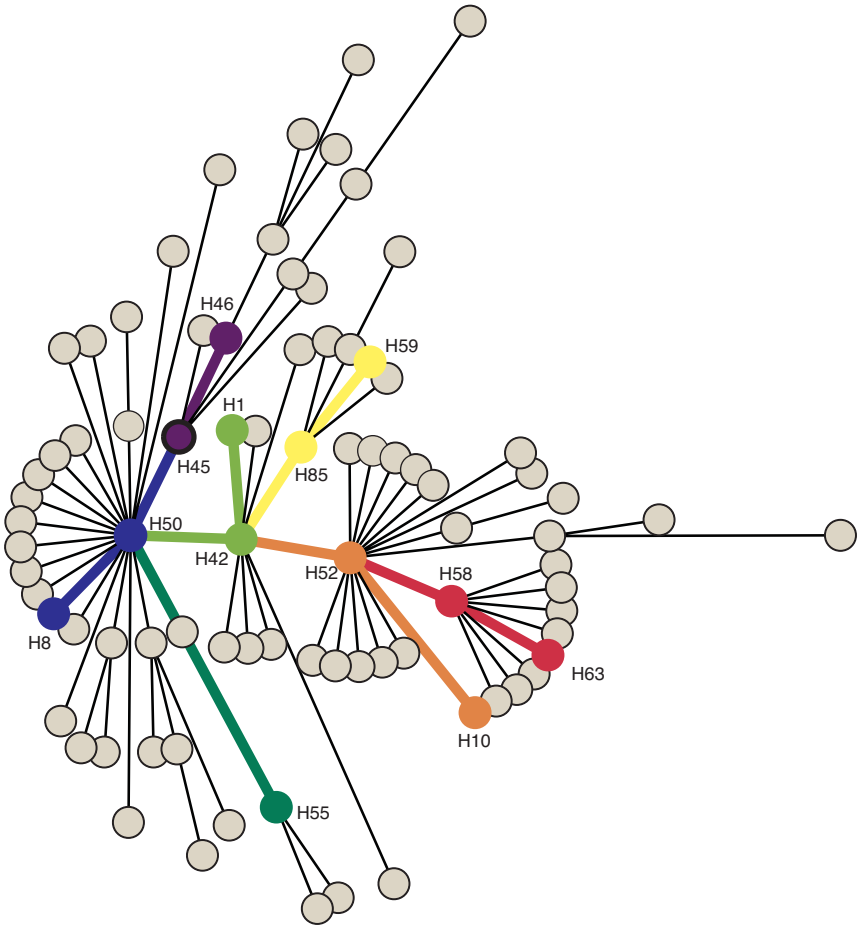

Figure 2.1: Phylogenetic tree guiding selection of Typhi isolates for sequencing - Phylogenetic tree defined previously from analysis of 97 SNPs in 481 Typhi isolates (2); root is H45. Haplotypes from which isolates were chosen for sequencing, and the branches joining them, are coloured according to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

2.1.1 Aims

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

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

2.2 Methods

2.2.1 Bacterial strains and DNA

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

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

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

2.2.2 DNA sequencing

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

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

2.2.3 Plasmid identification

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

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

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

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

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

2.2.4 Phylogenetic analysis

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

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

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

2.2.6 Transition bias

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

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

where:

i = synonymous, nonsynonymous or intergenic,

j = transition or transversion,

m = codon,

$n_{i,j,m}$ = number of mutations of type i, j possible starting from codon m;

N_m = number of times codon m appears in Typhi CT18 (excluding repeat regions).

Transition bias in each class was calculated as follows:

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

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

where:

$s_{i,j}$ = number of SNPs observed of type i, j,

n_i = number of possible mutations type i, that is $\sum_{j=1}^{2} n_{i,j}$,

$$p_i = \frac{s_{i,ts}}{n_i},$$

$$se_i = \sqrt{\frac{p_i(1-p_i)}{n_i}}.$$

2.2.7 Detection of recombination events

Didelot *et al.*(56) found the distribution of gene-wise divergence between Typhi and its closest relatives (other serovars of *S. enterica*) predominantly followed a normal distribution with mean 1%, and the authors used divergence below 0.3% as a cut-off for identifying potential recombination events with other serovars. In order to identify poten-

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

2.2.8 Evidence of expression from published microarray data

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

2.2.9 Accession codes

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

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

2.3 Results

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

2.3.1 Assessment of SNP detection methods

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

At 100 bp on average, 454 reads were long enough to be assembled *de novo* into contigs (i.e. without reference to any other sequence). Thus two approaches were available for the detection of SNPs from 454 data: (i) align reads directly to a reference sequence, or (ii) assemble reads into contigs and align contigs to a reference sequence. To determine the best method for analysing 454 data, SNP detection error rates were calculated using real and simulated reads, and real and simulated contigs. Reads were aligned to a reference using ssahaSNP (http://www.sanger.ac.uk/Software/analysis/ssahaSNP/). Contigs were aligned to a reference using MUMmer (v3.19, `nucmer` algorithm) (576). These free, opensource software packages combine alignment with SNP detection, and provide flexible parseable output which facilitates high-throughput analysis. Unlike ssahaSNP, most other software specifically developed for detecting SNPs from sequencing reads utilise raw fluorescence data generated by capillary sequencing (e.g. SNPdetector, NovoSNP, PolyBayes), which is not appropriate for analysing data generated by 454 pyrosequencing. 454 provide their own SNP detection algorithm within their mapped assembly software, however this proprietary software is something of a 'black

box', and appears to miss a large number of true SNPs (e.g. it detected on average 40 SNPs per Typhi genome compared to CT18, whereas the results of all other analyses reported here suggest 5-10 times this level of variation). MUMmer is also unique in that it facilitates fast and free whole-genome comparison combined with SNP detection; other options include diffseq (EMBOSS package, (577)) which can't handle genome rearrangements, Mauve (578) which does not include SNP detection, and the proprietary software package Kodon (Bionumerics).

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

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

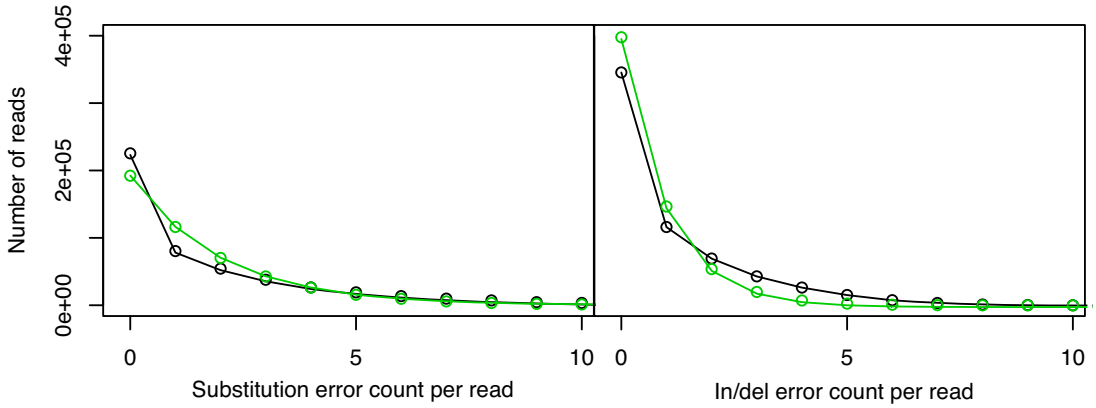

Figure 2.2: Error models for 454 reads - Real frequencies of errors in 454 reads from *Streptococcus suis* P_17 are shown in black; these were determined by mapping reads to the finished sequence. Note *S. suis* data was used because both 454 data and finished sequence was available for the same isolate;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

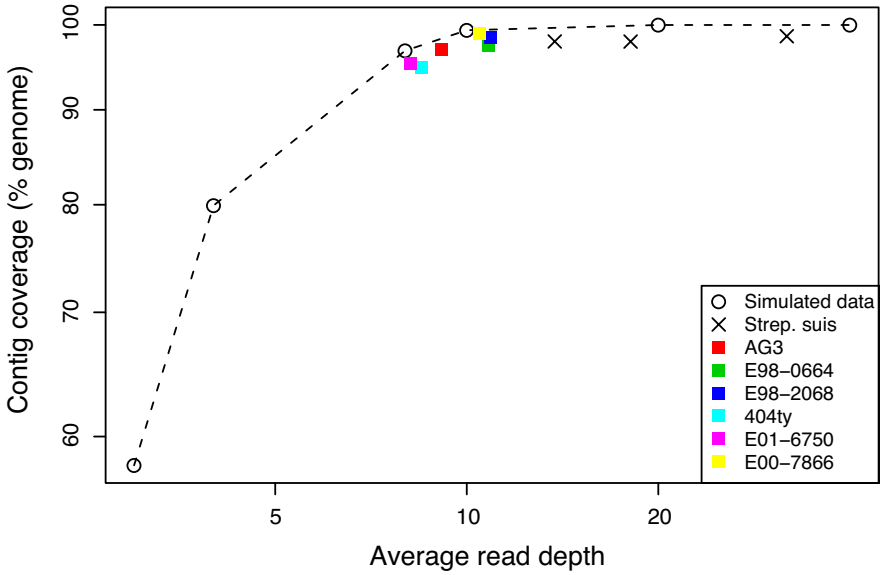

Figure 2.3: Read depth vs genome coverage for real and simulated 454 data - Coloured squares represent real data from Typhi strains.

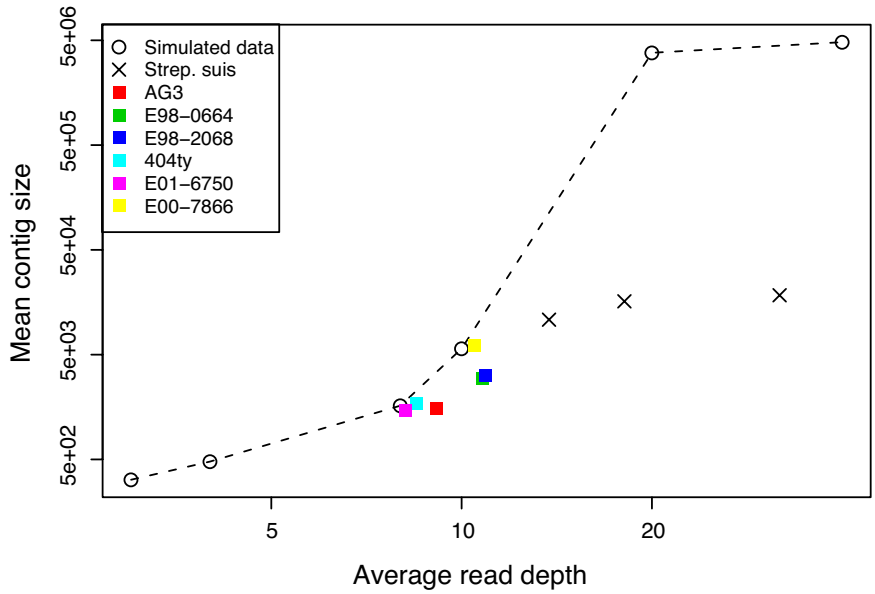

Figure 2.4: Read depth vs mean contig size for real and simulated 454 data - Coloured squares represent real data from Typhi strains.

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

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

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

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

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

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

reads simulated at 8-10x depth (Table 2.3), the former explanation appears more likely. Thus for the remainder of this study, SNP detection from 454 data was performed by analysing assembled contig sequences rather than reads.

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

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

2.3.1.2 Solexa data

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

2.3.1.3 Determining quality filters for SNP detection

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

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

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

2.3.1.4 Estimating error rates for SNP detection

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

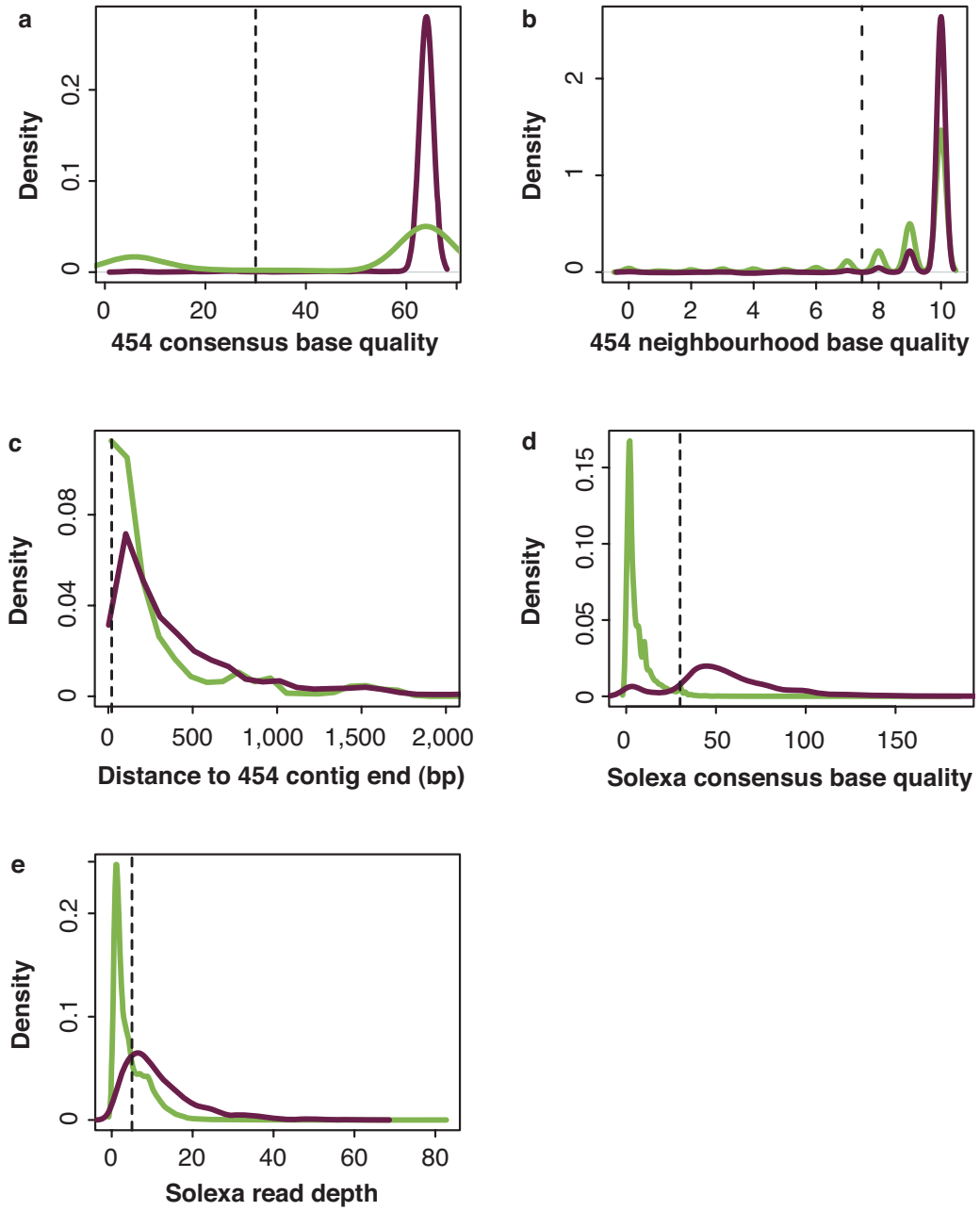

Figure 2.5: Distribution of quality parameters for SNP detection - Purple lines show distribution of each parameter among SNPs called in both 454 and Solexa data, pooled for three isolates (AG3, E98-3139, 404ty). Green lines show distribution of each parameter among SNPs called in only 454 or Solexa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

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

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

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

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

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

2.3.1.5 Minimisation of potential errors

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

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

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

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

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

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

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

2.3.2 SNP analysis

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

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

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

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

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

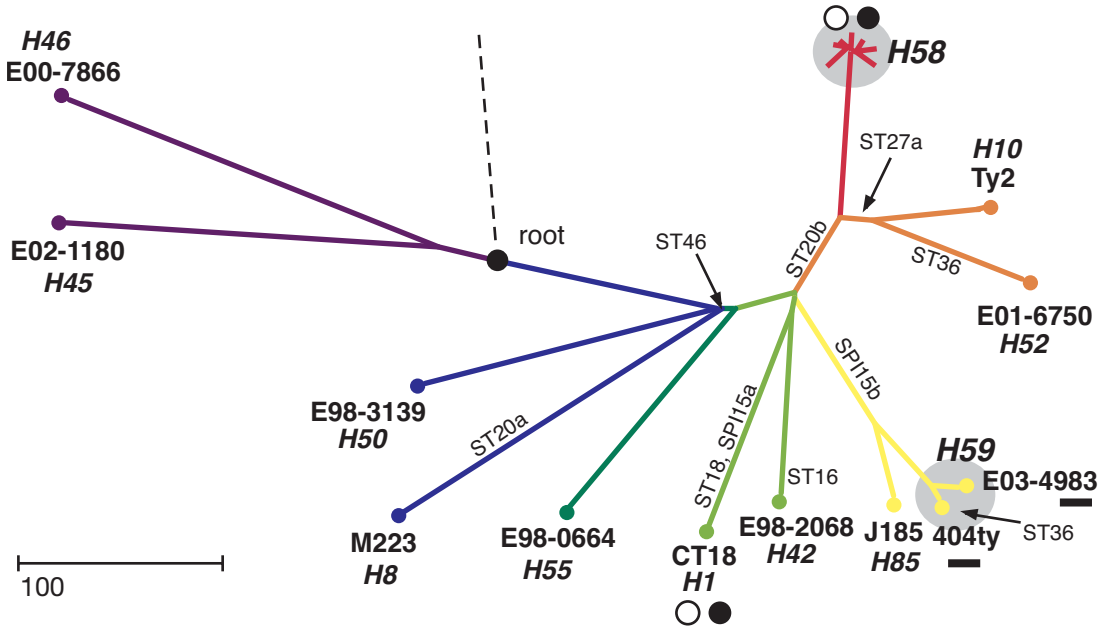

Figure 2.6: Phylogenetic tree of Typhi based on SNP data - Branch colours and lengths are consistent with Figure 2.11; branch lengths are measured in number of SNPs, scale as indicated. Black circle indicates the ancestral root, dashed line represents the link to other *Salmonella*; phage (ST) and SPI15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

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

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

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

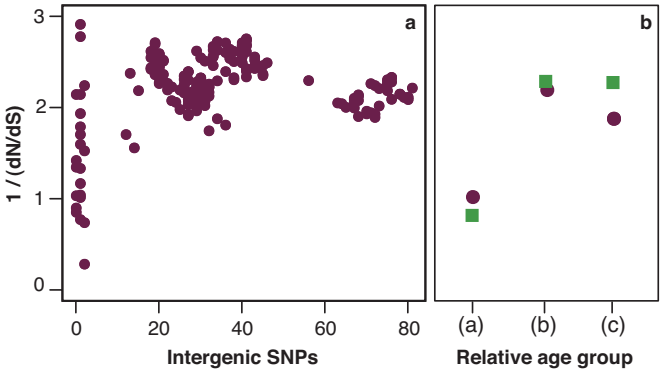

Figure 2.7: Trajectory of $\frac{dN}{dS}$ over time in Typhi - Y-axis is the reciprocal of $\frac{dN}{dS}$, or $1/\frac{dN}{dS}$. (a) Pairwise $1/\frac{dN}{dS}$ between 19 Typhi isolates vs pairwise number of intergenic SNPs. (b) $1/\frac{dN}{dS}$ for SNPs in three relative age groups (a=youngest, c=oldest),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

2.3.2.2 Potential signals of selection

There was very little evidence of adaptive selection in Typhi genes, which would be represented by an overabundance of nonsynonymous SNPs or independent changes in the same or nearby amino acid residues. Nearly three quarters (72%) of Typhi genes contained no SNPs and the distribution of SNPs per gene followed a Poisson distribution in the range 0-6 SNPs per gene, shown in Figure 2.8. However, there were a few exceptions, listed in Table 2.9. Three genes (*yehU*, *tviE* and STY2875) contained more than six SNPs, which deviates from the Poisson model. STY2875 is an exceptionally large gene (3,625 bp compared to the genome mean of 910 bp), which may account for the number of SNPs. However *yehU* and *tviE* are small (562-579 bp) and thus the high

number of SNPs may be evidence of diversifying selection in these genes, the second of which is encoded in SPI7 and involved with Vi synthesis (92). Ten SNPs did not fit the phylogenetic tree, which may indicate either recombination or convergent evolution, whereby the same mutation arose independently in different lineages (note however that independent SNP typing (Chapter 6) suggested that two of these SNPs, indicated in Table 2.9, were not actually homoplasic). If the latter explanation is true it would suggest the possibility of adaptive selection at these sites, which include nonsynonymous SNPs in two membrane proteins (STY1204 and *yadG*) and two nonsynonymous SNPs in *gyrA* that are known to increase resistance to fluoroquinolones, the class of antibiotics currently recommended for treatment of typhoid fever (2, 3, 397). Fifteen genes contained clusters of nonsynonymous SNPs, whereby two residues within five amino acids were mutated, which may indicate adaptive selection in localised regions of the encoded protein (Table 2.9).

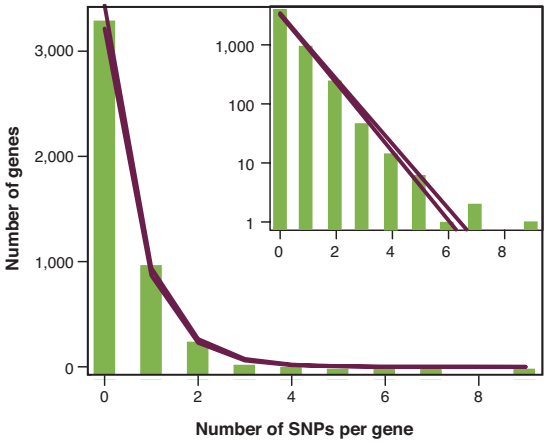

Figure 2.8: Distribution of number of SNPs per Typhi gene - Lines indicate 95% confidence interval of mean predicted values under a Poisson distribution fitted to the data shown in green. Inset shows gene count on a log scale to better show deviation from the Poisson model at high numbers of SNPs per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

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

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

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

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

2.3.2.3 Recombination

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

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

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

2.3.3 Gene acquisition

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

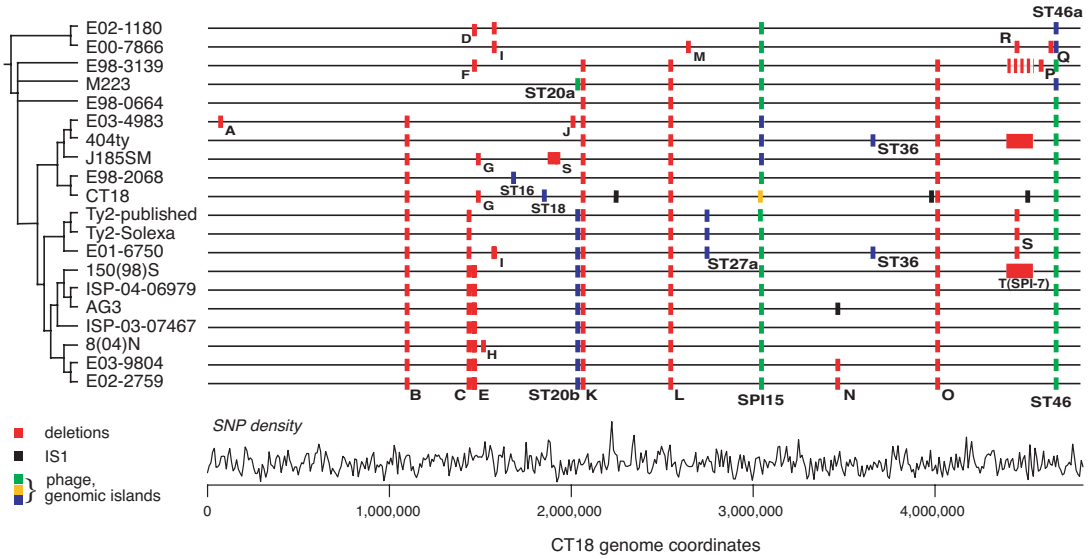

Figure 2.9: Distribution of prophage, IS elements and deletions in the Typhi genome and phylogenetic tree - The phylogenetic tree based on SNPs is shown on the left, the distribution of these SNPs in the genome is shown at the bottom. The genomic positions of deletions, prophage and IS insertions are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

2.3.3.1 Prophage sequences

CT18 harbours seven well defined prophage-like elements (Figure 1.4) (46, 224) and while some of these were conserved in all sequenced isolates, several novel phage were

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

2.3.3.2 Genomic islands

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

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

2.3.3.3 Plasmids

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

2.3.4 Loss of gene function

2.3.4.1 Genomic deletions

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

total deletion of 42 Typhi genes. The distribution of deletions is shown in Table 2.11 and illustrated in Figure 2.9.

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

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

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

In addition SPI7, which harbours genes required for synthesis of the polysaccharide Vi capsule (92) was missing from 404ty and 150(98)S. The isolate E98-3139 appeared to be a mixed population in regards to SPI7 as its coverage in both 454 and Solexa reads was approximately 25% that of genomic coverage (see Figure 2.10). Note that

the low mapping coverage in this region is most likely due to deletion of SPI7 rather than replacement with a similar island, as deletion is known to occur during culture (517, 568) and no alternative island could be assembled from 454 reads. No other SPIs were deleted from the sequenced Typhi, indicating they are relatively stable in the genome (although we observed three variants of the 6 kbp SPI15 as described above).

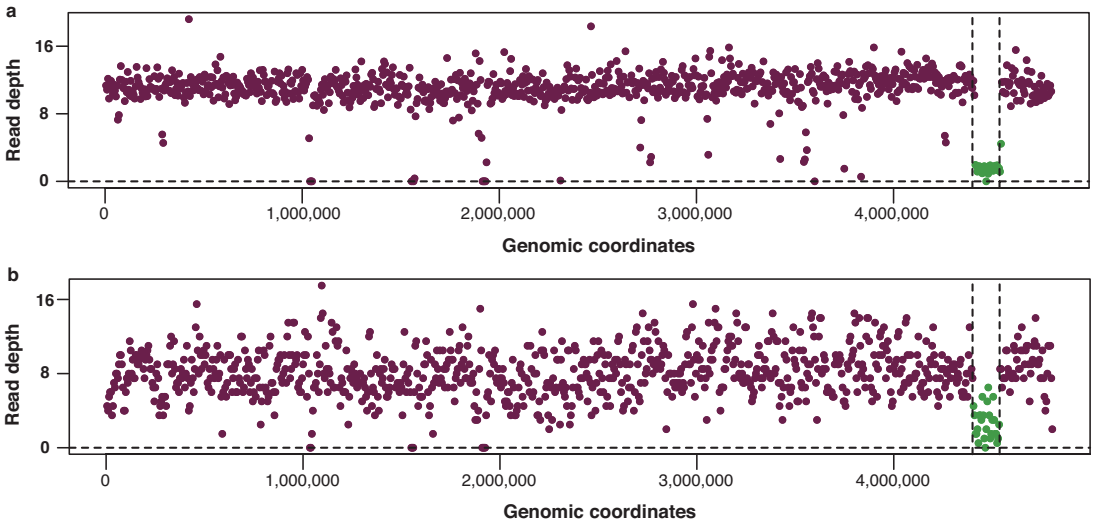

Figure 2.10: Coverage of SPI7 in Typhi isolate E98-3139 - Genome positions correspond to CT18 genomic coordinates, SPI7 is located between the two vertical dashed lines; coverage in this region is highlighted in green; horizontal dashed line shows zero coverage. (a) Read depth in 454 data. (b) Read depth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

2.3.4.2 Accumulation of pseudogenes

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

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

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

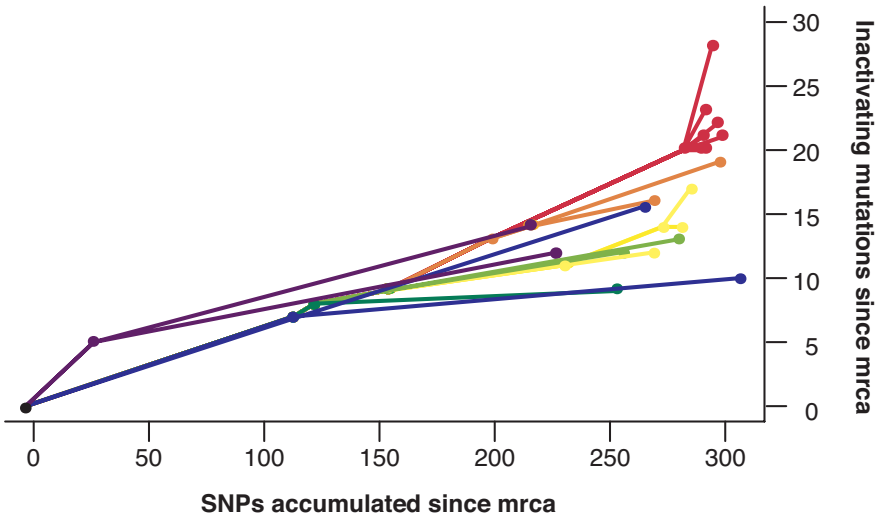

Figure 2.11: Accumulation of gene-inactivating mutations in Typhi lineages - Points correspond to bifurcations in the phylogenetic tree in Figure 2.6, y-axis shows the total number of genes inactivated by deletion or nonsense mutation up to that bifurcation. Each line represents the accumulation of mutations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

2.4 Discussion

2.4.1 Strengths and limitations of the study

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

For this study, the analysis of 454 data was of central concern, as this platform was used to sequence representative isolates from central and radial haplotype nodes thus forming the basis of phylogenetic analysis, as well as providing evidence of insertion and deletion events via assembled sequence. Analysis of simulated and experimental Typhi data suggested that analysis of assembled contigs provided more accurate SNP detec-

tion than direct analysis of 454 reads (Tables 2.3, 2.4; 2.3.1.1). This is likely due to the fact that the 454 *de novo* assembler Newbler 1.0.5.3 operates in flow space, meaning it assembles raw pyrosequencing signal data directly into consensus base calls as opposed to converting signals to base calls in individual reads and assembling the resulting read sequences. Thus a base call in the assembled contigs represents the consensus of multiple flow signals, which is expected to be more accurate than the collection of base calls in individual reads, and so contigs assembled in this way are expected to be more reliable for correctly identifying SNPs (Guido Kopal, Roche, personal communication, February 2007). Note however that both 454 data generation and the Newbler assembly software have changed since this study was completed, and so analysis of new data will require a reassessment of SNP detection approaches that take into account the platform and software in use at the time.

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

An independent test of false negative rates is much more difficult. Data simulation suggested this should be below 10% at the read depths used in this study (2.3), although

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

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

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

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

done about this using current data and software, this problem will become increasingly tractable using Solexa sequencing either to identify and correct errors in 454 data, or to detect small indels directly by alignment of reads to reference (587).

2.4.2 Differences between Typhi lineages

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

It is difficult to estimate how much of the underlying variation in the Typhi population has been captured in this sequencing study. However we do now have a much better picture of how much variation can be expected between two distinct Typhi lineages. Figure 2.12 shows the distribution of the number of SNPs detected between pairs of the 12 Typhi lineages shown in Figure 2.6 (using 404ty to represent the 404ty/E03-4983 lineage and AG3 to represent the H58 lineage). There are peaks at ∼300 SNPs and ∼550 SNPs, reflecting the very early divergence of E00-7866 (H46)/E02-1180 (H45) from the other lineages. Note roughly half of SNPs are nonsynonymous, thus lineages

differed by an average of 150 nonsynonymous SNPs. Any two lineages differed by an average of 5 deletions (range 0-15) and 1 or 2 prophage (range 0-5), the distributions are shown in Figure 2.13a-b. However prophage insertions tended to be specific to individual lineages, while deletions were more frequently conserved (see Figures 2.6 and 2.9, Table 2.11). Thus the deletions identified in this study would make good genetic markers whereas the prophage insertions would not. There is likely to be some functional variation between lineages due to nonsynonymous SNPs (in particular nonsense SNP)s, deletions and small indels. Lineages differed by an average of 150 nonsynonymous SNPs and four pseudogenes differentially inactivated by nonsense SNPs and deletions (Figure 2.13). Note that the true difference in pseudogene complement is likely to be larger, due to frameshifts which could not be detected from this data; more than 75% of pseudogenes that differed between CT18 and Ty2 were due to frameshifts, thus it is likely that the mean variation between lineages is more in the order of 12 pseudogenes than four. The effect of these mutations is unknown, however it is likely that they do contribute to some phenotypic differences. It is also possible that prophage variation contributes to differences in genetic function between Typhi lineages, although no obvious phage cargo genes were identified in this study. The variation observed in SPI15 may also contribute to phenotypic variation between lineages, although the function of its cargo remains a mystery.

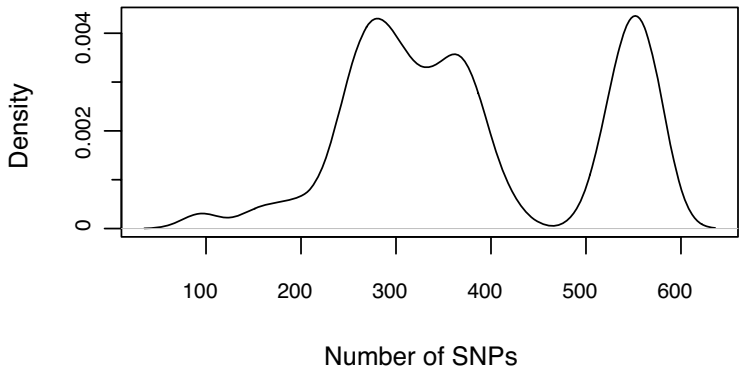

Figure 2.12: Distribution of number of SNPs between pairs of Typhi lineages - The number of SNPs between every possible pair of 12 Typhi lineages was calculated (using AG3 to represent the H58 lineage and 404ty to represent the H59 lineage), the distribution of SNP numbers between pairs is shown.

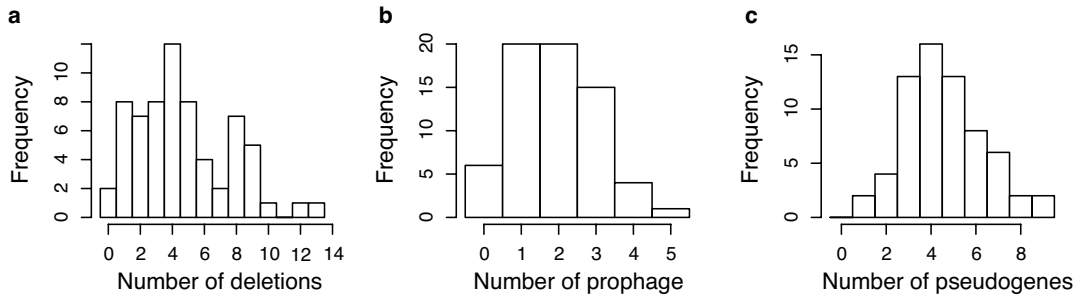

Figure 2.13: Distribution of number of deletions, prophage and pseudogenes between pairs of Typhi lineages - The number of deletions, prophage insertions and pseudogenes that differed between every possible pair of 12 Typhi lineages was calculated (using AG3 to represent the H58 lineage and 404ty to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

2.4.2.1 Antibiotic resistance and the H58 lineage

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

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

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

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

In this study 106 H58-specific SNPs were detected, including 57 nonsynonymous substitutions. Three of these introduce stop codons within genes encoding lipoprotein B precursor *rlpB* (STY0698), DNA-binding protein *stpA* (STY3001) and ATP-dependent RNA helicase *dbpA* (STY1410). The *stpA* gene also contains a SNP introducing a novel stop codon in E98-3139, which also harbours the GyrA-Phe83 mutation. StpA is a homolog of the transcriptional repressor H-NS involved in repressing transcription of the porin gene *ompS1* and possibly many other Typhi genes (588, 589). The occurrence

of independent inactivating mutations in *stpA*, on different haplotype backgrounds, suggests that this gene may be subject to negative selection. One hypothesis is that inactivation of *stpA* enables cells to better tolerate the mutations in GyrA, perhaps by the suppression of some response normally induced by StpA. This could be investigated by looking for evidence of an association between *stpA* inactivation and the GyrA-Phe83 mutation on different haplotype backgrounds and by selection on media containing nalidixic acid. In addition, the H58 isolates shared a deletion affecting the aminotransferase gene STY1507 and hypothetical gene STY1509, and they shared a SNP within the RNA gene *csrB* which regulates activity of the carbon storage regulator *csrA* (STY2947) (590). CsrA regulates the expression of SPI1 genes and genes secreted via the SPI1-encoded TTSS in Typhimurium (591), so the mutation in *csrB* may have an effect on virulence.

2.4.3 Insights into the evolution of Typhi

2.4.3.1 Adaptive selection in Typhi genes

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

The lack of evidence for adaptive selection in general is in contrast with the known adaptive selection for mutations in *gyrA* associated with fluoroquinolone resistance. The signal of selection in *gyrA* was detected in the present study as clustered, homoplasic nonsynonymous SNPs in neighbouring codons 83 and 87. Two other genes contained homoplasic nonsynonymous SNPs (Table 2.9), one of which (*yadG*) is the membrane component of an efflux protein in *E. coli* (592) and may therefore be associated with

antibiotic resistance in Typhi (efflux proteins can act as pumps to remove antibiotics from the bacterial cell (593)). However, no genes besides *gyrA* contained multiple homoplasic SNPs and few contain multiple nonsynonymous SNPs at all, consistent with the hypothesis of genetic drift in the Typhi genome. The adaptive mutations evident in the *gyrA* gene highlight the strong selective pressure on the Typhi genome associated with antibiotic use in the human population. This is not particularly surprising, as the fitness advantage associated with increased antibiotic resistance is likely to be very strong. However the lack of similar evidence for other adaptive mutations suggests that Typhi is under relatively little selective pressure from its host or the environment in general.

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

as yet unidentified mechanisms of generating antigenic diversity or that prophage genes, which were excluded from SNP analysis in this study, may play a role.

2.4.3.2 Evolutionary dynamics of the Typhi population

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

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

estimated population size (2) and the persistence of distinct haplotypes through time and geographical space (all nodes of the Typhi phylogenetic tree shown in Figure 2.1 were detected among a set of less than 500 extant Typhi isolates (2), indicating that the Typhi population is not shaped by clonal replacement).

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