3. Disentangling recurrent and mixed *Mycobacterium tuberculosis* infections

Based on work published in:

J. M. Bryant, S. R. Harris, J. Parkhill, R. Dawson, A. H. Diacon, P. van Helden, A. Pym, A. A. Mahayiddin, C. Chuchottaworn, I. M. Sanne, C. Louw, M. J. Boeree, M. Hoelscher, T. D. McHugh, A. L. C. Bateson, R. D. Hunt, S. Mwaigwisya, L. Wright, S. H. Gillespie and S. D. Bentley (2013). "Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study." The Lancet Respiratory Medicine. DOI: 10.1016/S2213-2600(13)70231-

And:

C. U. Koser, J. M. Bryant, J. Becq, M. E. Torok, M. J. Ellington, M. A. Marti-Renom, A. J. Carmichael, J. Parkhill, G. P. Smith and S. J. Peacock. (2013) "Whole-genome sequencing for rapid susceptibility testing of *M. tuberculosis*." The New England journal of medicine **369**(3): 290-292.

Statement of contribution:

REMoxTB study: I carried out all bioinformatics analyses and interpretation. The study was initiated by SHG and AHD, RD, AP, AAM, CC, IMS, CL, MJB and MH. TDMcH, ALCB, RDH, LW, SM, PvH, and SHG were responsible for microbiological design, conduct, and review of the study. Sequencing was managed by SDB, and SRH gave advice on bioinformatics.

XDR patient study: I carried out all bioinformatics analyses. CUK and SJP initiated the study. Sequencing was carried out at Illumina Ltd. (Cambridge) by JB and GPS. Drug susceptibility testing was carried out Public Health England National Mycobacterium Reference Laboratory (London) except for Clofazimine which was carried out by the Scottish Mycobacteria Reference Laboratory, Edinburgh.

3.1. Introduction

Recurrent tuberculosis infection is defined as a second episode of disease after successful treatment of a previous episode. Recurrence is low globally with the WHO reporting it for 5% of the 6.2 million tuberculosis cases in 2010 (WHO 2012). However, it has been well documented that in high incidence regions such as South Africa, recurrent tuberculosis is more dominant, and is associated with HIV status (Glynn, Murray *et al*. 2010). Recurrence can arise via two routes: relapse of the primary infection that treatment has failed to eradicate, and re-infection with an unrelated exogenous strain.

Until recently re-infection was considered to be rare, as a traditional assumption of tuberculosis epidemiology was that an infection episode is caused by a single strain and that subsequent episodes are caused by re-activation of the endogenous strain (Stead 1967). However there is an increasing appreciation that this is often not the case, and that both mixed infections and exogenous re-infection do frequently occur. This change in thinking is due to the development of genotyping techniques and their application to recurrent tuberculosis disease in a clinical setting, which makes it possible to distinguish if the primary and secondary disease episodes were caused by the same genotype. In a study in India, it was estimated that 88% and 9% of recurrence cases were due to re-infection in HIV positive and negative patients respectively (Narayanan, Swaminathan *et al*. 2010). In South Africa it was noted that the incidence of re-infection was higher than the incidence of new infections, where 77% of recurrence was classed as re-infection (Verver, Warren *et al*. 2005). Similarly mixed infections have also been found to be more common than first thought, with one study in South Africa finding at least two different strains in 19% of patient samples (Warren, Victor *et al*. 2004).

Although these typing techniques have been useful in revealing the possible extent of mixed and re-infections, they can lack resolution as discussed more generally in section 1.4.3. Further to this, mixed samples can be very difficult to detect using traditional techniques, as the signal can be unclear or undetectable if one of the strains is present in too low quantities, or are too similar. This impacts on our understanding of recurrent disease as it would be difficult to disentangle complex scenarios such as

an apparent re-infection which may in reality be a mixed infection followed by endogenous re-activation of one of the strains. The high depth of coverage that can be obtained with whole genome sequencing should allow the detection of mixed infections, and allow us to pick up on these scenarios more accurately.

Here, two studies are presented which both use whole genome sequencing to disentangle the different routes that can result in multiple infections and disease episodes of tuberculosis. The first is based on pairs of samples collected from patients with recurrent disease during a multi-centre clinical trial, REMoxTB. The second is based on a single patient from Addenbrooke's hospital diagnosed with XDR tuberculosis.

3.2. Methods

3.2.1. REMoxTB study

REMoxTB was a phase three clinical trial that aimed to test two four-month moxifloxacin containing regimens compared to standard treatment. 1,931 patients underwent randomised treatment across sites in South Africa, India, Tanzania, Kenya, Thailand, Malaysia, Zambia, China and Mexico (Gillespie, Crook et al. 2014). At the time of the analysis the trial was still ongoing and researchers were blinded to the treatment regimen. The first 50 paired isolates available from participants enrolled in the trial were used: composed of the initial sample upon diagnosis and a post week 17 of treatment sputum sample from patients with relapse or bacteriological failure. Eligible patients were adults diagnosed with previously untreated, drug-sensitive, smear-positive, pulmonary tuberculosis without severe co-morbidities. HIV-positive patients with a CD4-count below 250/µl or those already on antiretroviral treatment were excluded. All subjects providing informed consent were treated for tuberculosis for 26 weeks with one of three different regimens of 4 or 6 months duration that could contain rifampicin, isoniazid, ethambutol, pyrazinamide, moxifloxacin and/or placebo. The total observation period including treatment and follow-up was 18 months.

To distinguish cases due to treatment failure and those resembling recurrent disease, the complete clinical history was reviewed (carried out by A. Bateson, University College London), thereby taking into account all culture results and all clinical

information available. Single isolated positives were also included in order to investigate their clinical relevance, as this is currently unclear. These are cases where a positive culture was followed by at least two negative cultures without re-treatment having been initiated by a physician and the patient remaining symptom free throughout the remainder of follow-up.

Both DNA extraction and MIRU VNTR were performed by R. Hunt and A. Bateson (University College London). MIRU-VNTR typing analyses the number of repetitive DNA sequences at multiple independent genetic loci (ETR-A, B, C, D, E and MIRU-02, 10, 16, 20, 23, 24, 26, 27, 39, 40) as described previously (Supply, Allix *et al*. 2006).

Samples were pair-end sequenced with a read length of 100bp on the Illumina HiSeq platform. The raw sequencing data was mapped to H37Rv and variant calling was carried out as described in the Methods 8.2 and 8.3. Mixed based calls were detected as described in Methods 8.9.

3.2.2. XDR patient study

Sputum specimens taken at the Cambridge University Hospital were processed by laboratory staff at the Cambridge Public Health England Microbiology Laboratory. DNA was extracted by Claudio Köser (University of Cambridge) from one half of a Mycbacterial growth indicator tube (MGIT) culture grown from the first sputum specimen obtained on admission to Cambridge University Hospital. DNA was also extracted from *M. tuberculosis* grown from subculture of the MGIT tube onto a Löwenstein–Jensen (LJ) slope. Library preparation and DNA sequencing (paired-end, 150 bp reads, Illumina MiSeq platform) were performed by Illumina Cambridge Ltd. Mapping and variant calling were carried out as described in Methods 8.2 and 8.3. Mixed base calls were detected as described in Methods 8.9.

3.3. Results – REMoxTB

3.3.1. Overview

Paired samples from 50 patients were sequenced (see Appendix 9.2 for meta-data on pairs). For 96 of the samples (representing 47 patient-pairs plus two singletons where one sample of the pair failed to sequence) an average coverage of 120 fold was obtained, with the remaining four excluded due to poor coverage or contamination with a non-mycobacterial source. Based on the 10,354 variable positions detected, a maximum likelihood phylogeny was built revealing the presence of four of the globally recognized lineages (Gagneux, DeRiemer *et al*. 2006) (Figure 13).

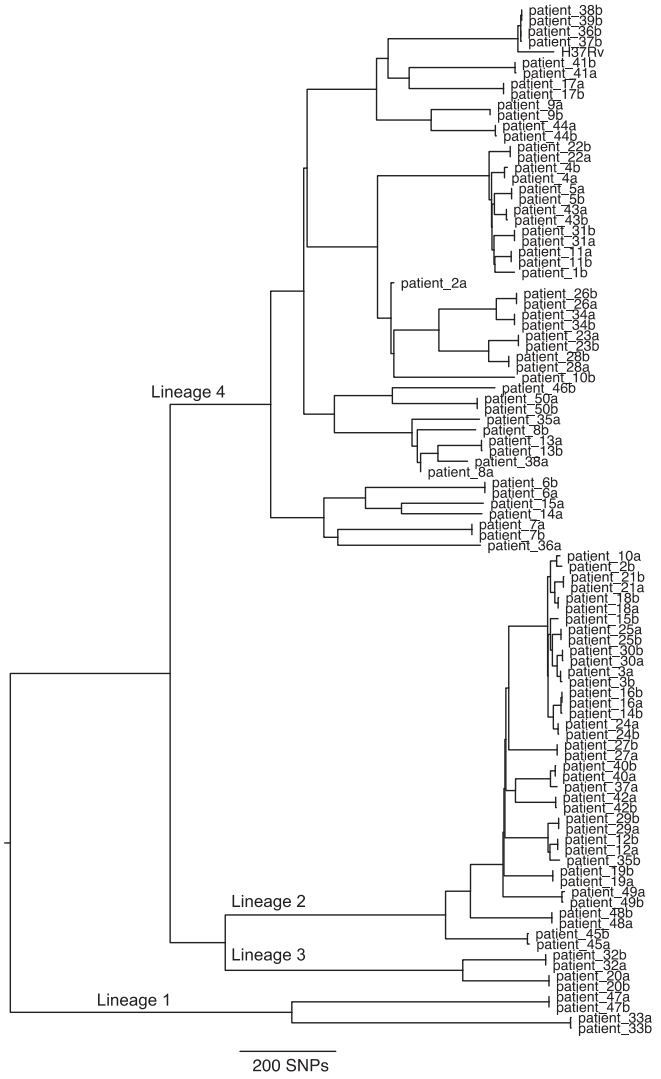

Figure 13 – Maximum likelihood tree of all successfully sequenced isolates in the dataset. Four of the major lineages of the MTBC (Gagneux, DeRiemer *et al*. 2006) are marked. Sample 2a and 8a sit close to internal nodes – they were later found to be a mix of two strains.

Using the observed SNPs between the initial and recurrence strains, cases were defined as relapse (n=33), re-infection (n=3) or mixed infection (n=6) (Figure 14). The rationale behind making these designations are discussed below.

3.3.2. Distinguishing relapse and re-infection

There was a clear distinction between pairs with a low SNP difference (<=6), and those with a high SNP difference (=>1306) (Figure 15). Previously it was observed that within-patient diversity didn't exceed 14 SNPs (Walker, Ip *et al*. 2012), which supports the inference that the low SNP distance pairs represent relapse, and the high SNP distance pairs represent re-infection. Relapse was identified in 33 cases (70% 33/47) with pairs differing by a mean of 0.4 SNPs and the majority (n=27) having no polymorphisms. For three pairs (7%), which all had SNP differences greater than 1306, their recurrence stain was defined as a re-infection. All three involved isolates belonging to different lineages: either the Euro-American or East Asian type. The mean SNP distance between the re-infection pairs was 1355 (Figure 15) which is significantly larger than the mean pairwise distance observed between all isolates in the dataset (972), when compared using the Wilcoxon test (P=0.044).

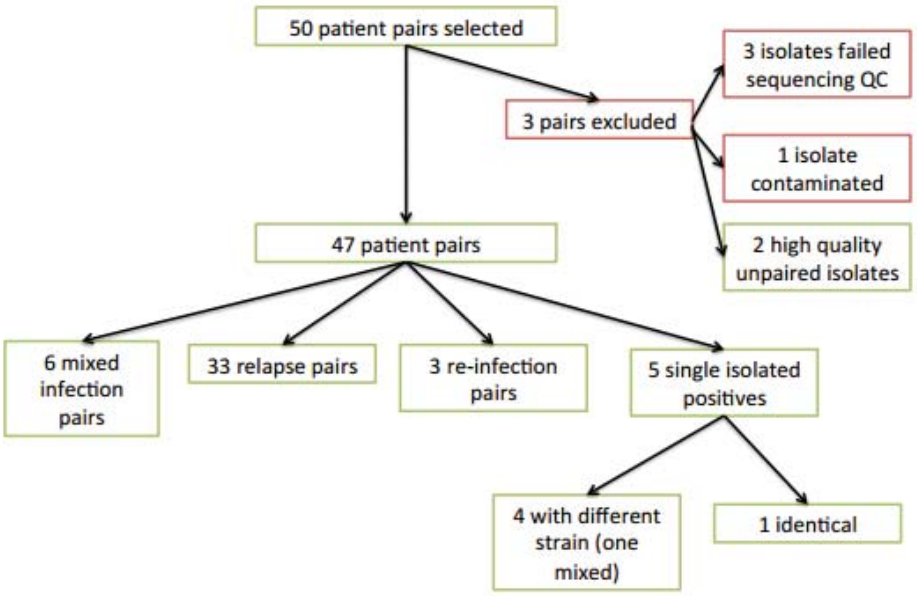

Figure 14 - Summary of sequencing results. Green boxes indicate isolates included in the analysis, red were excluded due to sequencing failure or contamination.

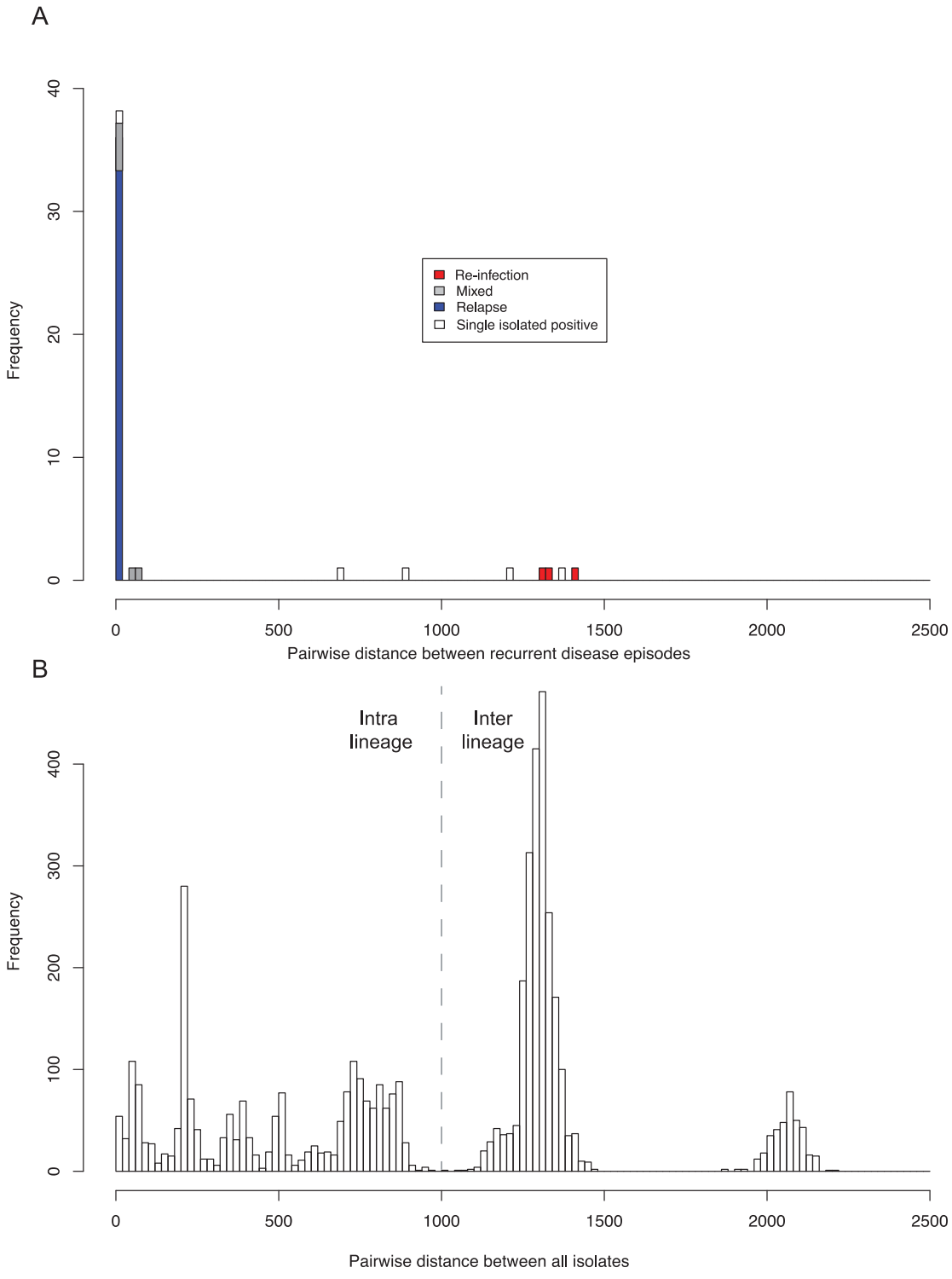

Figure 15 - Histograms of genome-wide pairwise SNP distance between isolates in the dataset. A) Pairwise difference between same patient pairs. B) Pairwise differences between all isolates in dataset, with a dotted line representing the point where the difference represents the distance between two lineages.

3.3.3. Mixed infections

For the majority of samples (n=87) fewer than 40 sites with a mixed base call were identified across the genome, and these were likely to be due to mapping error. In seven sample pairs, however, there were outliers with more than 80 sites (Figure 16) which were manually inspected to look for mixed base calls at lineage defining positions (Stucki, Malla et al. 2012), or where SNPs had been identified in the other isolate of the pair. A total of six patients showed evidence of a mixed infection of which four were mixed in the first sample of the patient pair. They were found to be heterogeneous in positions where a SNP was identified in the second sample, indicating that the initial sample was composed of the strain found only in the secondary isolate plus a sequence from another lineage. Two patient pairs had evidence of two distinct strains only in the secondary isolate (Table 2), one of which was the same strain found in the initial sample, which could be interpreted as relapse and super-infection. An additional sample was also found to have evidence of a mixed population, but was defined clinically as a single isolated positive (see below).

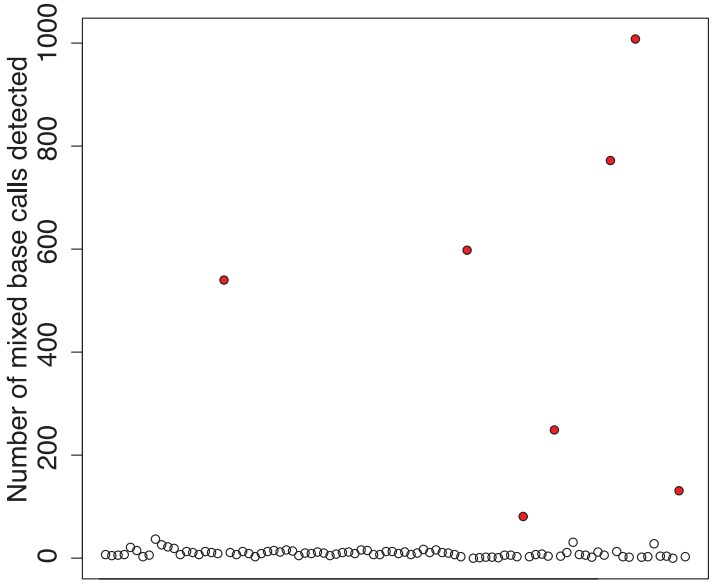

Figure 16 - Number of mixed base calls identified for all isolates in the study. Each dot represents a isolate, arranged in a random order along the x-axis. Red dots represent those identified as mixed.

Table 2 - Proportion of reads matching lineage defining SNPs identified in the mixed infections
Lineage specific SNPs were identified using informative positions previously defined (Stucki, Malla *et al*. 2012) Frequencies represent the proportion of reads that match the base that defines the lineage. * Sample 8a is composed of two Euro-American strains divergent by at least 132 SNPs in a 50% mix ** Sample 42b is composed of a Typical Beijing isolate identical to 42a (95%) plus an Atypical Beijing strain (Schurch, Kremer *et al*. 2011) (5%). Manual inspection of 42b also reveals reads matching the Atypical strain (~2%). The mixed sample from a single isolated positive was excluded.

Patient sample	1. Indo oceanic 3920109 (G->T)	2. East Asian 1834177 (A->C)	3. East African Indian 301341 (C->A)	4. Euro American 3326554 (C->A)	5. West African 1 1377185 (C->G)	6. West African 2 2427828 (C->G)
2a	0	0.16	0	0.84	0	0
2b	0	1	0	0	0	0
8a	0	0	0	1*	0	0
8b	0	0	0	1	0	0
23a	0	0.39	0	0.7	0	0
23b	0	0	0	1	0	0
42a	0	1	0	0	0	0
42b	0	0.96**	0	0	0	0
45a	0	1	0	0	0	0
45b	0	0.93	0	0.08	0	0
50a	0.26	0	0	0.69	0	0
50b	0	0	0	1	0	0

3.3.4. Single isolated positives

Cases were defined clinically as single isolated positives on five occasions. These are incidences where a single sample is found to be sputum positive for *M. tuberculosis*, and in the absence of treatment all subsequent samples are negative. These are usually attributed to lab cross-contamination. Out of the five cases, three of them were with a strain unrelated to the primary case (>500 SNPs), one was mixed and one differed by only three SNPs. The small SNP distance in the latter suggests that this case represents a true relapse and not contamination.

3.3.5. Correlation with MIRU-VNTR data

MIRU-VNTR, one of the most commonly used typing techniques, was carried out on all samples. The three cases identified as re-infection by whole genome sequencing differed by 1-13 loci. Twenty-seven of the relapse cases had an identical MIRU-VNTR type, but five differed by one or more loci. There were six cases identified by

genome sequencing as possible mixed infections but MIRU/VNTR identified four of these as re-infections and two as relapse.

3.4. Results – XDR patient

Two samples were sequenced from a male patient diagnosed with XDR tuberculosis at Addenbrooke's hospital. They were isolated from different culture techniques: first from a MGIT tube, and the second from an LJ slope which are both standard techniques used to select for and grow mycobacteria. When mapped against the *M. tuberculosis* reference genome, a high number of mixed base calls (n=421, raw unfiltered) were called in the MGIT sample but not the slope sample. Manual inspection of these positions revealed that there was an apparent mixture of an isolate that was highly similar or identical to the slope sample, together with another isolate in a ratio of approximately 70:30. The mapping data was filtered for high quality mixed base positions (n=224) and the alleles were separated into two by sorting the alternative alleles for each position into those that matched the slope sample and those that didn't. This enabled the mixture to be separated into "slope-like" and "non-slope-like" as shown Figure 17.

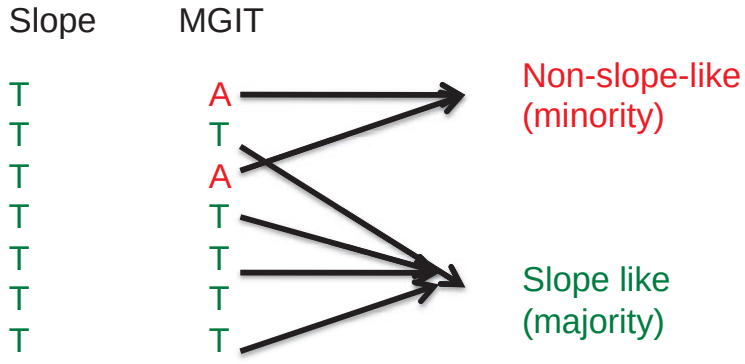

Figure 17 - Rationale used to separate mixed MGIT sample from the XDR patient

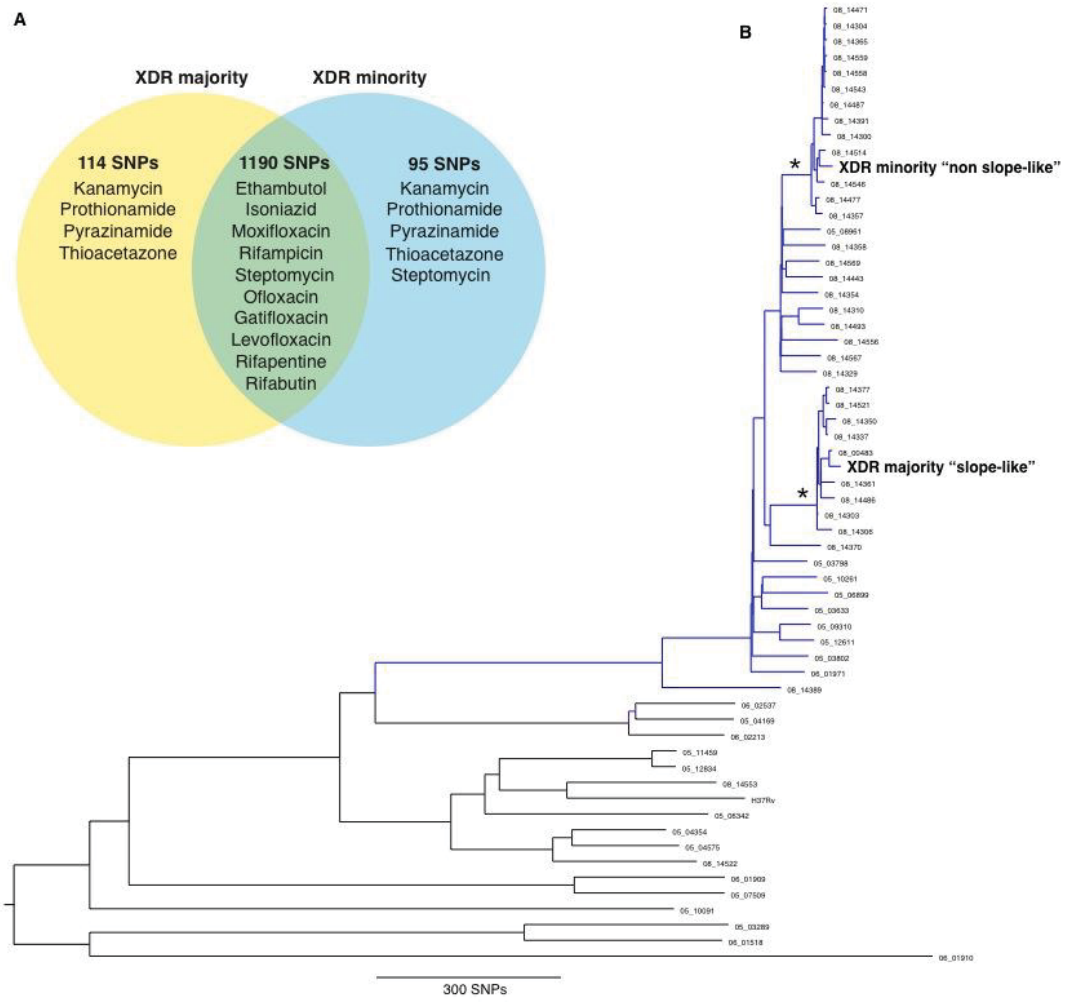

Figure 18 - A mixed extensively drug resistant (XDR) infection. A) Antibiotics with evidence of resistance mutations present in majority and minority strains. The same mutation was present in both strains for ten drugs (green intersect), but different mutations in each strain accounted for resistance to five drugs (yellow and blue). Streptomycin is listed twice as an additional resistance associated variant was found in the XDR minority B) Maximum likelihood tree showing the phylogenetic position of the XDR minority and majority strain in the Beijing (blue) lineage. Contextual strains were from Samara, Russia (Casali, Nikolayevskyy *et al*. 2012). The bootstrap support for the blue clade, and the two clades containing the separated strains (node marked with *) were all 100%

The presence of lineage determining SNPs (Stucki, Malla *et al*. 2012) suggested that they both belonged to the Beijing lineage of *M. tuberculosis*. To place the two strains in context, a phylogeny was built using the Beijing lineage samples from a previous study of tuberculosis in Samara, Russia (Casali, Nikolayevskyy *et al*. 2012). The two strains were found to be paraphyletic (Figure 18b); confirming that this mixture was not a result of diversity generated during an infection.

Drug susceptibility testing concluded that the sample was XDR. However this observed phenotype could be the result of just one of the strains in the mixture, both or a combined result of resistance phenotypes contributed by both strains. For some of the tested antibiotics, resistance mechanisms have been well characterised; for others, very little is known. Possible genes (and in some cases specific codons) associated with resistance to the tested antibiotics were identified through an extensive literature search (carried out by C. Köser). These sequences were then checked for the presence of variants with respect to the H37Rv reference (which is fully susceptible to the drugs of interest). If a possible resistance-causing variant was identified it was called as XDR majority if it matched the variant called in the slope sample (yellow –Figure 18a), XDR minority if it didn't match (blue), and both if present in 100% of the reads (green). This confirmed that both strains in the mixture were XDR, and were found to be resistant to the same antibiotics but were due to different independent mutations in five cases.

3.5. Discussion

A traditional assumption of tuberculosis research is that an infection episode is caused by a single strain and that subsequent episodes are caused by re-activation of the endogenous strain (Stead 1967). However, most tuberculosis clinicians and researchers now appreciate that this can often not be the case, and that a number of different scenarios could be underlying a disease episode. This is particularly true for endemic regions, such as South Africa, where HIV may be a driving force and Eastern Europe, where poor infection control and treatment failure may be resulting in multiple infections. The fact that super-infections (resulting in mixed infections) and new infections (resulting in re-infections) can occur in non-HIV positive individuals suggests that the immune protection conferred by the first infection may not always be strong or durable enough to protect against subsequent infections. Immunity to tuberculosis is poorly understood, but we can speculate that this could be due to either the diversity of the host immune response, or diversity of the pathogen.

There is a possibility that the immune protection conferred by one tuberculosis strain may not extend to more distantly related strains due to differences in their antigenic profile. Hints that this may be the case come from studies of the Bacillus Calmette–

Guérin (BCG) vaccine, a live attenuated form of *M. bovis*. Estimates of its efficacy have varied wildly, ranging from 0 – 90% (Fine 1995), and have been attributed to a variety of factors including host diversity and exposure to NTMs. However, these studies lack knowledge of the prevalence and diversity of circulating *M. tuberculosis* strains, leaving this diversity unaccounted for. Interestingly both animal (Lopez, Aguilar et al. 2003, Tsenova, Harbacheuski et al. 2007) and human (Kremer, van-der-Werf *et al*. 2009) studies have both supported the provocative idea that the highly successful Beijing lineage may represent a BCG vaccination escape variant (Abebe and Bjune 2006). This strain-specific variation in the efficacy of BCG suggests that different lineages of *M. tuberculosis* may confer differential immune protection. Was there any evidence for this in the re-infection cases in this dataset? Unfortunately, three cases are not enough to make any robust conclusions. But it's noteworthy that all the re-infection cases were with strains from a different lineage, and that the SNP difference was significantly larger than would be expected by chance if re-infection were equally likely for all strains in the dataset. In another study focusing on applying whole genome sequencing to transmission chains in Uganda, two re-infection cases were identified. One of these involved two strains from the same lineage (lineage 4), and the other was with strains from two different lineages (lineage 4 and 3) (Clark, Mallard *et al*. 2013). Clearly further studies on larger datasets will be required to address this question, which may have important consequences for vaccine design.

Future studies will need to use whole genome sequencing to accurately distinguish the scenarios of relapse, re-infection and mixed infections. This is reflected by the fact that this study found that 11/47 cases came to different conclusions than those using the MIRU-VNTR data. In the context of a clinical trial, this means that 6/33 cases were misclassified as relapse: the primary end-point in a clinical trial. This high level of misclassification could also impact on our understanding of the prevalence of these scenarios. Previous studies have used a cut-off of greater than one locus to conclude re-infection (Narayanan, Swaminathan et al. 2010, Martin, Herranz et al. 2011), and this would have resulted in the misclassification of two of the relapse pairs (differing by 2 and 3 loci), which means that re-infection may have been over estimated in these cases.

One major limitation of all genotyping techniques, including whole genome sequencing when attempting to classify recurrent disease, is that an apparent relapse may be due to re-infection with a closely related strain (from a family member for example). It's not known how often this occurs, and would be impossible to estimate using the approach described here. In order to accurately quantify the rates of these two processes, future analyses may need to incorporate modeling approaches and epidemiological information collected from patients.

The detection of mixed infections is important for individual patient management in addition to increasing our understanding of tuberculosis epidemiology. The XDR case described here demonstrates that different infecting populations in the same patient can have different resistance profiles, and that whole genome sequencing provided clarity in this respect. Mixed infections are expected to be more difficult to treat and more likely to lead to acquired resistance (Cohen, van Helden *et al*. 2012). Furthermore, miss-identification of a mixed infection could lead to errors during epidemiological investigation, when failure to detect both strains in the index case could lead to failure to define a transmission event to secondary cases. On the population level, mathematical models predict that a high preponderance of mixed infections will lead to the survival of less fit strains, which will persist longer than they would in the absence of mixed infections (Cohen, van Helden *et al*. 2012). More complex models also predict that if mixed infections were common, control interventions that target latent infection (such as isoniazid preventative therapy) would be more likely to lead to the emergence of drug resistant strains (Colijn, Cohen *et al*. 2009).

This study identified six mixed infections, despite the bacteriological methods being orientated towards the isolation of a single strain, suggesting that this is an underestimate of the real burden. Previous estimates of the prevalence of mixed infections were based on genotyping techniques, and were often limited to one sample per patient. A study in Georgia, found that out of the 26 mixed infection cases that were identified using genotyping of multiple samples, all or 14 (RFLP typing or PCR respectively) of them would have been missed entirely based the analysis of a single pre-treatment isolate (Shamputa, Jugheli *et al*. 2006). This demonstrates that to really understand the prevalence of mixed infections in different settings, whole genome

sequencing will need to be carried out on multiple colonies or non-colony-purified cultures of multiple samples. The XDR study also highlights the possibility that laboratory handling of samples may result in selection for one of the strains, as the mixture was only identified in the MGIT sample, and not the LJ slope sample.

Of the five cases identified as single isolated positives in this study, four were likely due to cross contamination and one provides evidence for the first time, that positive cultures originating from the patient's own infection may be cultured and the patient's infection resolved without further treatment. Cross contamination is a well-recognised challenge in myco-bacteriology laboratories, accounting for up to 3.9% of samples (Glynn, Yates *et al*. 2004). It usually occurs in less than 1% of positive samples with more than half of laboratories achieving a rate of less than 2.5% (Ruddy, McHugh *et al*. 2002). In the clinical trial setting there is a need to ensure that adequate molecular methods are in place to identify the origin of single isolated positive samples correctly.

The ability to accurately distinguish relapse, re-infection and mixed infections is of critical importance for an understanding of tuberculosis epidemiology, determining end points in clinical trials and for patient management. This study provides a proof of principle demonstrating that whole genome sequencing can distinguish these different scenarios unequivocally. Larger scale studies will now be required in order to quantify these processes in different geographical, clinical and social contexts.