Chapter 3

Genomic sequence variation in Paratyphi A

3.1 Introduction

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

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

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

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

3.1.1 Aims

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

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

3.2 Methods

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

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

3.2.2 SNP detection

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

3.2.3 Phylogenetic network analysis

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

3.2.4 Detection of insertion/deletion events and plasmid sequences

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

3.2.5 Gene ontology analysis

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

3.2.6 Accession codes

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

3.3 Results

3.3.1 Comparison of seven Paratyphi A genome sequences

3.3.1.1 Whole genome sequencing

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

3.3.1.2 SNP analysis

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

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

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

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

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

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

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

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

3.3.1.3 Gene acquisition

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

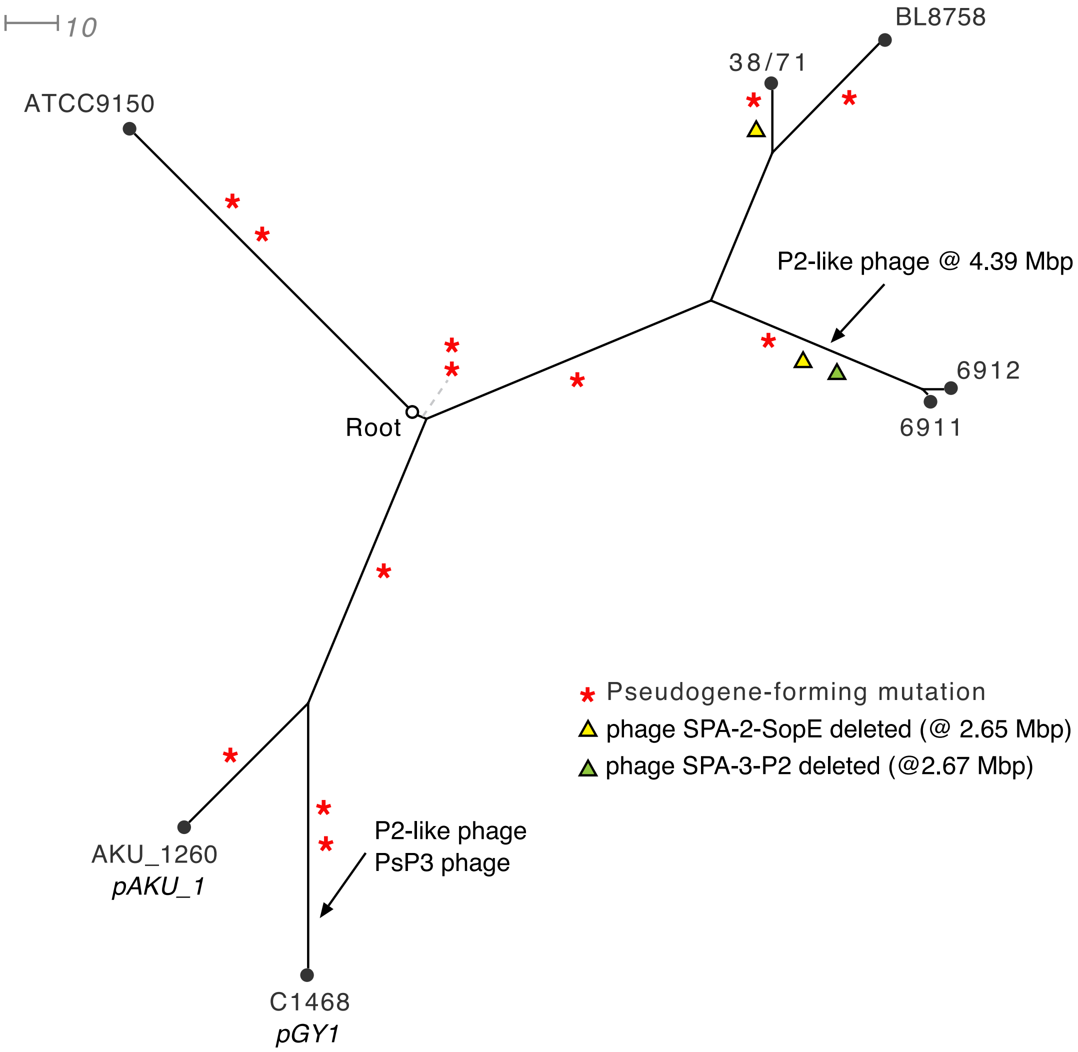
10
ATCC9150
38/71
BL8758
P2-like phage @ 4.39 Mbp
6912
6911
Root
* Pseudogene-forming mutation
△ phage SPA-2-SopE deleted (@ 2.65 Mbp)
△ phage SPA-3-P2 deleted (@2.67 Mbp)
P2-like phage
PsP3 phage
AKU_1260
pAKU_1
C1468
pGY1

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

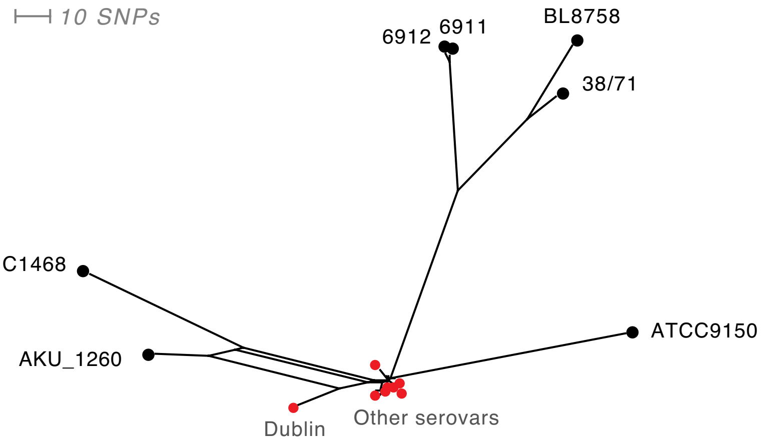

Figure 3.2: Phylogenetic network of seven Paratyphi A isolates including seven serovars as outgroups. - Serovars Typhi, Typhimurium, Enteritidis, Paratyphi B, Choleraesuis, Dublin, Galinarum and Pullorum were used as outgroups (red nodes).

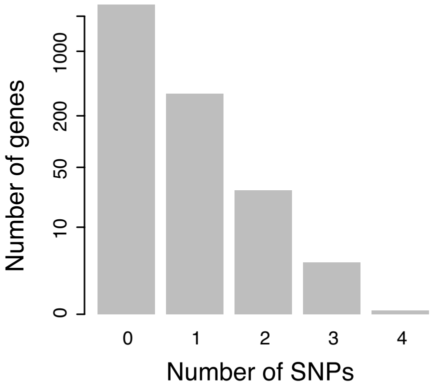

Figure 3.3: Distribution of SNPs per Paratyphi A gene - Note the y-axis, number of genes, is on a natural logarithmic scale.

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

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

3.3.1.4 Insertion/deletion mutations

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

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

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

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

3.3.1.5 Loss of gene function

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

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

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

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

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

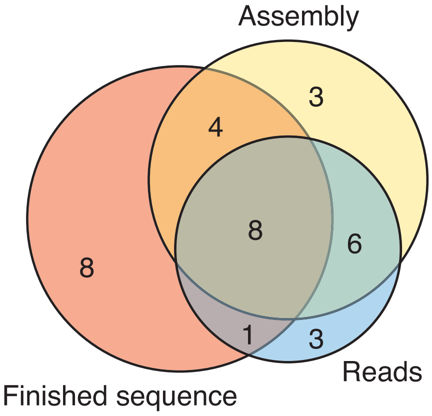

Figure 3.4: Detection of small indels from short read data - Indels of <20 bp detected from finished sequence (pink) compared to those detected from alignment of *de novo* assembled contigs (yellow) or reads (blue) to the reference.

3.3.2 Optimisation of SNP detection from pooled sequence data

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

3.3.2.1 SNP detection and frequency estimation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.3.2.2 Comparison of potential methods

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

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

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

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

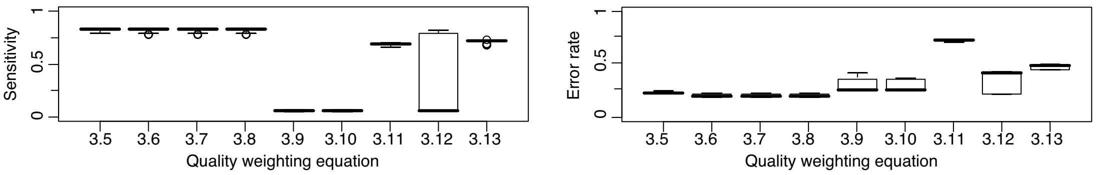

Figure 3.5: Sensitivity and error rates for different weighting measures - The distribution of each measure is shown as a box-and-whisker plot where black bars indicate median values, boxes indicate interquartile ranges, and whiskers indicate the full range of values observed using each weighting equation.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

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

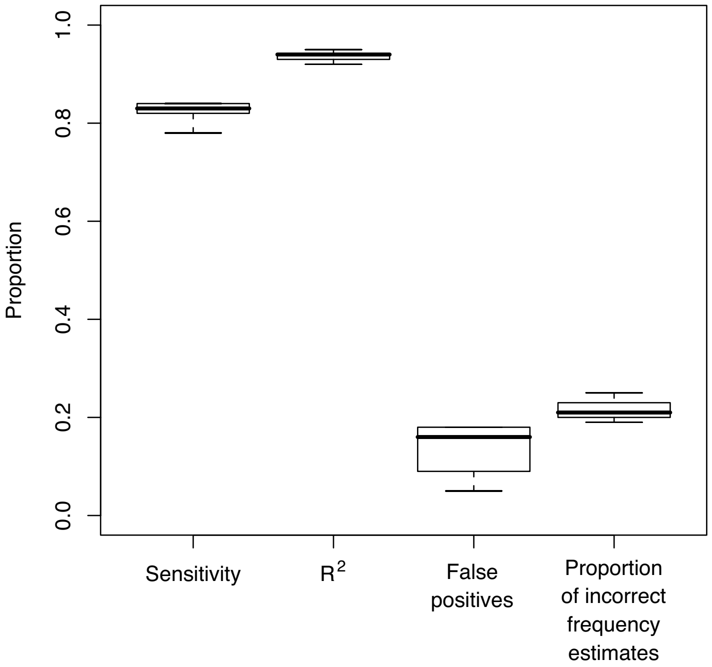

Figure 3.6: Ranges for each accuracy measure - The distribution of each measure is shown as a box-and-whisker plot where black bars indicate median values, boxes indicate interquartile ranges, and whiskers indicate the full range of values observed. Each measure was calculated for every combination of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

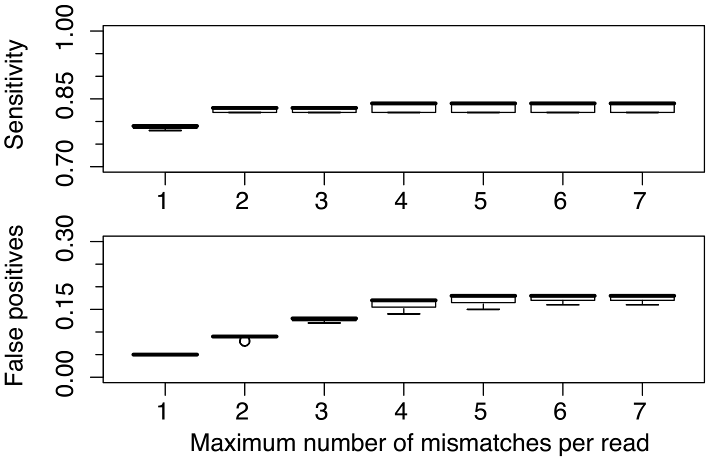

Figure 3.7: Sensitivity and false positive rates for different assembly parameters - The distribution of each measure is shown as a box-and-whisker plot where black bars indicate median values, boxes indicate interquartile ranges, and whiskers indicate the full range of values observed. Each measure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

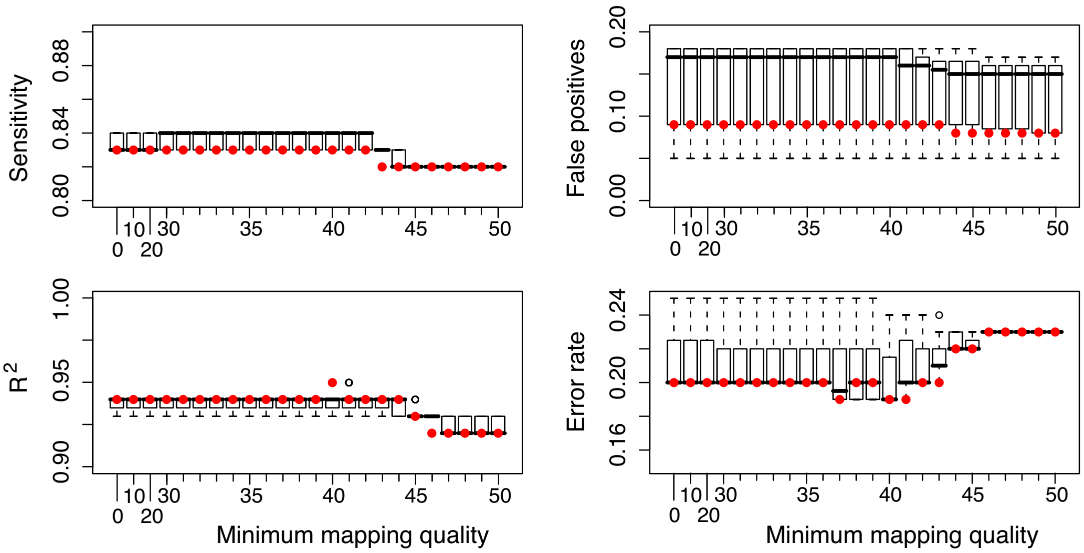

Figure 3.8: Accuracy measures for different pileup parameters - The distribution of each measure is shown as a box-and-whisker plot where black bars indicate median values, boxes indicate interquartile ranges, and whiskers indicate the full range of values observed. Each measure was calculated for all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

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

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

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

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

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

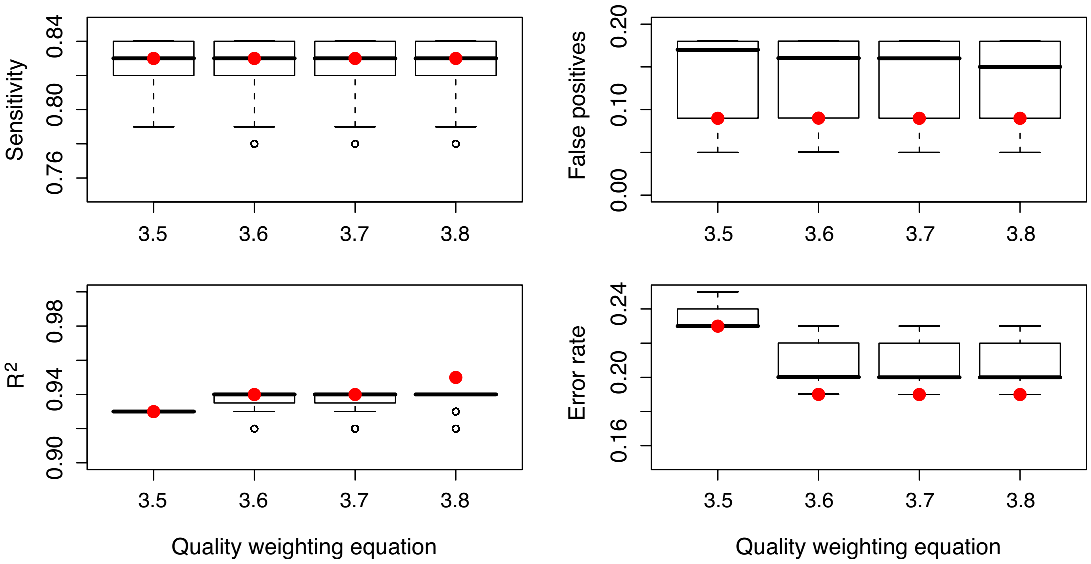

Figure 3.9: Accuracy measures for different weighting equations - The distribution of each measure is shown as a box-and-whisker plot where black bars indicate median values, boxes indicate interquartile ranges, and whiskers indicate the full range of values observed. Each measure was calculated for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

3.3.2.3 Performance of optimised method

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

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

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

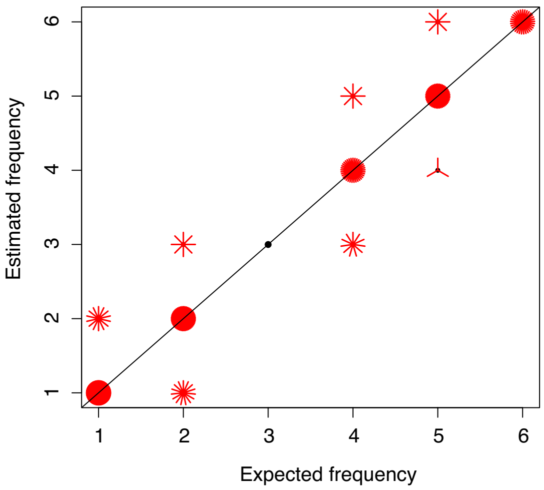

Figure 3.10: Expected frequencies vs SNP frequencies estimated from Paratyphi A test pool sequence data - The sunflower plot is designed to aid visualisation of correlations between discrete variables. It is similar to an x-y plot, but uses radial red lines to indicate the number of data points that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

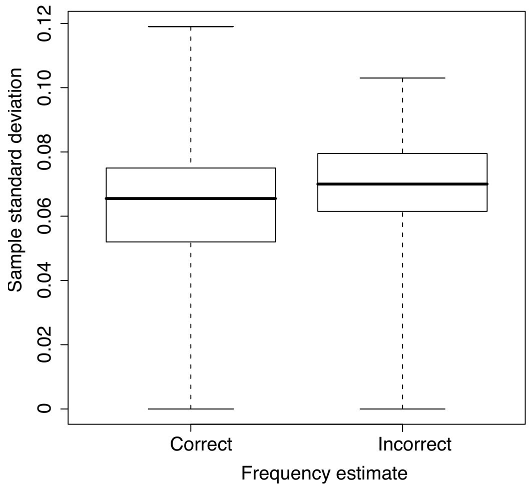

Figure 3.11: Distributions of sample standard deviations calculated among SNPs with correct and incorrect frequency estimates - Black bars indicate median values, boxes indicate interquartile ranges, and whiskers indicate the full range of values observed.

3.3.2.4 Performance of optimised method over a range of read depths

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

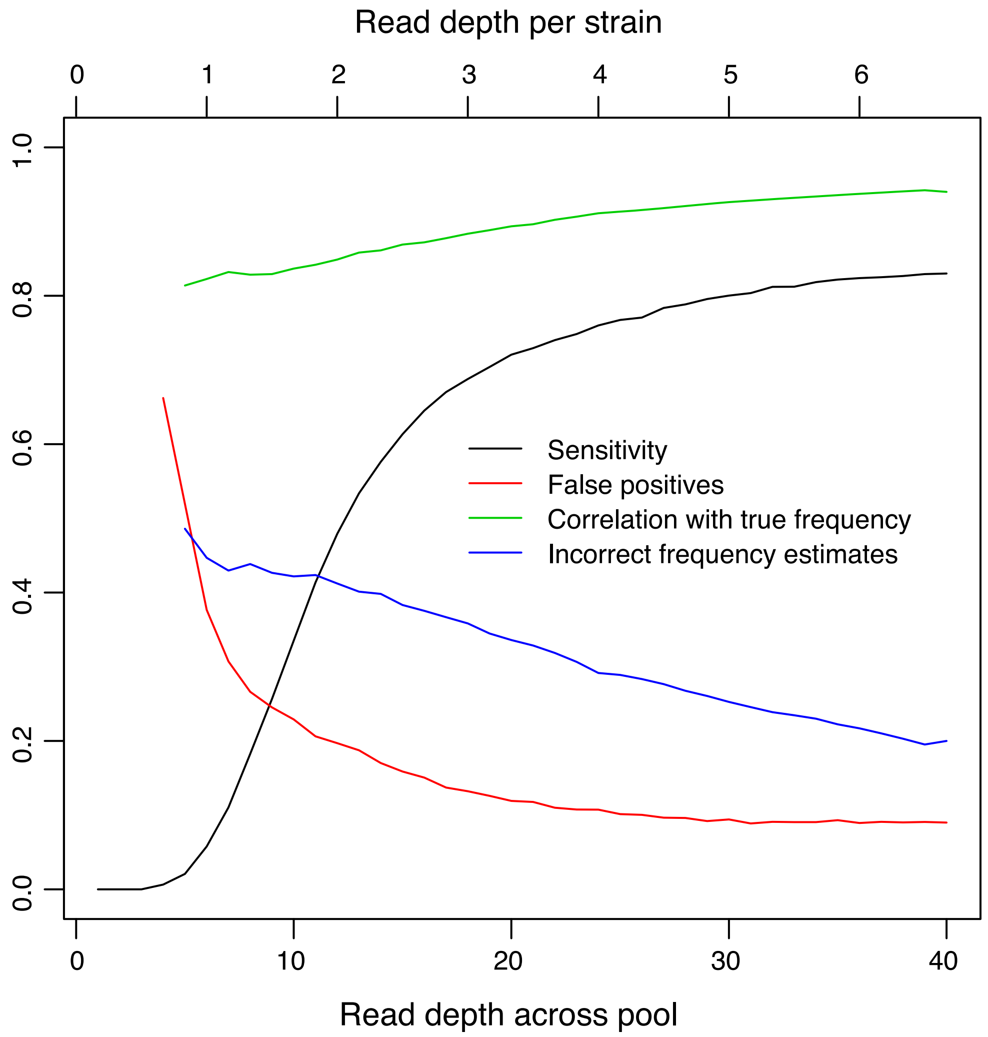

Figure 3.12: Error rates expected at different levels of read depth - Accuracy estimates were based on random sampling of reads from the test pool data at different read depths (1x, 2x, ... 39x across the pool). Data points plotted here show the mean value observed across fifty random samples at each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

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

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

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

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

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

3.3.3.1 SNP detection

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

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

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

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

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

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

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

3.3.3.2 Distribution of SNPs among pools

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

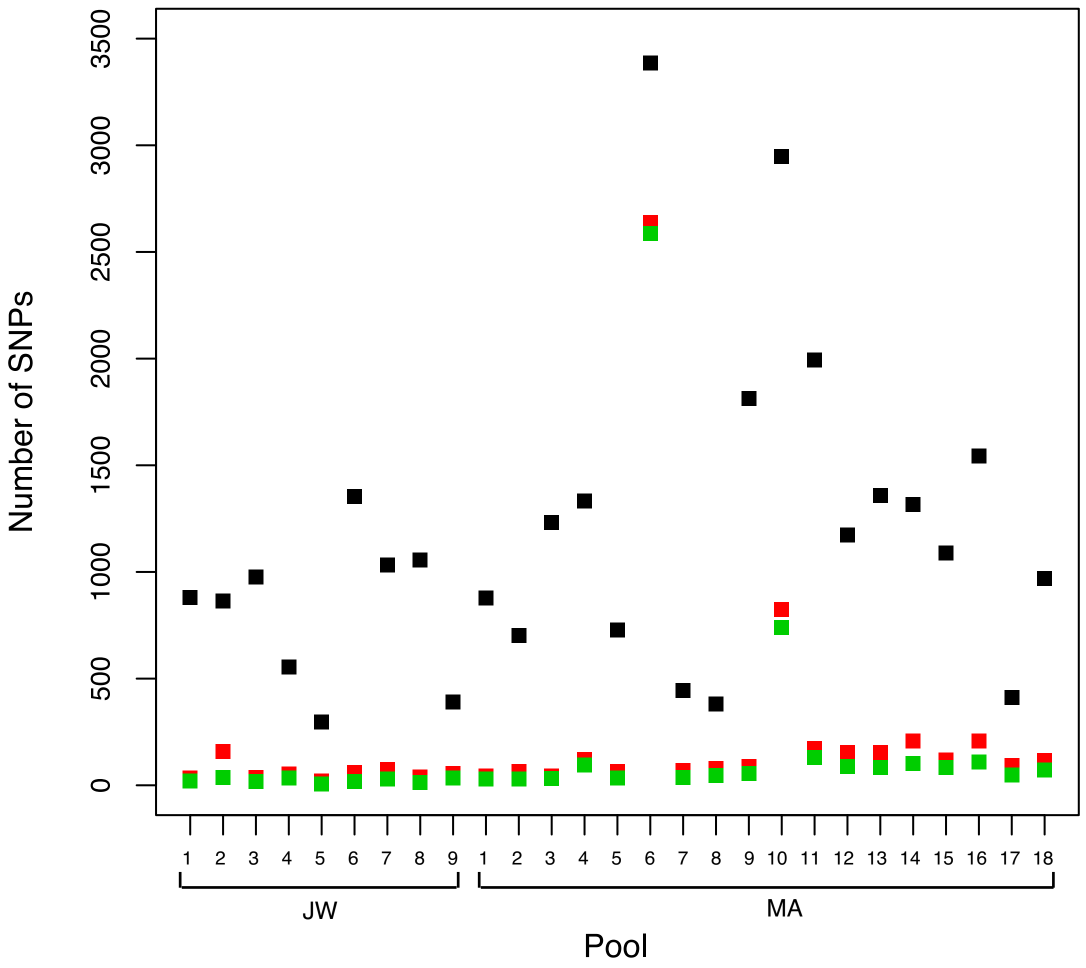

Figure 3.13: Numbers of SNPs detected in each Paratyphi A pool - Numbers are based on SNP detection analysis in all Paratyphi A pools using optimised methods. Total SNPs (black) = total number of SNPs detected at any frequency within the pool. SNPs unique to pool (red) = total number of SNPs detected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

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

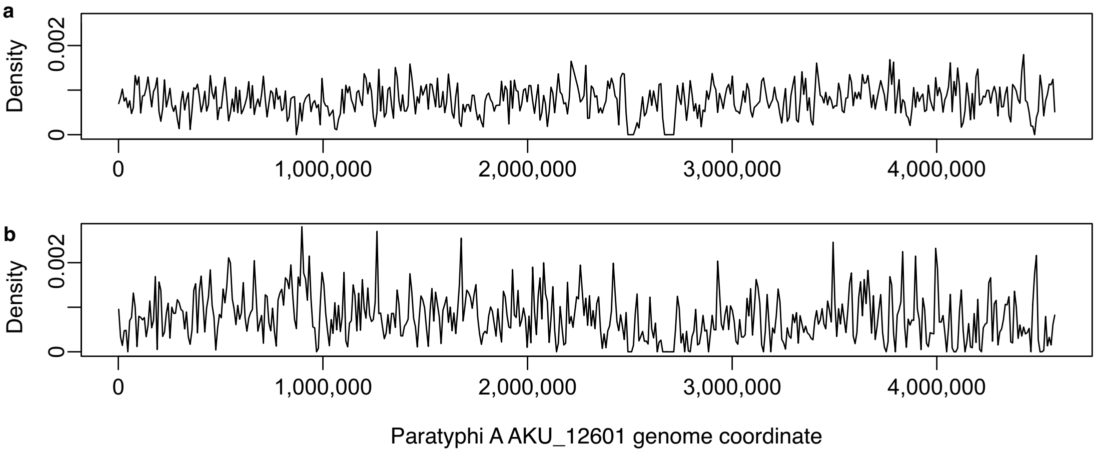

Figure 3.14: Distribution of SNPs detected uniquely in pools MA6 and MA10. - Distribution in the Paratyphi A genome of SNPs detected uniquely in MA6 (a) and MA10 (b) pools, with an estimated frequency of one strain. If the SNPs were caused by recombination from another serovar, we would expect there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

3.3.3.3 Distribution of SNP frequencies

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

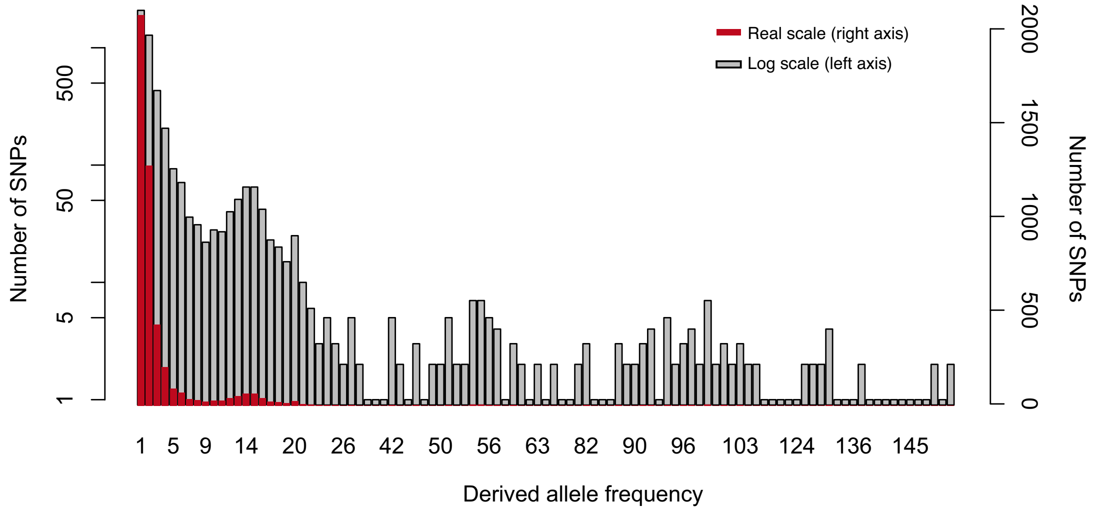

Figure 3.15: Distribution of estimated Paratyphi A SNP frequencies - Note that the same distribution is plotted on a log scale (grey bars, left axis) and non-log scale (red bars, right axis).

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

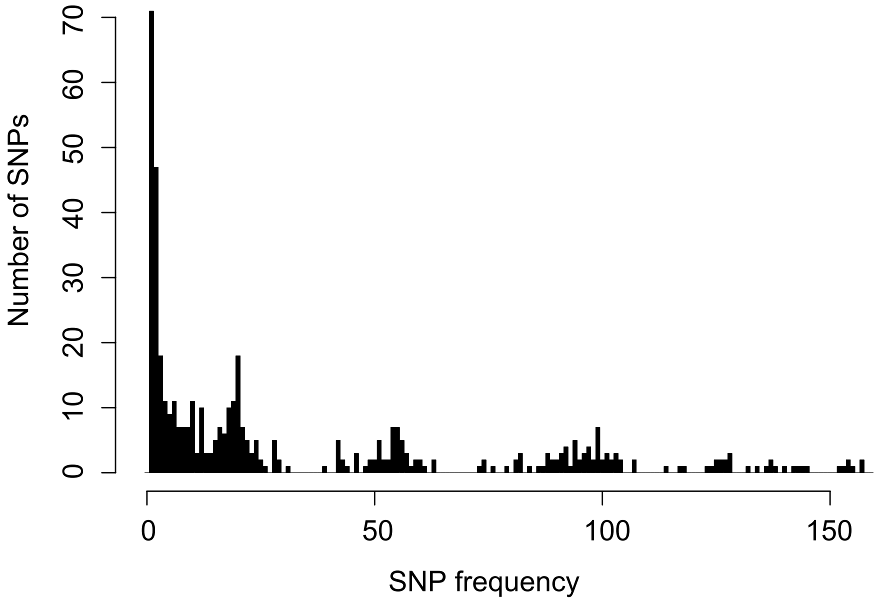

Figure 3.16: Distribution of frequencies across pools for SNPs originally detected among seven individually-sequenced Paratyphi A isolates - Frequencies shown are for 403 SNPs originally detected among seven individually-sequenced isolates.

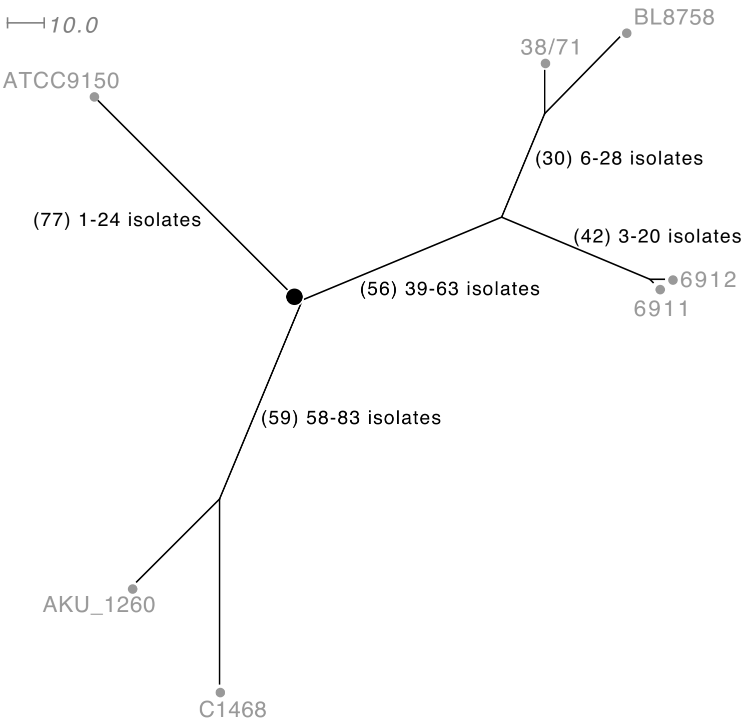

Figure 3.17: Pool-wide frequencies of SNPs defining different branches of the seven-strain phylogenetic tree of Paratyphi A - Scale bar is 10 SNPs. The number of SNPs defining each branch is shown in brackets before the pool-wide frequencies for those SNPs

3.3.3.4 Distribution of SNPs in the Paratyphi A genome

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

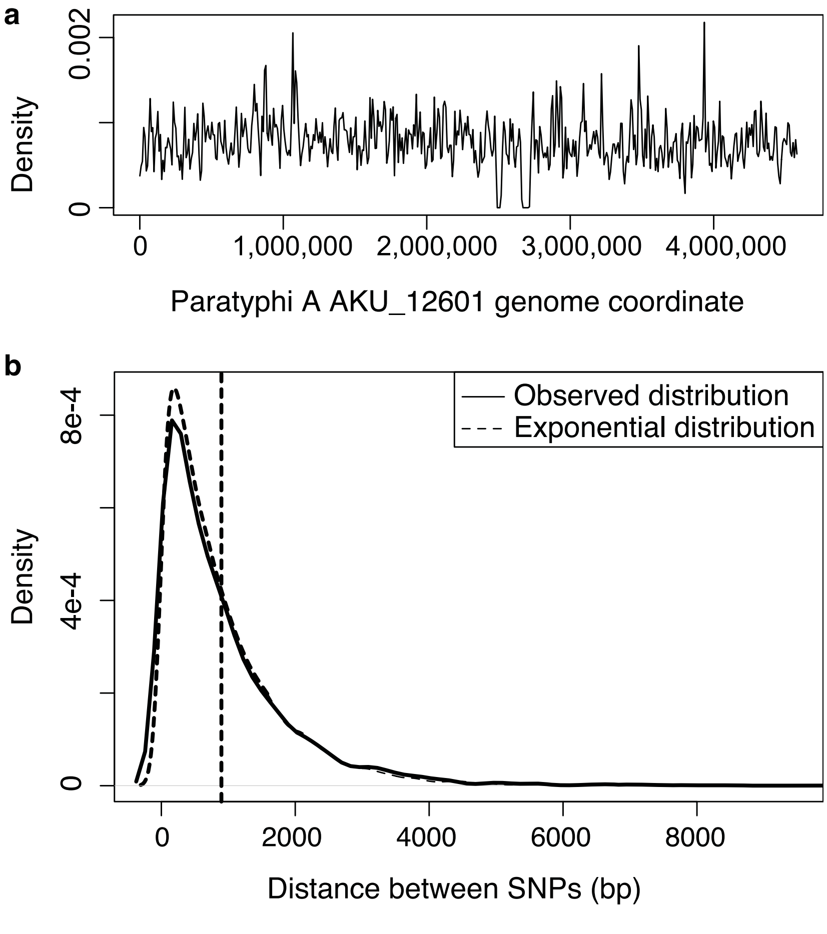

Figure 3.18: Distribution of SNPs within the Paratyphi A genome - (a) Distribution of SNPs in the genome, including those in C1468 and 38/71 and excluding SNPs detected uniquely in pools MA6 or MA10. (b) Distribution of distances between SNPs. Vertical dashed line shows the mean distance, 897 bp.

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

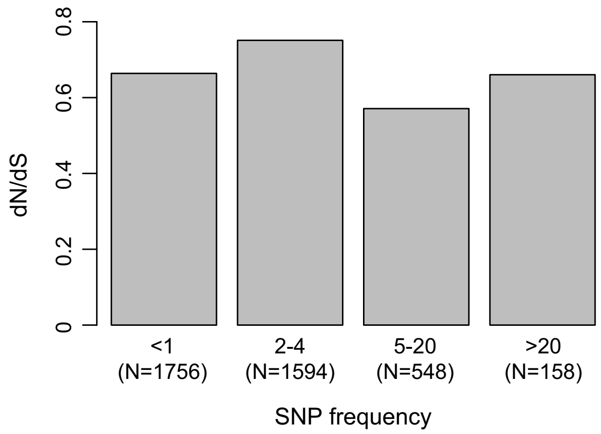

Figure 3.19: $\frac{dN}{dS}$ plotted against SNP frequency in Paratyphi A - $\frac{dN}{dS}$ values calculated for SNPs with different frequency ranges.

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

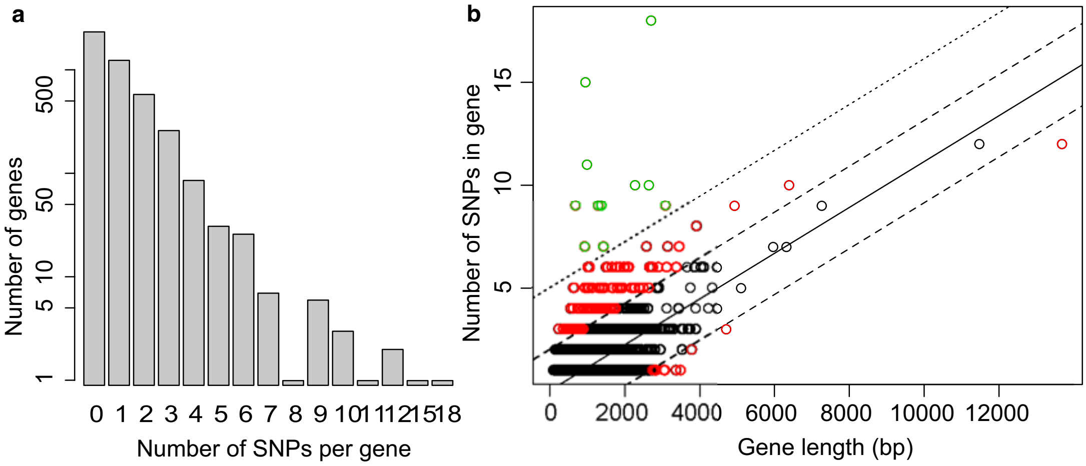

Figure 3.20: Number of SNPs per gene in Paratphi A from pools - (a) Distribution of number of SNPs per gene. (b) Gene length vs. number of SNPs. Solid line shows expected number of SNPs as a function of gene length; dashed line = 2 SNPs more or less than expected based on gene length (red data points);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

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

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

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

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

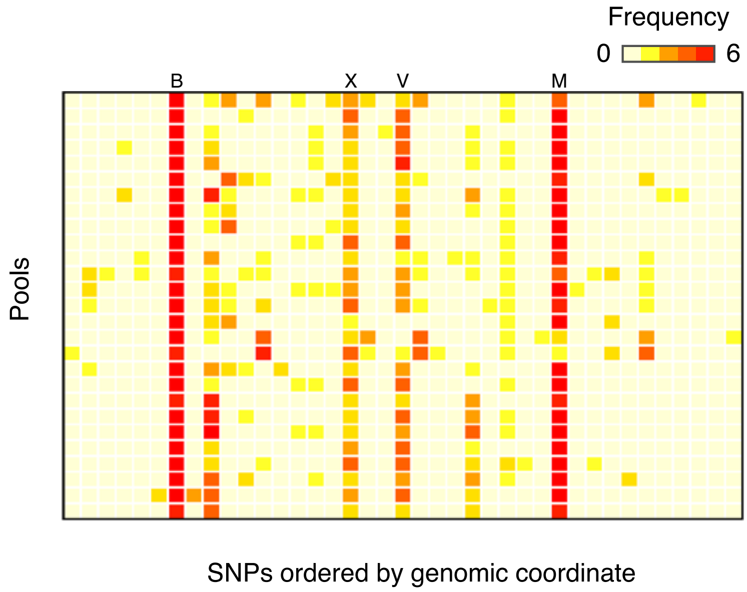

Figure 3.21: Distribution of *wba* cluster SNPs in Paratyphi A pools - B, X, V, M indicate SNPs in genes *wbaB, wbaX, wbaV, wbaM.*

3.3.3.5 Novel pseudogene-forming mutations

A total of 158 nonsense SNPs were detected, introducing stop codons within the coding sequence of 147 genes (see Appendix D). Thirteen of these genes were already pseudogenes in Paratyphi A, and the additional nonsense SNPs should be considered secondary mutations. In total, 153 genes were differentially inactivated in a subset of the Paratyphi A isolates via nonsense SNPs or deletions (see Appendix D). As described previously (2.3.4.2), these novel pseudogenes may be the result of adaptive selection by gene loss (negative selection) or simply tolerance for the loss of genes whose functions are no longer necessary in the host-restricted niche. The genes containing nonsense SNPs were generally more divergent than those without nonsense SNPs, with mean divergence 0.259% compared to 0.191% among all genes containing SNPs (T-test, p = 9.5x10^{-5}). Again this could be explained by either negative or neutral selection pressures. Gene ontology analysis of these 'variable' pseudogenes found that genes encoding protein kinases were overrepresented (N=6, p=0.007) as were those involved more generally in two-component signal transduction systems (N=9, p=0.038). It was not possible to reliably detect frameshift mutations from the short reads data, but given that comparison of the finished AKU_12601 and ATTC9150 genomes identified twice as many frameshift mutations as nonsense SNPs (3.3.1.5), it's likely that there are many more variable pseudogenes present within the Paratyphi A pools.

3.3.3.6 Detection of IncHI1 plasmids

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

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

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

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

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

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

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

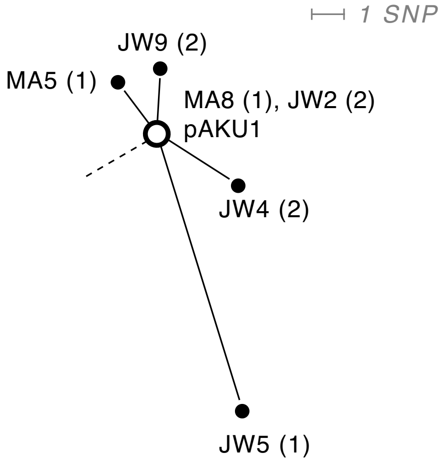

Figure 3.22: IncHI1 SNPs detected in Paratyphi A pools - Split network based on 16 IncHI1 SNPs detected in Paratyphi A pools. Plasmid sequences R27 and pHCM1 were used as outgroups to root the tree; the open circle represents this root.

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

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

3.3.3.7 Detection of plasmid pGY1

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

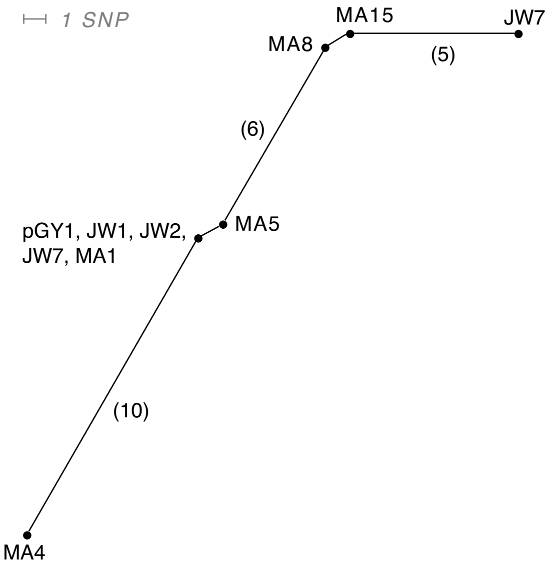

Figure 3.23: Phylogenetic network of pGY1 plasmids detected in Paratyphi A pools - Note that the pGY1 reference sequence matches identically pGY1 sequences in four pools. Branch lengths represent the number of SNPs; scale bar is one SNP; branches longer than one are labelled with the number of SNPs in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

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

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

3.4 Discussion

3.4.1 Strengths and limitations of the study

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

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

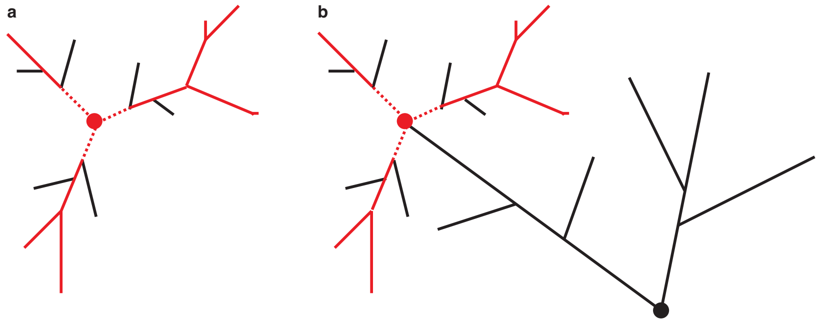

Figure 3.24: Difference in SNP frequencies given biased and unbiased sampling - Red branches indicate the phylogenetic tree of seven individually sequenced Paratyphi A isolates. Black branches indicate hypothetical branches described by the wider population of Paratyphi A sampled in the pools, under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

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

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

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

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

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

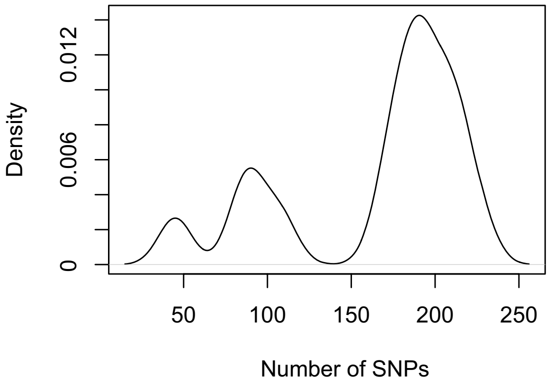

Figure 3.25: Distribution of number of SNPs between two Paratyphi A lineages - The number of SNPs between every possible pair of 6 Partayphi A lineages was calculated (treating 6911 and 6912 as a single lineage), the distribution of SNP numbers is shown.

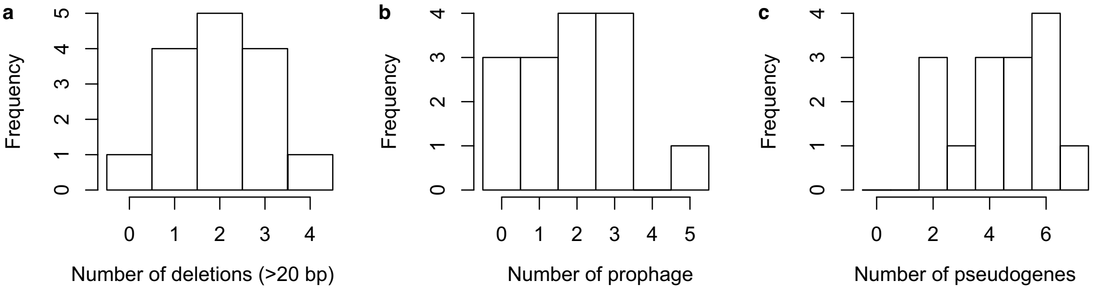

Figure 3.26: Distribution of number of deletions, prophage and pseudogenes between two Paratyphi A lineages - The number of deletions, phage insertions and pseudogenes between every possible pair of 6 Paratyphi A lineages was calculated (treating 6911 and 6912 as a single lineage). The distribution of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

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

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

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

3.4.3 Adaptive selection in Paratyphi A genes

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

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

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

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

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