Chapter 7

Final discussion

In this thesis Typhi and Paratyphi A, the chief agents of enteric fever, were investigated at the population genomic level using novel technologies for sequencing and genotyping. Before the study began, there were two Typhi genome sequences and one Paratyphi A genome available (46, 47, 49). These finished, annotated sequences provided a wealth of information about the genomes of these organisms, facilitating the identification of novel pathogenicity islands (in particular Typhi's SPI7 (46, 92)), prophage insertions (47, 224) and patterns of gene inactivation (46, 49). Comparative analyses of the Typhi and Paratyphi A genomes revealed patterns of similarities, yet they were not the sorts of similarities one might have expected. The genomes shared several genes that were rare among *Salmonella*, yet there were no obvious "smoking guns" with regards to their shared ability to cause systemic infection in humans (49). They each contained over 200 pseudogenes, yet most of the inactivated genes were different (46, 47, 49). Evidence of large-scale recombination was detected, yet the timing and mechanism was unclear (56). The genome sequences, along with that of Typhimurium (50), allowed the development of oligonucleotide arrays, leading to the first attempts to compare gene content between different strains of Typhi or Paratyphi A (49, 511). These array studies, along with comparison of the two Typhi genomes, confirmed what had long been suspected: the Typhi and Paratyphi A populations were both remarkably monomorphic, displaying very few substitution, deletion or insertion mutations that could give clues as to the evolution or population dynamics of these pathogens, or be targeted in genomic epidemiology studies. While the introduction of MLST led to rapid progress in population genetics studies for many bacterial pathogens (460), the analysis of Typhi,

Paratyphi A and other monomorphic pathogen populations lagged behind (1, 471). It became clear that to pursue population-level studies in these bacteria would require the interrogation of vast amounts of sequence. In 2006, a landmark study addressing the global population structure of Typhi was completed by Roumagnac *et al.* (2), using a technique analagous to MLST but involving more than 25 times the number of loci normally targeted in MLST schemes. The study revealed several important aspects of the Typhi population - it was clonal in structure, with no evidence of recombination between lineages; clones were globally distributed, persisting side-by-side over decades on every continent; and a single clone, H58, had recently come to dominate the South East Asian population (2). The <100 SNPs identified in the study were hard-won, with two-thirds of the examined loci yielding no variation at all (2). Fortunately, it was at this point that two novel sequencing technologies came on the scene, bringing with them the possibility of examining entire populations at the whole-genome level. This is where the present study began, with the aim of exploiting the new genomics technologies to study the Typhi and Paratyphi A populations at high resolution.

For Typhi, the 2006 study (2) provided an unbiased framework for selecting isolates for whole-genome comparison. Seventeen novel isolates were chosen for sequencing, to complement the CT18 and Ty2 genomes that were already finished (Chapter 2). The ∼2,000 SNPs identified among these genomes described a single phylogenetic tree, strengthening the suspicion that the Typhi population is essentially clonal and uncomplicated by recombination between lineages. The whole-genome comparisons identified novel prophage over and above those discovered in Typhi CT18 and Ty2 (47, 224), but found no other insertions - just a host of deletions. This is reminiscent of the *M. tuberculosis* population, in which insertions even of prophage or plasmids are virtually unheard of (783, 784) but deletions are quite common (512). In *M. tuberculosis* this genetic isolation is easy to explain, as the bacterium can be isolated for decades inside encapsulated lung granulomas during asymptomatic infection of human hosts. In Typhi, it points towards a similar explanation, suggesting that long-term carriage in the gall bladder is the niche that really matters for the long-term survival of Typhi. This niche is less isolated than the granulomas inhabited by *M. tuberculosis*, which may account for the mid-level frequency of genetic exchange in Typhi: higher than that of *M. tuberculosis*, in the form of phage and plasmids, but lower than that of other *S.*

enterica serovars which show signs of recombination and more extensive plasmid and phage variation (178, 245). The deletions identified in the *M. tuberculosis* genome are conserved within lineages and have proved useful targets for typing (493). A similar scheme is currently in development for Typhi using some of the deletions identified in Chapter 2, which may provide a low-tech way of discriminating between Typhi lineages. The novel prophage identified among the Typhi genomes also provide an opportunity for further research. These regions are currently divided into many contigs but with a little additional sequencing could be improved to the point where cargo genes could be identified, which may prove an interesting source of genetic variation in the Typhi population. The Typhi genome comparisons were not particularly fruitful in terms of identifying genes under selection. However a relatively high level of variation was detected in the genes *yehT* (four SNPs, all nonsynonymous) and *yehU* (nine SNPs, all nonsynonymous) which together encode a two-component regulatory system. In the study by Roumagnac *et al.* (2), 13 SNPs were identified among 105 strains in the 500 bp fragment analysed from the sensor kinase gene *yehU*, compared to 0-5 in other gene fragments of the same size. The function of this regulatory system is currently being investigated using a combination of mouse models (with Typhimurium), phenotypic screening and gene expression analysis.

Typing of the Typhi SNPs identified by comparative genome analysis was applied in Chapter 6 to the analysis of Typhi populations in localised endemic areas. This was the highest resolution sequence-based interrogation of Typhi ever performed, and offered novel insights into the structure and dynamics of the Typhi populations in each area, which could be directly compared. These studies confirmed that multiple lineages of Typhi co-circulate in very narrow windows of time and space, but also revealed fluctuations in the composition of the population over time. The former is attributable to the central role of asymptomatic carriers, each of whom provides an independent and persistent source of infection with their own particular Typhi strain. The latter demonstrates the importance of other factors for short-term changes in the Typhi population, likely including things like climate, human behaviour and immunity in the human population, as well as random chance. Geographically, there was evidence of global dissemination of Typhi haplotypes over the long term, but also rather localised expansions of H58 sublineages in the short term (evident in the localised studies in

Vietnam (sublineage C) and in Nepal and India (sublineage B) spanning the last five years). There was also evidence of long term trends in the Typhi population, with SNP typing in both the global and Kenyan collections showing H58 becoming more and more dominant over time. However, the resolution offered by the current set of SNPs is inadequate to study the expansion of H58 in really fine detail. For example, isolates assigned to node H58-B undoubtedly harbour additional variation that happened not to be present among the seven sequenced isolates. Thus by SNP typing, we are blinded to this variation and may be tempted to assume that the H58-B isolates identified as potentially escaping the effects of the Vi conjugate vaccine in India represent a single clone, which is very closely related to that recently arrived in Kenya. However this may well be wrong, and can only be resolved by additional sequencing. Until it becomes feasible to engage in whole-genome sequencing of entire collections of Typhi isolates (which may not be too distant a prospect, discussed below) a multi-step approach may be best, where isolates are typed first using a subset of phylogenetically informative SNPs to discriminate major clusters, followed by additional sequencing to identify additional SNPs able to discriminate within those clusters. For example, based on the Typhi populations studied here, a sensible approach may be to (i) type a few SNPs that can discriminate between H58, H42 and other lineages plus all SNPs known to discriminate within H58, then (ii) select a manageable subset of isolates within the H58 (and potentially H42) clusters (maximising variability by pre-screening with PFGE) and sequence them, perhaps using a pooled approach to limit costs, followed by (iii) additional typing within clusters at SNP loci identified in step ii. This approach is currently being trialled in a large study of Typhi isolates from Kathmandu, using Sequenom for SNP typing (which allows rapid design and execution of SNP typing assays in under two weeks) and Solexa sequencing for SNP detection.

One outstanding issue with respect to the expansion of Typhi H58 is the role of the ST6 IncHI1 plasmid. The association between strain and plasmid is remarkable, with every single ST6 plasmid identified in this study (N=235) found within H58 host strains. However at this point it is not clear whether the apparent success of the ST6 plasmid subtype is simply hitchhiking on the expansion of H58, or whether it played a role in the success of the clone. The evidence from *IS*1 insertions suggests that the common ancestor of extant H58 lineages carried the plasmid (6.3.7), so the latter is at least

possible. One potential mechanism for selection of ST6-containing strains, apart from drug resistance which is essentially identical to that conferred by plasmids of other IncHI1 subtypes, is the *betU*-carrying transposon *Tn*6062. The osmoprotectant BetU may help host cells to survive in the environment or within the urinary tract of human hosts, facilitating long-term carriage and transmission. Competition experiments in a variety of different media may help to detect phenotypic differences between ST6-containing and ST1-containing Typhi strains, although care must be taken to control for Typhi strain background.

Prior to this study, there was barely any phylogenetic information available for Paratyphi A. Comparison of gene content using microarrays had detected five chromosomal deletions including three prophage deletions, but no SNPs were known. A second Paratyphi A genome sequence was finished at the Sanger Institute, providing the first opportunity for comparison at the nucleotide level. However by this time, 454 and Solexa sequencing was also available, thus the first comparative analysis of Paratyphi A genomes included not two but seven isolates (Chapter 3). Selection of these isolates was random and we still do not know how representative they really are, but it seems they do capture a lot of the structure present in the larger global sample of isolates sequenced in pools (3.4.1). The level of variation detected among these genomes was markedly less than that observed between Typhi genomes (4.3.1), supporting previous suggestions that the Paratyphi A population harbours even less genetic variability than Typhi (479). The pooling strategy was designed to maximise genome-wide variation detection in the absence of any guiding phylogenetic framework like that available for Typhi. This was essentially a compromise between individual interrogation of a smaller number of isolates, which provides not only variation detection but also haplotypes appropriate for phylogenetic analysis, and maximal sensitivity to detect variation around the entire chromosome, which was expected to be very low. Haplotypes and phylogenetic structure could then be resolved using high throughput SNP typing. This last step was not pursued during the course of this study, partly due to time constraints and partly due to rapid changes in the relative cost-effectiveness of sequencing and SNP typing. As of mid-2009, it is increasingly feasible to sequence tens to hundreds of individual bacterial genomes of 4-5 Mbp in a single Solexa run using indexed libraries (785). As throughput continues to increase, thanks to improvements in sequencing

chemistry that allow ever-longer paired-end reads (recent runs at the Sanger Institute exceed 30 Gb), it is likely that sequencing of even more genomes will be feasible very soon. Given that sequencing combines detection of known and novel variants, the cost of high throughput SNP typing will need to be significantly lower than that of sequencing in order to be of any benefit. The further pursuit of population structure in Paratyphi A currently involves SNP typing of ∼100 of the SNPs identified in Chapter 3, and will soon move to individual sequencing of the isolates so far sequenced only in pools.