2. Genomic diversity of *Mycobacterium tuberculosis* over short time scales

The majority of this work has been published in:

J. M. Bryant, A. C. Schurch, H. van Deutekom, S. R. Harris, J. L. de Beer, V. de Jager, K. Kremer, S. A. van Hijum, R. J. Siezen, M. Borgdorff, S. D. Bentley, J. Parkhill and D. van Soolingen (2013). "Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data." BMC infectious diseases **13**(1): 110.

Statement of contribution:

I carried out all bioinformatic analyses and interpretation. The study was initiated by JP, KK and DVS. Advice and bioinformatic training was provided by SRH. Help with interpretation and metadata was provided by ACS. Strain selection and sample preparation were carried out by the Tuberculosis Reference laboratory for the Netherlands.

2.2. Introduction

Many consider the routine clinical application of whole genome sequencing to *M. tuberculosis* to be inevitable in the near future (Koser, Ellington et al. 2012, Walker, Ip et al. 2013). Contact tracing and outbreak investigation will particularly benefit, as these currently rely on the information gathered from epidemiological investigation and molecular typing techniques to determine all possible contacts and routes of transmission for each patient. The higher level of resolution that whole genome sequencing offers could enable a more informed ability to include or refute possible links. However, in order to achieve this an understanding of the molecular clock over short time scales in the context of direct patient-patient transmission is essential. By knowing how much variation is generated within and between patients we can begin to judge whether direct transmission is likely to have occurred between individuals, given what we know about their contact.

Estimations of mutation or substitution rate are very much dependent on the evolutionary scale on which they are being measured on. The term mutation rate commonly refers to the basal *de novo* generation of mutations in the absence of selection. Instead here, where the variants detected in the context of patient-patient transmission are being sampled, the term substitution rate is used. In this context mutations have risen to a high enough frequency to be sampled (so highly deleterious mutations have been removed), but may not become permanently fixed in the *M. tuberculosis* population. Currently our knowledge of the intra-patient mutation rate (closer to basal mutation rate than inter-patient comparisons but not selectively neutral) of *M. tuberculosis* is limited to a single study based on infection experiments of macaques. Based on 15 isolates collected from four infections, an estimate of 0.39 (0.16-0.80 95% CI) single nucleotide polymorphisms (SNPs) per genome per year was estimated (Ford, Lin *et al*. 2011). More recently, an estimate of 0.5 SNPs/genome/year was made on the basis of within and between patient sampling of 93 patients from a larger study in the UK (Walker, Ip *et al*. 2012). This suggests that the intra-patient mutation rate may be similar to the inter-patient substitution rate.

This study aimed to further characterize the rate of change of *M. tuberculosis* observed between patients over short time scales, and to explore the strength of this technique to refute and confirm direct transmission links.

2.3. Methods

199 *M. tuberculosis* isolates were chosen by the Municipal Health Service in Amsterdam, comprising of isolates from 151 patients with 97 known epidemiological links between them, and another 48 patient isolates from the same RFLP clusters but with no evident epidemiological link. Isolates were chosen to represent a broad phylogenetic range, belonging to 42 RFLP clusters. All genotyping including IS*6110* RFLP, Spoligotyping and 24-locus variable number of tandem repeats (VNTR) typing were performed using standardized methods (van Embden, Cave et al. 1993, Kamerbeek, Schouls et al. 1997, Supply, Allix et al. 2006) by the Tuberculosis Reference laboratory the RIVM, the Netherlands.

The genomic DNA libraries were subjected to paired-end sequencing on the Illumina Genome Analyzer GAIIx platform. Thirty-three of the isolates were sequenced with a read length of 76 bases and the remaining 166 with a read length of 108 bases.

The raw sequencing reads were mapped to a corrected version of the H37Rv reference and variants were called as described in Methods 8.2 and 8.3. Mapping and SNP calling were also carried out independently at the Center of Molecular and Biomolecular Informatics (CMBI), Radboud University, using RoVar (Robust Variant detection in genome sequences using Next Generation Data from various platforms: Jager, B.A.M. Renckens, R.J. Siezen, and S.A.F.T. van Hijum, unpublished). The mapping results were compared using the epidemiological linked pairs as a test set. Most SNPs were found to agree except those found in regions flanking insertions. As short insertions and deletions are difficult to call in general, only SNPs were considered for all subsequent analysis. The genetic distance was calculated between epidemiologically linked pairs by comparing the SNPs called in each isolate. A SNP difference was only counted where there was high confidence in the base call in both isolates.

A maximum likelihood tree was constructed as described in Methods 8.6. Path-O-Gen was used to plot root to tip distances against time (Rambaut 2007) (see Methods 8.7). This program uses linear regression to root trees with date information at the position that is most compatible with the assumption of the presence of a molecular clock.

2.4. Results

2.4.1. Overview of mapping results

For the 199 samples, sequencing reads covered an average of 95.6% of the genome to a depth of approximately 100 fold. With respect to H37Rv, 11,879 positions had a SNP called in at least one of the isolates. A maximum-likelihood phylogeny based on the variants called, revealed four of the globally dominant lineages (Figure 5). The 97 linked pairs had a mean SNP difference of 3.42 (range of 0-149) and 37 of the pairs had no detectable SNP difference.

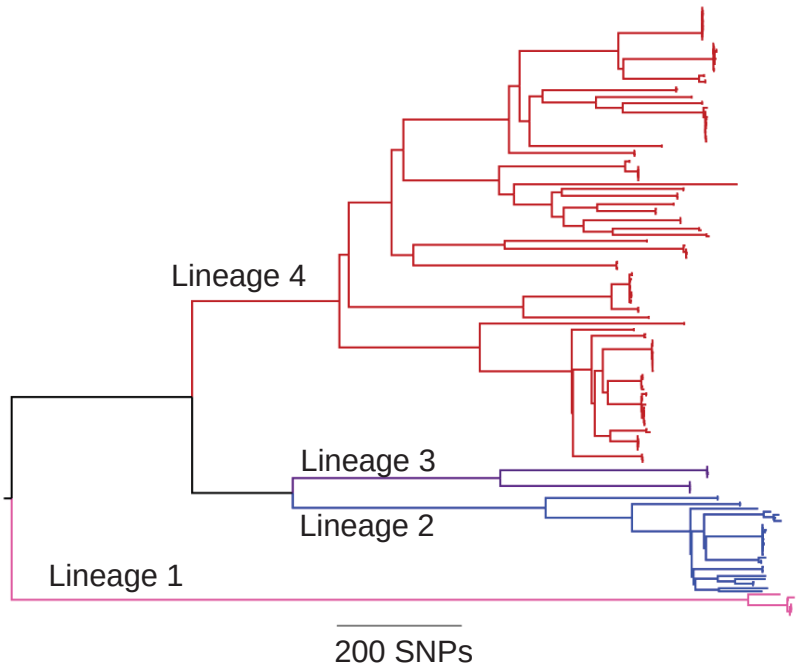

Figure 5 - Whole genome maximum likelihood phylogeny of 199 samples in dataset.

2.4.2. Homoplasic SNPs

In total, 16 homoplasic non-synonymous SNPs were identified (Table 1). Only two synonymous homoplasies were identified (Rv0161, which encodes an oxidoreductase and Rv2005c and stress related protein), suggesting that the high number of non-

synonymous homoplasies observed was unlikely to have occurred by chance alone. Five of these found in genes *rpoB* (Ramaswamy and Musser 1998), *gyr*A (Takiff, Salazar *et al*. 1994), *rrs* (Maus, Plikaytis *et al*. 2005), *kat*G (Heym, Alzari *et al*. 1995) and *emb*B (Sreevatsan, Stockbauer *et al*. 1997) have previously been associated with drug resistance. Two of the other homoplasic SNPs occurred in genes thought to be involved in pathogenicity: *ino1* (Movahedzadeh, Smith *et al*. 2004) and *opcA* (Jiang, Zhang *et al*. 2006) and three (*opcA*, Rv2082 and Rv3077) were observed in a recent study focusing on convergent evolution in *M. tuberculosis* (Farhat, Shapiro *et al*. 2013). It is likely that homoplasy has occurred in these genes due to recurring selective pressures for traits such as antibiotic resistance. This suggests that the 11 homoplasic SNPs with no ascribed function deserve further investigation, as they may possibly represent previously un-described pathogenicity, antibiotic-resistance or associated compensatory mutations.

Table 1: Homoplasic SNPs identified in this study

Gene	Number of branches	Amino acid change
gyrA	2	D94G
rpoB	2	S450L
Rrs	2	C517T
katG	5	S315T
embB	2	M306V
ino1	2	G190R
Rv0750	2	L27V
ribG	2	K30N
opcA	2	A103T
Rv1760	2	M397T
lldD2	5	V253M
Rv2082	2	L53R
Rv2709	2	E42K
Rv3077	2	R452H
Rv3463	2	G94D
aspB	2	R358Q

2.4.3. Deriving a molecular clock

For the 97 epidemiologically linked pairs, the relationship between time and the number of SNPs accumulated was investigated. Only SNPs accumulated in the secondary case in each of the linked pairs were used and SNPs found only in the

primary case isolate were excluded, as these are likely to represent either variation in the source host population that is not present in the transmitted population, or SNPs generated via laboratory passage. SNPs conferring drug resistance were also excluded (n=7), as these are likely to be subjected to a strong selection pressure and be less clock-like in the rate in which they are appear. In addition three pairs were excluded based on the phylogenetic evidence discussed below in section 2.4.4.

There was a poor correlation between the number of SNPs accumulated and the time elapsed for each patient pair (Figure 6). However, when drug resistant and sensitive pairs were plotted separately, there was an improved correlation for the sensitive pairs, but not the drug resistant pairs (Figure 6). The reason for this is unclear but a possible explanation could be based on the differing selection pressures and effective population sizes of the two groups.

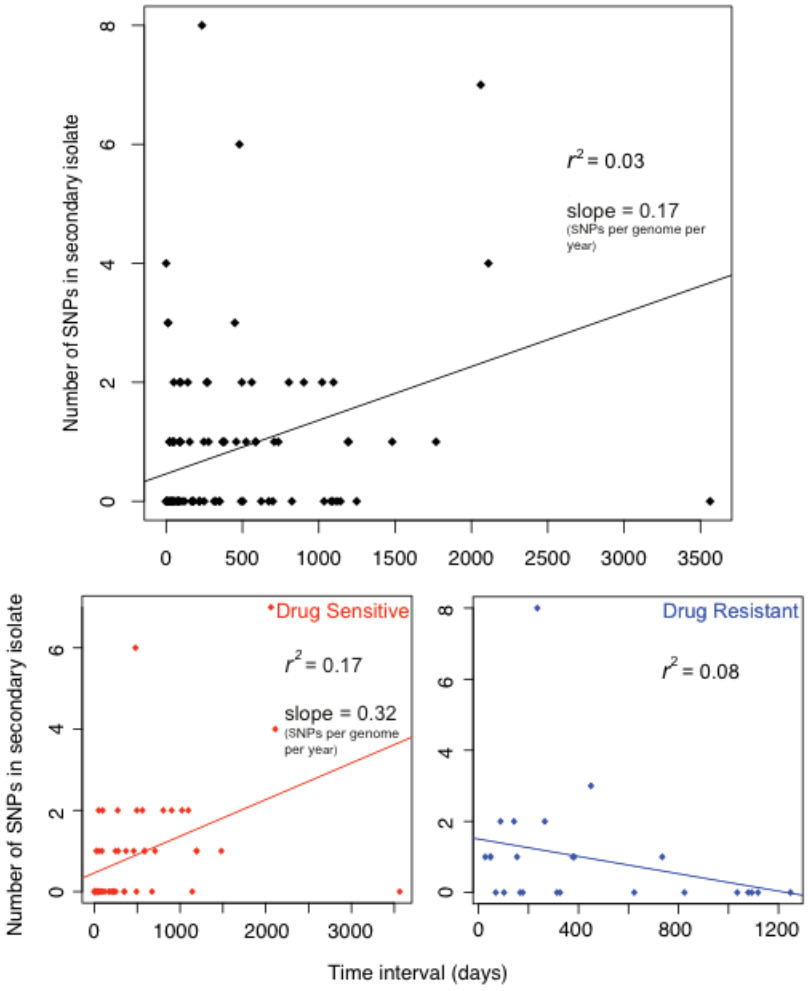

Figure 6 - Poor correlation between time and number of SNPs accumulated in the secondary case isolate for drug resistant and sensitive isolates. Correlation coefficient for linear regression models are shown

The slope of the graph provides an estimate of substitution rate, and for the sensitive isolates this was 0.32 SNPs per genome per year but with a large degree of variation around the mean reflected by an r^2 value of 0.17. This variability could reflect a number of sources of error that have to be taken into account. The first possible source of error is that the epidemiological inference could be incorrect and that direct transmission did not occur between the pairs. Secondly, there is an unknown degree of error regarding how well the date of transmission is represented by the date of isolation. Transmission is likely to have occurred prior to the isolation date (as transmission is generally considered unlikely for a patient receiving treatment), but due to the slow progression of the disease the degree of error could be very large. Finally, for 26 pairs SNPs were detected in the primary case (averaging 0.64 SNPs per pair) that were not found in the secondary, suggesting that the sampled isolate is unlikely to represent the transmitted population. This suggests that genetic diversity in the infecting population of *M. tuberculosis* exists in patients, of which only a small proportion has been sampled.

To control for the sources of error described above, a substitution rate was inferred from the entire dataset, thus not requiring assumptions about the routes of transmission. The presence of a clock-like signal was investigated using Path-O-Gen (Rambaut 2007) for the 197 samples for which a date of isolation was available. Lineage specific phenotypes have been frequently proposed (Brown, Nikolayevskyy et al. 2010, Krishnan, Malaga et al. 2011), and due to the possibility that the different lineages may have different mutation rates, the analysis was carried out per lineage. There was a complete lack of correlation between the accumulation of SNPs and time for all the lineages (Figure 7).

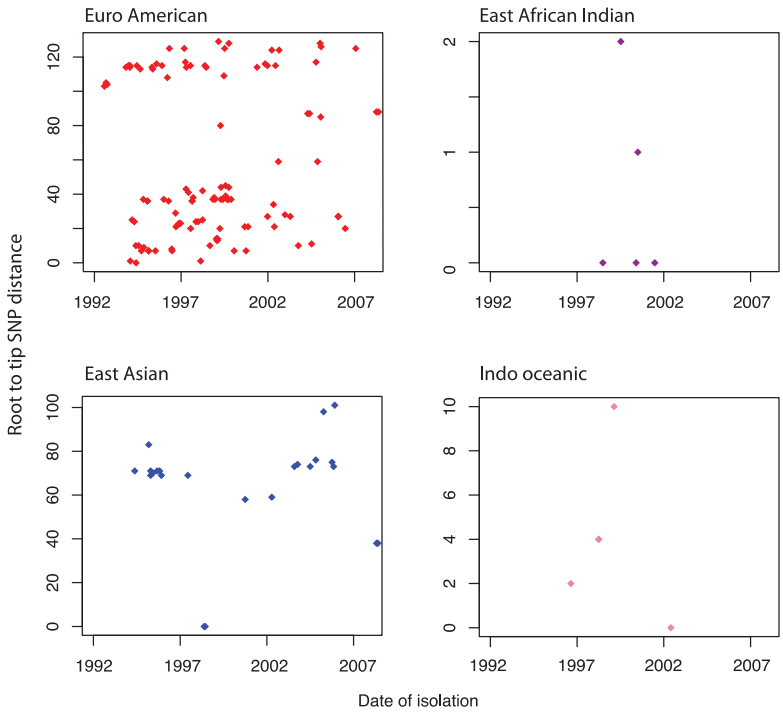

Figure 7 - Per lineage root to tip distance vs. time. The correlation coefficient of the linear regression models was poor for all lineages, with r^2 values of 0.002,0.03,0.006 and 0.06, clockwise.

This is perhaps unsurprising when the age of these lineages is considered, which would exceed thousands of years if the out-of Africa model of *M. tuberculosis* expansion is correct (Comas, Coscolla *et al*. 2013). The time dependency of the molecular clock is a well described phenomenon (Ho, Shapiro *et al*. 2007), and only the more recent variation nearer the tips of the tree is likely to be neutral and clock-like. For this reason, this analysis was then carried out on five of the largest within-lineage clusters that are more likely to represent a neutral accumulation of SNPs (Figure 8a). The linear regression slope ranged from 0.08 to 0.43 SNPs per genome per year, with this variation probably reflecting the small number of isolates and SNPs observed. When the cluster data was combined, a mean rate of 0.27 SNPs per genome per year (95% CI 0.13, 0.41) was estimated (Figure 8b). Additionally, when the age of the clusters was plotted against the number of SNPs accumulated (Figure 9), controlling for the number of isolates, a similar rate of 0.34 SNPs per genome per year was obtained.

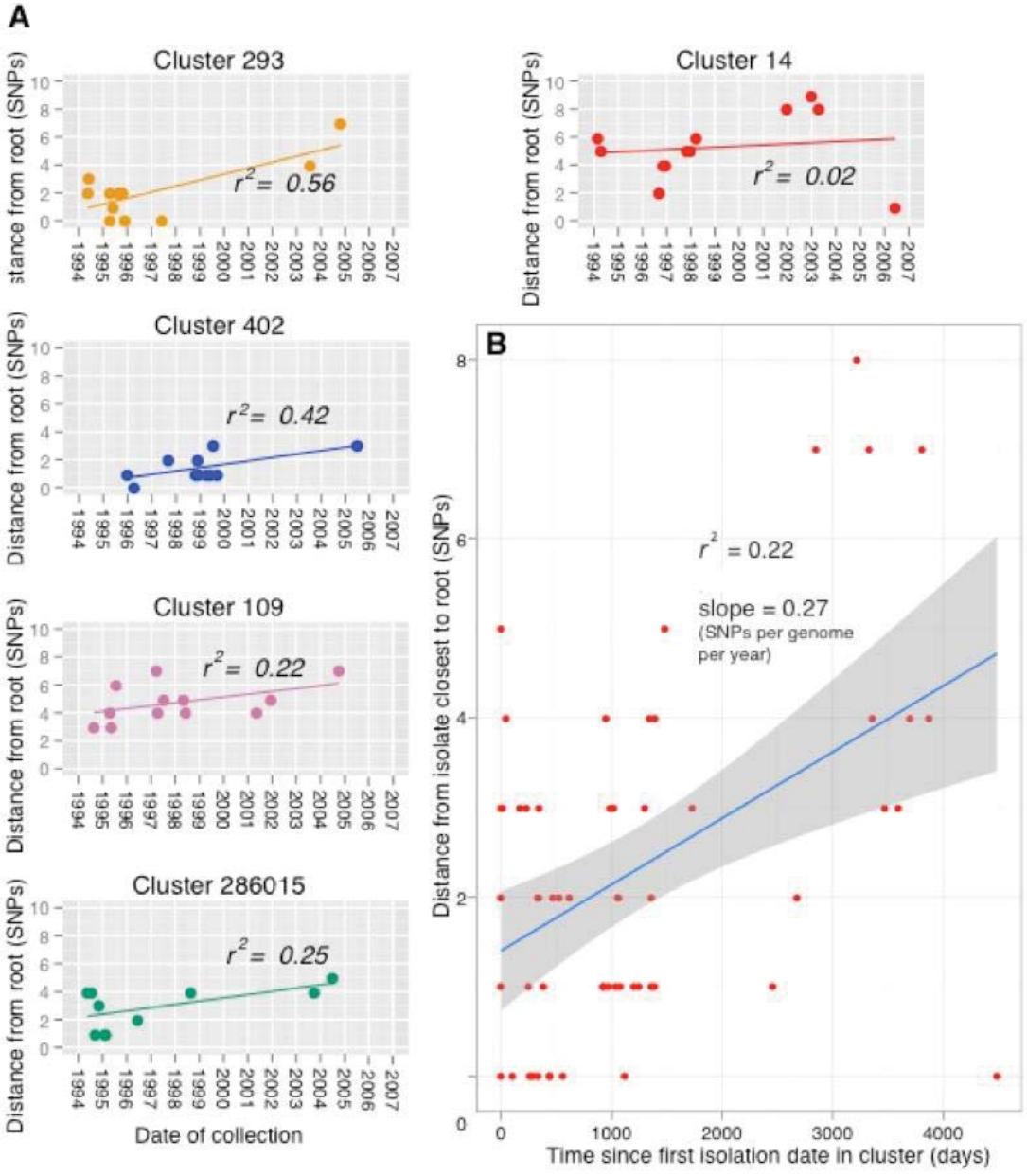

Figure 8 - Date of collection vs. root to tip SNP distance of the 5 largest clusters. B. Data combined from A. Time represents days since first isolation in the cluster. Shaded area indicates 95% confidence of linear regression model.

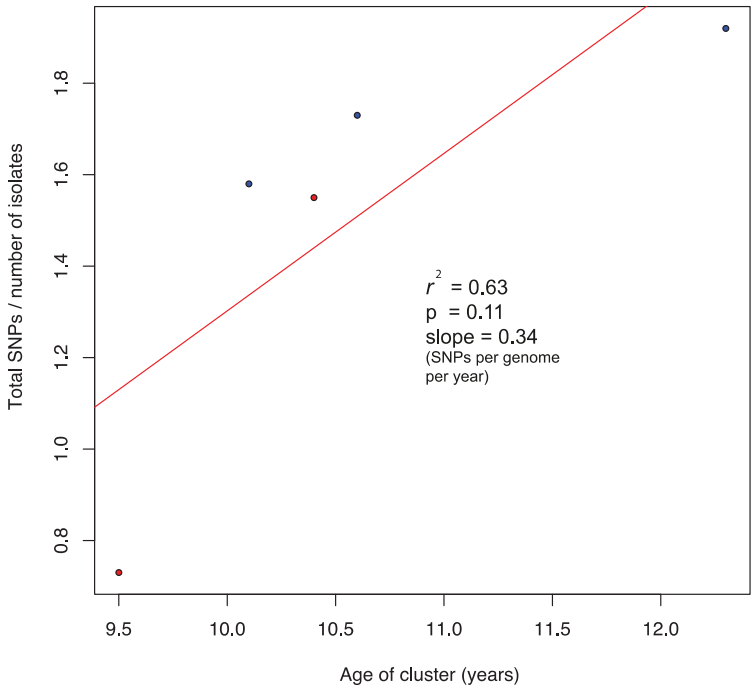

Figure 9 - Diversity vs. age for the five largest clusters. The total number of SNPs was corrected for no. of isolates in the cluster. Red indicates drug resistant clades and blue are drug sensitive.

In summary, three methods agreed on an average rate of ~0.3 SNPs per genome per year which is remarkably similar to that estimated by Ford *et al* (Ford, Lin *et al*. 2011) using the tuberculosis macaque infection model. However, the correlation co-efficient was consistently poor (maximum r^2 value of 0.6 in one case) and the level of variation observed at the isolate level was high. In addition the statistical significance of these correlations could not be assessed, as the root-to-tip distances are not independent. A more rigorous way to estimate the substitution rate is to use Bayesian coalescent analysis. However none of the datasets discussed here successfully converged using BEAST v1.7.5 (Drummond and Rambaut 2007) (see section 8.8). This is likely due to small sample size, and sampling frame and the very low rate and stochastic nature of SNP accumulation, and indicates that this estimate needs to be used with caution.

2.4.4. Using phylogeny to exclude direct transmission

Instead of looking at possible transmission events in isolation, deep sampling of a phylogenetic cluster can provide context, which can be used to make more confident inferences. Thus the structure of a phylogeny can be used to assess whether a direct transmission event is likely to have occurred. Isolates that represent a recent

transmission event are expected to be adjacent on the tree and share a most recent common ancestor, as shown in Figure 10d. If other isolates occupy the common nodes between the linked isolates in question, then this is evidence against direct transmission. This scenario was identified for two of the pairs in the study (Figure 10a and b). However, it is possible that the source case could have been carrying an infection with a diverse *M. tuberculosis* population, comprised of several sub-lineages as observed previously (Sun, Luo *et al*. 2012). In such a scenario, the entire cluster may in fact represent within patient diversity and each patient isolate is effectively a sample of this. As liquid cultures (i.e. not colony purified) were used in this study, this heterogeneity may be preserved at the variable positions. However, no evidence for this was found, and in the absence of multiple samples from each patient, this strongly suggests that these pairs do not represent direct transmission events.

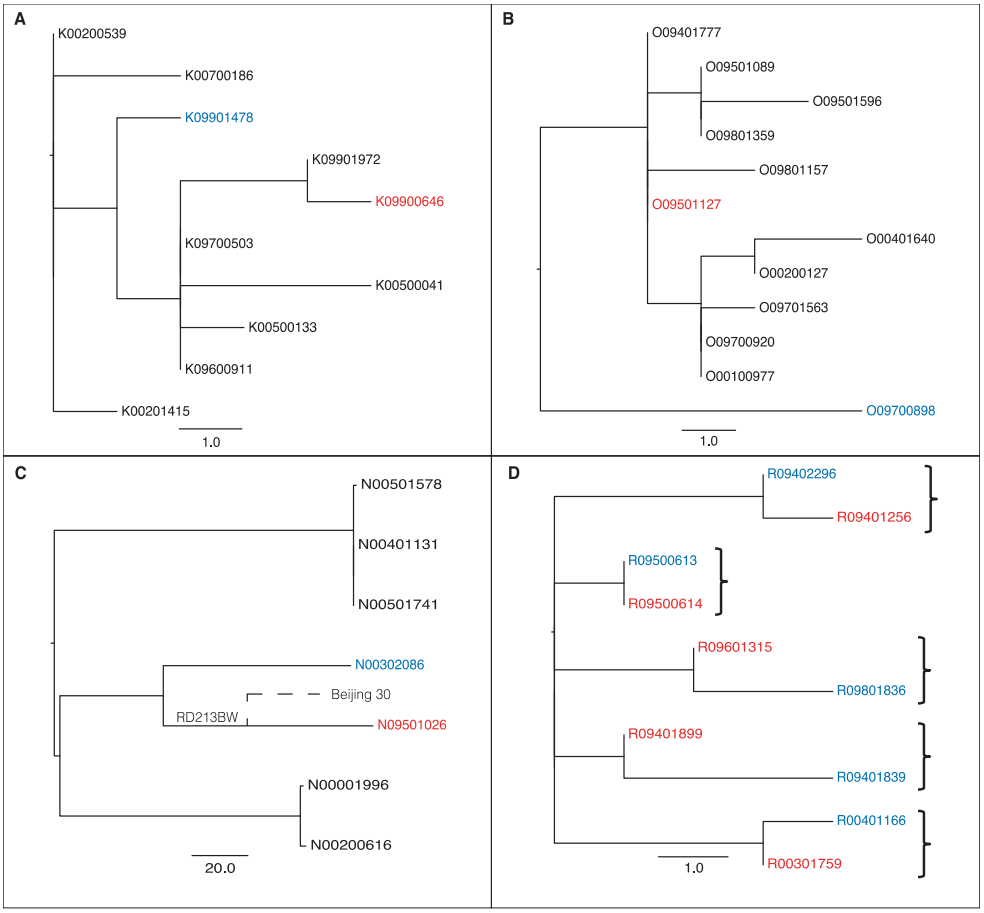

Figure 10 - Exclusion of epidemiologically linked pairs based on phylogenetic position. Red indicates primary case isolate, blue is the secondary case isolate. Maximum likelihood trees were rooted using the nearest non-clustered isolate as an outgroup. A: excluded pair 1. B: excluded pair 2. C: Excluded pair with SNP difference of 149. D: Example of expected phylogenetic positioning of direct transmission pairs, brackets indicate paired isolates.

One pair from the East Asian linage had a particularly large SNP difference of 149 (Figure 10c). The suspected source case patient lived in the same street as the suspected secondary case patient. It is unclear, however, if they were in direct contact with each other. Both isolates shared an IS*6110* RLFP pattern but their 24-locus VNTR pattern differed in 6 loci. With no detectable evidence of recombination or SNPs in possible hypermutator genes, the sequencing data was examined more closely and two independent deletions were identified: each unique to either isolate (Figure 11). The large deletion of part of the *pks1* gene found in the source isolate was found in another East Asian strain, Beijing 30, in a previous study (Tsolaki, Gagneux *et al*. 2005), suggesting that a more recent common ancestor exists than between these two isolates. This evidence along with the large SNP difference means that the possibility of recent direct transmission can be confidently excluded. In the absence of whole genome sequencing, the clear genetic separation of these isolates would have been un-detectable.

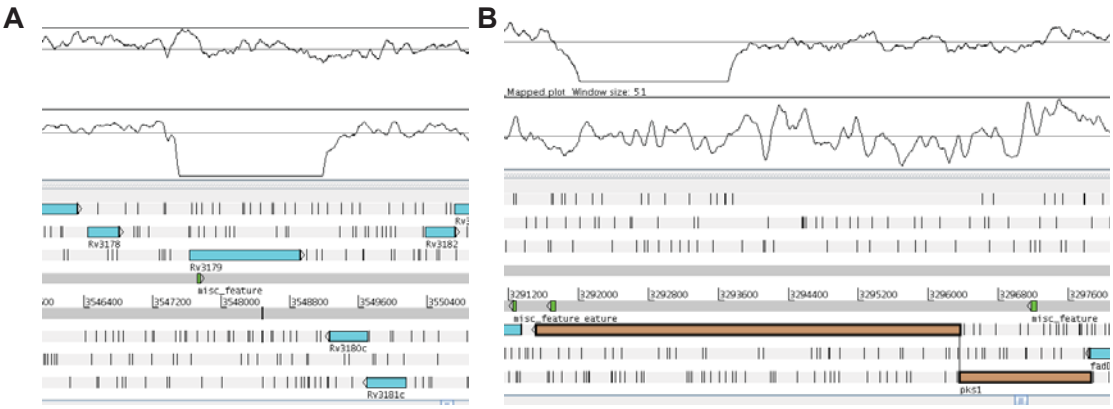

Figure 11 - Deletions identified in isolates N09501026 (top row) and N00302086 (bottom row). Plots represent mapping coverage to H37Rv reference genome. Figures adapted from Artemis (Carver, Harris *et al*. 2011). A) Deletion in Rv3179 in N00302086. B) Deletion in *pks1* in N09501026.

2.4.5. Identifying novel transmission events

In low incidence countries, identical RFLP types are often used as an indicator of possible transmission. In this dataset, 572 pairs of isolates had identical RFLP types, which had SNP distances ranging from 0-149, with a median of two SNPs. Figure 12b further confirms that the linked pair with a SNP distance of 149 is a clear outlier showing that it is distinct from the rest of the same-RFLP and epidemiologically

linked pairs. 95% of same-RFLP pairs have SNP distances under 11, indicating that in general RFLP type is a good indicator of phylogenetic relatedness. However, whole genome sequencing provides a higher resolution. For example in Figure 10d, all of the isolates in this cluster would be indistinguishable via RFLP, but at the whole genome level individual transmission events can be inferred. Figure 12b demonstrates that many pairs of isolates of the same RFLP type, with currently no known epidemiological link, have SNP distances that overlap with the range observed between the 94 linked pairs.

Strikingly, 82 pairs of these non-linked isolates, of the same RFLP type, had a SNP difference of zero. This suggests that amongst these pairs there may be previously un-detected transmission events. The range of date intervals between these pairs ranged from zero days to almost 5 years. In the absence of epidemiological evidence, and the low and variable mutation rate observed, it would be difficult to assess whether direct transmission has occurred in these cases, however this information would provide valuable evidence in a clinical setting, informing further investigation and contact tracing.

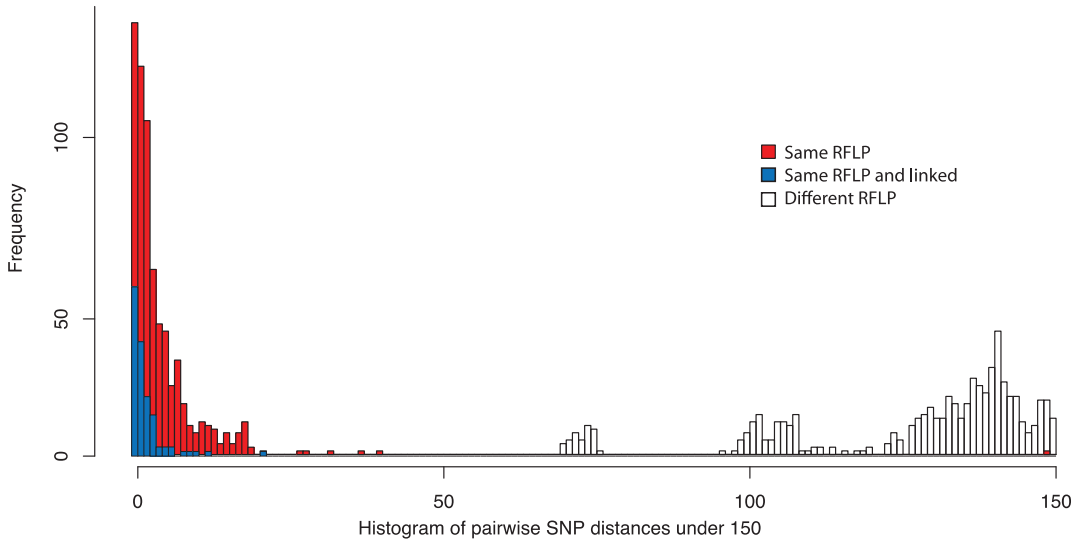

Figure 12 - Pairwise SNP differences between isolates. Only pairwise SNP distances under 150 are shown. There are many unlinked pairs that have SNP distances which overlap with the distribution of SNP distances for linked pairs.

2.5. Discussion

At the cluster level an average substitution rate of 0.3 SNPs per genome per year was estimated, which is remarkably close to estimates made in the macaque model (Ford,

Lin *et al*. 2011). This is also highly similar to the rate estimated in Oxfordshire of 0.5 (Walker, Ip *et al*. 2013), and subsequent work in Germany which concluded an average of 0.4 SNPs per genome per year (Roetzer, Diel *et al*. 2013) . This confirms the extremely low rate of accumulation of variation that characterizes *M. tuberculosis*, which is approximately 3 times and 44 times slower than that observed in *Escherichia coli* and *Staphylococcus aureus*, respectively (Didelot, Bowden *et al*. 2012).

This analysis was unable to detect a clock-like signal at a larger phylogenetic scale (at the lineage level), probably reflecting the different processes of fixation and substitution having variable influences on different parts of the evolutionary history. This confirms that the inter-patient substitution rate estimated here cannot be extended to deeper evolutionary histories, where the species-wide substitution rate is likely to be different. However, even at the intra-cluster level and between the epidemiologically linked pairs where we were able to estimate a rate, there was a large level of variation around the mean, which is in contrast to observations of other bacteria such as *Staphylococcus aureus* and *Vibrio cholerae* (Harris, Feil et al. 2010, Mutreja, Kim et al. 2011). There are a variety of factors that may have contributed to this noisy signal. Latency is common in tuberculosis infection and could result in considerable discrepancies in the apparent rate of substitution over time. However, work by Ford *et al*. showed that the substitution rate during latency in macaques was similar to that during active infection (Ford, Lin *et al*. 2011). Within host selection for factors such as drug resistance may also result in variation in the accumulation of SNPs over time. However, perhaps the most important factor is the low substitution rate itself, meaning over short time scales only a weak signal of a molecular clock can be detected.

This lack of a strong signal means that although a molecular clock may be detectable over longer time frames, it is only an aggregate measure and should be used with extreme caution when applying it to infer local transmission or date recent evolutionary events. Furthermore, while substitution rate can be used to strengthen or exclude epidemiological links, it cannot be used alone to infer direct transmission, particularly due its slow rate. There was no evidence of hyper-mutation in this dataset, and this has not been reported in clinical *M. tuberculosis* isolates to my knowledge.

However it is possible that treatment may impose selection pressures on isolates that could affect the observed rate of fixation, and this should be considered.

The Oxfordshire study concluded that a cutoff of ≤5 SNPs for cases less than three years apart may be appropriate for concluding transmission (Walker, Ip *et al*. 2013). This more in-depth analysis of the molecular clock reveals a lack of a strong signal with a high degree of variation around the mean rate, suggesting that using a simple cut-off may not be entirely appropriate for confirming transmission, but that the phylogenetic context provided by deep sampling of clusters may be more informative. This is highlighted by the fact that direct transmission between one epidemiologically linked pair separated by 5 SNPs was excluded, based on the presence of intersecting unrelated strains in the phylogeny (Figure 10a). Phylogeny and the context from other strains are important tools that can be used to further inform us on the likelihood direct transmission has taken place.

In summary, the slow molecular clock of of *M. tuberculosis* means that even at the highest resolution provided by whole genome sequencing it is still difficult to confidently affirm the inferences of transmission made by traditional epidemiological techniques. This means it is very difficult to determine transmission inclusively. However, whole genome sequencing does in some cases allow us to exclude direct transmission, by using the phylogenetic context provided by other strains. Understanding the limitations and strengths of this approach will be important for future clinical applications, and has also informed on the rest of work discussed in this dissertation.