

Use of whole-genome sequencing to distinguish relapse from reinfection in a completed tuberculosis clinical trial

Adam A. Witney^{1*}, Anna L.E. Bateson², Amina Jindani¹, Patrick P.J. Phillips³, David Coleman¹, Neil G. Stoker¹, Philip D. Butcher¹, Timothy D. McHugh², RIFAQUIN Study Team.

¹ Institute for Infection and Immunity, St George's University of London, London, UK

² UCL Centre for Clinical Microbiology, Royal Free Campus, UCL, London, UK

³ MRC Clinical Trials Unit at UCL, London, UK

* Corresponding author

Adam A. Witney	awitney@sgul.ac.uk
Anna L. E. Bateson	a.bateson@ucl.ac.uk
Amina Jindani	ajindani@sgul.ac.uk
Patrick P. J. Phillips	patrick.phillips@ucl.ac.uk
David Coleman	dcoleman@sgul.ac.uk
Neil G. Stoker	stoker.neil@gmail.com
Philip D. Butcher	butcherp@sgul.ac.uk
Timothy D. McHugh	t.mchugh@ucl.ac.uk

Abstract

Background

RIFAQUIN was a tuberculosis chemotherapy trial in southern Africa including regimens with high-Dose rifapentine with moxifloxacin. Here, the application of whole genome sequencing (WGS) is evaluated within RIFAQUIN for inferring new infections in treated patients as either relapses or reinfections, and is further compared with MIRU-VNTR typing. This is the first report of using WGS to evaluate new infections in a completed clinical trial where all treatment and epidemiological data were available for analysis.

Methods

DNA from 36 paired samples of *Mycobacterium tuberculosis* cultured from patients before and after treatment were typed using 24 locus MIRU-VNTR, *in silico* spoligotyping and WGS. Following WGS, the sequences were mapped against the reference strain H37Rv, the single nucleotide polymorphism (SNP) differences between pairs identified and phylogenetic reconstruction performed.

Results

WGS indicated that 32 paired samples had a very low number of SNP differences (0-5; likely relapses). One pair was a likely mixed infection with a pre-treatment minor genotype which was highly related to the post-treatment genotype, this was reclassified as a relapse, in contrast to the MIRU-VNTR result. The remaining three pairs had very high SNP differences (>750; likely reinfections).

Conclusions

WGS and MIRU-VNTR both similarly differentiated relapses and reinfections, but WGS provided significant extra information. The low proportion of reinfections seen, suggests that

in standard chemotherapy trials with up to 24 months of follow-up, typing the strains brings little benefit in terms of the trial outcome purely in terms of differentiating relapse and reinfection. However, there is benefit in using WGS as compared to MIRU-VNTR in terms of the added genotype information obtained, in particular defining the presence of mixed infections and the potential to identify known and novel drug resistance markers.

Keywords: Whole genome sequencing, tuberculosis, clinical trial

Background

Evaluations of drug trials for tuberculosis (TB) are complicated by the fact that a recurrence of disease can either be due to endogenous relapse of disease or to subsequent exogenous infection with a new strain (reinfection). Historically, during the major tuberculosis chemotherapy trials of the 1960s-1980s (reviewed by Fox *et al* (1)), it was not possible to differentiate isolates, and all new infections that occurred after the trial conclusion were labelled as relapses.

From the 1980s, a series of genomic-based methods for typing strains of *Mycobacterium tuberculosis* were developed, in particular IS6110 RFLP, spoligotyping and MIRU-VNTR typing (2–4). Some trials therefore began to use molecular methods to differentiate relapses from reinfections. This was initially through IS6110 RFLP typing (5–7) and then MIRU-VNTR typing (8), while others continued without any differentiation (9).

MIRU-VNTR became the favoured typing approach because it combined reasonable discrimination with a readout that could both be easily measured, and also be described in a digital form (3). More recently, whole genome sequencing (WGS) has enabled the identification of single-nucleotide polymorphism (SNP) differences, thus leading to far greater discrimination in TB epidemiological studies (10–13).

Two groups have recently used WGS to evaluate paired samples, comparing SNP differences between the original infections and new infections following treatment (14,15). The study by Bryant (14) was based on an ongoing clinical trial (16) which was being carried out in sub-Saharan Africa, south and east Asia, and central America. 33 out of 36 paired were found to be highly similar (≤ 6 SNPs; classed as relapses) and 3 highly divergent (≥ 1306 SNPs; classed as reinfections).

The report by Guerra-Assunção (15) was not based on a clinical trial, but was taken from the Karonga Prevention Study, a long-term population based programme in Malawi. 60 paired samples collected over a 15 year time period were sequenced, and while they also found a clear division in SNP numbers between relapses and reinfections, it was not as marked as in the Bryant study. Thus, they classed 46 samples with 0-8 SNP differences as relapses, and 14 with >100 SNP differences as reinfections.

In this study WGS was performed and SNPs analysed comparing pre- and post-treatment isolates from the completed RIFAQUIN clinical trial (17), a study evaluating high dose rifapentine with moxifloxacin, carried out in sub-Saharan Africa. 36 pairs of samples of *M. tuberculosis* recovered before treatment and from those patients showing positive cultures at 6 months were successfully sequenced, and compared with MIRU-VNTR data. The results agree with general findings from the two studies referred to above, that the overwhelming majority of secondary cases are relapses. Importantly, WGS was further able to monitor possible epidemiological connections and sample errors during the trial, which were not detected using MIRU-VNTR. Given the added benefit of WGS in this context, it is suggested that WGS should be routinely used as the method of choice in such trials.

Methods

RIFAQUIN trial

The RIFAQUIN chemotherapy trial, in collaboration with six institutions in southern Africa, has been previously described (17). Between August 2008 and August 2011, patients with newly-diagnosed smear-positive drug-sensitive tuberculosis were randomly assigned to either:

Control regimen: 2 months of daily ethambutol, isoniazid, rifampicin, and pyrazinamide followed by 4 months of daily isoniazid and rifampicin;

4-month regimen: Isoniazid replaced by moxifloxacin daily for 2 months followed by 2 months of twice-weekly moxifloxacin and 900mg rifapentine; or

6-month regimen: Isoniazid replaced by moxifloxacin daily for 2 months followed by 4 months of once-weekly moxifloxacin and 1200mg rifapentine.

Sputum was examined by microscopy and culture at regular intervals for treatment failure or relapse. Patients had up to 18 months follow-up post randomisation, with the patients recruited last having follow-up to 12 months post-randomisation. Samples from patients with two or more consecutive *M. tuberculosis* positive cultures after six months (or end of treatment) were selected for WGS.

MIRU-VNTR determination and assignment

The 24 loci MIRU-VNTR typing of these isolates was previously described (17). Briefly, a 10µl loop was used to pick up a sample of *M. tuberculosis* colonies by sweeping across growth on a Lowenstein-Jenson (LJ) slope. Bacteria were heat-killed and DNA extraction performed using lysozyme and proteinase K digestion followed by phenol-chloroform extraction and ethanol precipitation (18). The 24 MIRU-VNTR loci were amplified in 8 labelled multiplex PCR reactions, and the amplicons sized, with MapMarker 1000 standard (Biosciences), by capillary electrophoresis on the sequencer (3130 Genetic Analyzer, Applied Biosystems). Analysis was carried out using the GeneMapper software (Applied Biosystems), which assigned alleles based on the customised bin-sets (fragment sizes and dyes) used to define each allele. For some samples there was variable coverage across the MIRU-VNTR loci using the sequencer, so where possible, any missing loci were confirmed by single-plex PCR with products sized by standard agarose gel electrophoresis. Where possible paired samples (pre- and post-treatment) from a given patient were run in parallel.

Whole Genome Sequencing

50µL, containing at least 250ng, genomic DNA from each sample was sheared using the Covaris E220 for a target size of 200bp (PIP: 175, Duty Factor: 10%, Cycle/burst: 200, Temp: <8°C Time: 120s). Libraries were prepared from sheared DNA using the NEB DNA Ultra kit in accordance to standard protocol (New England Biolabs, UK). The NEB adapters were substituted for the set described by Kozarewa *et al* (19). Libraries were quantified using the Qubit High Sensitivity DNA assay and pooled equimolarly (Invitrogen, UK). The pools were subjected to paired-end sequencing carried out on a single lane of the Illumina HiSeq 2500 (v3 chemistry, read length 100bp). Samples which produced low-yield were re-pooled and sequenced on a single MiSeq run (v2 chemistry, read length 250bp).

Sequence analyses

Sequence reads were mapped to the H37Rv reference genome (RefSeq accession: NC_000962) using bwa mem v0.7.3a-r367 (20), alignments sorted and duplicates removed with samtools v0.1.19 (21). Site statistics were generated using samtools mpileup and variant sites filtered based on the following criteria: mapping quality (MQ) above 30, site quality score (QUAL) above 30, at least 4 reads covering each site with at least 2 reads mapping to each strand, at least 75% of reads supporting site (DP4), allelic frequency (AF1) of 1. Sites which failed these criteria in any isolate were removed from the analysis. Phylogenetic reconstruction was performed using RAxML v8.2.3 (22) with a GTR model of nucleotide substitution and a GAMMA model of rate heterogeneity, branch support values were determined using 1000 bootstrap replicates. Relapse or reinfection calls were made by applying the above filtering criteria to the individual patient paired samples. INDELS were identified using samtools mpileup as above, but setting the minimum fraction of gapped reads for candidates to 0.05.

In silico spoligotyping and sub-lineage typing

Spoligotypes were generated using SpolPred (23). Sub-lineages were further determined using the presence or absence of a set of 62 lineage-defining SNPs as derived by Coll *et al* (24).

Mixed infections

For each isolate sequence a count of the percentage of reads supporting a variant base at each genome position was plotted. Mixed isolates can be identified by the presence of an extra peak suggesting the presence of two genotype populations in the sequenced sample. Base calls for the majority and minority strain were separated based on the percent reads and pseudo-sequences generated and subsequently included in phylogenetic reconstruction as above.

Results

Samples studied

Figure 1 shows a flowchart of the samples studied. A total of 827 patients, with newly diagnosed, microscopy positive pulmonary tuberculosis were enrolled in South Africa, Zimbabwe, Botswana and Zambia in the trial. Fifty-one patients had positive cultures in post-treatment follow up and therefore required genotyping to distinguish relapse and reinfection (as per the RIFAQUIN protocol (17)). DNA was available to generate MIRU-VNTR data for 44 pairs of samples (pre- and post-treatment). The DNA remaining was passed for WGS, and good quality sequence (>20x coverage) was generated for both pre- and post-treatment samples of 36 patients.

SNP differences were determined between the pairs of isolates, and a comparison with MIRU-VNTR differences is shown in Table 1. Two main groups can be identified: 32 pairs of isolates

had five or fewer SNPs, four pairs of samples had much higher numbers of SNPs (range 737-1329); a single pair of isolates differed by 57 SNPs, but this was probably because the pre-treatment isolate contained a mixed infection, as discussed below.

Phylogenetic reconstruction of SNPs

Phylogenetic reconstruction of variant SNPs (Figure 2A) showed that the majority (32/36) of the isolate pairs have low numbers of SNP differences and were therefore clearly determined cases of relapse. One isolate pair was identified as a mixed infection (see below). The remaining three isolate pairs that had high numbers of SNP differences appear quite divergent on the tree (marked in green) and were determined as likely reinfections.

There were also isolates that mapped closely to other patient isolates on the tree, and these merited closer attention to see if there were genuine connections or unexpected problems caused by possible laboratory handling errors.

Figure 2, panels B and C show one class of pattern that was observed with clustered isolate pairs, where there are no SNP differences between each member of a pair, but each pair was very closely related to another pair. In both panels, the two pairs of samples came from different centres (panel B: 005 and 014; Harare and Marondera, both in Zimbabwe; panel C: 008 Harare, Zimbabwe and 001 Francistown in, Botswana on the borders of Zimbabwe (Table 2) suggesting a laboratory processing error was unlikely; alternatively highly similar local strains were circulating in the two relatively close regions and evolving independently over time.

Panels D and E show a different type of pattern, where a pair of isolates from one patient clustered together, as expected for relapses, but was also identical to a single isolate from another pair, suggesting a possible transmission event. In Figure 2D, a post-treatment sequence for isolate 009 was identical to isolate pair 012; the two 009 isolates differed by

1233 SNPs. In Figure 2E, a pre-treatment isolate 004-1 was identical in sequence to both isolates of patient 003; the two 004 isolates differed by 737 SNPs. All four patients received treatment in the same city, Harare (Table 2). While it is not impossible that these genotypes were genuinely isolated from the two patients, 009 and 004, another possible explanation is some form of laboratory processing error. Indeed, in one case the patients visited the hospital on the same day, and in the other results were reported at the same time. This combined with their geographical co-location, would further support the possible processing error interpretation. It is also worth noting that if indeed these are in fact errors, they would normally be invisible to the analysis without the resolution of WGS.

Mixed infections

One patient's pair of samples (035) displayed 57 SNPs between the pre (035-1) and post-treatment (035-2) isolates and was therefore initially classified as a reinfection. However, further analysis of the WGS data showed evidence of a mixed infection in the pre-treatment isolate (035-1; Figure 3A) corresponding to an approximately 75%/25% combination of two genotypes. Using this majority/minority ratio of read coverage, it was possible to separate the two genotypes and further phylogenetic reconstruction suggested that it was likely that the minority genotype (035-1-min) was closely related to the post treatment isolate (035-2; Figure 3B; Figure 2A), suggesting that this was in fact a case of a relapse of a previously unidentified minority genotype, rather than a reinfection as previously assigned.

Initially there appeared to be 57 SNP differences between the pre and post-treatment isolates (035-1, 035-2), which would have been an unusual result as the previous studies had only identified reinfections with very high SNP differences, and nothing at an intermediary level. The observation of mixed genotypes would explain this discrepancy as one of the main filtering criteria in the site-calling algorithm is to remove sites with mixed genotype calls (<75% read support for the call), so the real number of SNP differences between the isolates is likely to be

higher. After separating the genotypes it was estimated that the number of SNP differences between the pre-treatment minority genotype and the post treatment isolates was 869 SNPs. The pre-treatment minority genotype and the post treatment isolate appeared to differ by 245 SNPs; however the genotype separation algorithm used was relatively crude, filtering based on parameter cutoffs, so it was not possible to completely separate the genotypes at all mixed genome sites, reflecting the overlapping shape of the two distributions (Figure 3A). However, the proximity of their placement on the tree (Figure 2A) suggests they are highly related and thus this patient was a likely relapse.

Comparing WGS with MIRU-VNTR data

Figure 4A shows there is a stark difference in the number of SNP difference between cases of relapse and reinfection, an observation also made by Bryant *et al* (14). Table 1 and Figure 4B show the distribution of MIRU-VNTR differences. The majority of pairs have no MIRU-VNTR differences (out of up to 21 loci determined), but some had a maximum of 7 loci different. We experienced technical difficulties which meant that the number of loci amplified varied (Table 2; see discussion).

The relationship between SNP and MIRU-VNTR differences is shown in Table 2 and Figure 4C. There was a clear MIRU-VNTR difference between those labelled relapses using WGS (0-2 MIRU-VNTR differences) and those that were labelled as reinfections (7-8 MIRU-VNTR differences). However, within the relapse group, there was no obvious relationship between these two measures: all samples with 2-5 SNPs had no MIRU-VNTR differences, whereas there were four with no SNP differences and one MIRU-VNTR difference. Overall WGS largely agreed with MIRU-VNTR (Table 3), with only the likely mixed infection causing a possible discrepancy. That was based on a decision in the trial to classify pairs with two or more MIRU-VNTR differences as reinfections.

In silico spoligotyping and sub-lineages

Human *M. tuberculosis* strains have been divided into six global lineages, and further into sub-lineages, some of which may have distinct infection phenotypes (24). In addition to the whole genome SNP-based methodology used above, analysis using a set of 62 lineage-defining SNPs (24) were also used to assign sub-lineages (Supplementary Table S1). The three reinfections observed all involved different sub-lineages in the pair (patient 004 Euro-American LAM → Euro-American S type; patient 009 Euro-American S-type → East Asian; patient 015 Euro-American T → East Asian).

In silico spoligotyping was also performed (Supplementary Table S1). 24/32 relapse pairs had identical spoligotypes, the remaining 8 having 1-7 spacer differences; all three reinfections had different spoligotypes (9-29 spacer differences).

Antimicrobial resistance

Drug susceptibility testing showed that only one isolate (004-2) had a drug resistance phenotype, confirmed by genotype (RIF^R: *rpoB* S450L; INH^R: *katG* S315T; EMB^R: *embB* M306V), while the pre-treatment isolate (004-1) was susceptible to all drugs tested. Therefore, there was no evidence of any acquisition of antibiotic resistance during the trial in the samples that were tested with WGS.

SNPs in relapse isolates

While it would be expected that most SNPs that arise in a strain between treatment and relapse to be random, as long as they are not deleterious, it would be a reasonable hypothesis that some SNPs may actively help the bacteria survive. Comparing the relapse pairs, 18/30 SNPs were synonymous and 12/30 non-synonymous (Table 4). Of the 12 non-synonymous SNPs and two INDELS, none were in a gene associated with antibiotic resistance, in accord with the fact that no phenotypic resistance was seen. However, two SNPs lay in genes that

are implicated in pathogenesis, both associated with *esx* Type 7 secretion systems (T7SSs (25)) (discussed below).

Discussion

Relapse v reinfection

In this study, high quality genome sequence was generated for 36 pairs of isolates. The majority of pairs (32/36) were shown to have very few SNPs (≤ 5) between pre- and post-treatment *M. tuberculosis* isolates, suggestive of relapse and thus treatment failure.

On initial inspection, the other four pairs (4/36) had significant SNP differences between samples (57, 737, and two >1000), indicative of reinfection. However, phylogenetic analyses cast doubt on two pairs, where a single isolate of each pair was highly related to another patient's isolate in the study. While it is possible that these reflect transmission events, it is difficult to rule out some form of laboratory processing error; indeed, a transmission event so similar to another pair of samples in the trial (in one case the pre-treatment and in the other the post-treatment samples) would be relatively uncommon though not impossible, but such a pattern would be expected if there were a sample processing error and patient samples were swapped. A similar event was suggested by Casali *et al* (26). Indeed, trials inserting negative samples into the TB diagnostic process showed that errors can occur (27), but strain typing methods allowed actual contamination to be detected. A review by Burman *et al* (28) indicated a median false-positive rate of 3.1% in published studies. WGS can thus help identify when processing errors have occurred, thereby improving overall trial data quality and acting as a quality control measure of trial procedures.

The case with 57 SNPs between the isolate pair was probably a mixed infection, and while accurate SNP figures could not be obtained, the data were consistent with a relapse from one of the two pre-existing strains. These are described as a major/minor strain within the

sequencing data, but that may not accurately reflect the relative levels in the patient; it could, for example, be affected by colony size on the LJ slopes, and the actual loop sample taken for DNA preparation. The isolate pair were initially identified as being different from each other at a number of SNP differences (57) higher than would be expected for a relapse, but at an unusually low level of SNPs for a reinfection compared to other reported examples. This is likely to be due to the mixed infection causing many genuine SNPs to be discarded as uncertain by the site-calling algorithm. Reports of similar cases of mixed infections in previous studies (14,15,29) support the likelihood that it may be genuine, thus suggesting that it is important to assess isolates for evidence of mixed infections before calling relapse/reinfection.

Therefore, from the 36 pairs of isolates sequenced, there was strong evidence that 32 were relapses, one was a mixed infection masking a likely relapse, and three were reinfections, although two of these may have been the result of laboratory processing errors. This proportion (32/35 (91%) relapse: 3/35 (9%) reinfection; excluding the possible mixed infection) compares to 92:8 (14), which was also a chemotherapy trial, and 73:27 (15) in the rather different situation of a long-term study, and longer post-treatment follow-up (over 12 years in some cases). This latter study indicated that relapses occurred towards the start of the follow-up, and particularly within the first two years, and therefore is consistent with the study reported here.

SNP differences in this and previous studies

The number of SNP differences in the relapse and reinfection groups were comparable to previous pre- and post-treatment studies (Table 5). Casali *et al* (26) also found up to 4 SNP differences over 4 years in intra-patient studies. In each of the previous relapse studies, there was a large gap between the number of SNPs in presumed relapses and reinfections. This both lends support to their definition as relapse or reinfection, and also gives weight to the suggestion by Bryant *et al* (14) that there is some immunity to reinfection by very similar

strains. The same pattern was observed in this study, even though the phylogenetic tree showed that highly similar strains were circulating. Guerra-Assunção *et al* (15) showed less SNP diversity in reinfections (100 rather than 1000 SNPs), and it would be interesting to determine if there was an effect of time, with similar strains only reinfecting after a longer passage of time. Casali *et al* (26) demonstrated that there is strain diversity within a single sputum specimen, with up to 10 SNP differences seen when individual colonies were sequenced. The methodology described in this study deliberately took a sweep of colonies which meant that much of this strain diversity within a single specimen would not be seen in WGS at the depth of coverage used.

SNPs seen in relapse isolates

For 16 of the 32 relapse pairs sequenced, SNPs were identified between the isolates (Table 4; excluding the mixed infection). While it is likely that many or most of these will not be advantageous to the bacteria, it is a plausible hypothesis that some of them might have a survival advantage.

Of the 12 non-synonymous SNPs observed in relapse isolate pairs, two were in gene systems that have proven involvement with pathogenesis: the two T7SSs. *esx1* and *esx3*. One lay in *eccB3*, which is a gene in the ESX3 T7SS, and is essential for growth. This system is involved in pathogenesis, partly through control of iron acquisition that appears to have a role in metal homeostasis (30). The other was located in *mce1B*, which is a gene in the ESX1 T7SS, which is essential for virulence and exports the well characterized ESAT-6/CFP10 complex (25). The previous study by Bryant *et al* (14) reported two genes with SNPs had functions associated with oxidative stress, and that of Guerra-Assunção (15) reported an association with *katG*, well known both as being involved in resistance to oxidative stress, and also isoniazid resistance. Clearly these may just be chance associations, but they also indicate potential avenues for studying bacterial survival during chemotherapy. The scale of investment in

Phase 2 and 3 trials is such that there is an obligation to extract as much information as possible from the study and the contribution of WGS is fundamental to understanding the bacteriology under treatment.

Mixed infections

A potential confounder in differentiating relapse from reinfection is that of mixed infections. If either the initial or subsequent infection is mixed, then sampling just one isolate could give a misleading designation. One likely mixed infection was identified with a 75:25 genotype ratio, although this ratio may not represent the ratio of the mixture in the bacterial population *in vivo*.

Of course, these methods would only reveal mixed infections with significant proportions of each strain, and it cannot formally exclude the possibility that other infections were also mixed, but at a very low levels. Bryant *et al* (14), Guerra-Assunção *et al* (15), Casali *et al* (26) and Köser *et al* (29) all identified mixed infections using WGS. Other studies have demonstrated them using alternative techniques, including MIRU-VNTR (31–34), but WGS is more powerful, and Bryant (14) found that WGS detected more mixed infections than MIRU-VNTR.

The definition of a mixed infection is less clear with the finding that at least 10 SNP differences can be found within a single sputum sample (26), and the observation that very similar strains circulate in high prevalence settings (e.g. Figure 2). However, the data here and in the previous relapse studies (14) suggest that some sort of immunological protection might exist that makes successful co-infection with a similar strain less likely.

Comparing WGS to MIRU-VNTR and spoligotyping

Previously, due to its speed and digital output, MIRU-VNTR has been preferred to the earlier IS6110 profiling as a means of typing *M. tuberculosis* isolates; indeed it was only recently described as “the new reference standard for molecular epidemiological studies” (35).

In this study there was a correlation between SNP and MIRU-VNTR differences for isolates predicted to be relapses (0-5 SNPs; 0-2 MIRU-VNTR loci) and reinfections (SNP > 1000; 7+ MIRU-VNTR loci) cases, in contrast with the study of Bryant *et al* (14) where three reinfection pairs had 1-13 loci different, although the Bryant study was an interim analysis performed prior to final data resolution and unbinding which may have impacted on the ultimate assignment of the patients. Furthermore, Casali *et al* (26) found that two MIRU-VNTR differences could correspond to a significant number of SNPs. A transmission study by Walker (11) only examined isolates with successful 24 locus MIRU-VNTR data, showing that up to a difference of 100 SNPs, isolates could have 1-3 MIRU-VNTR locus differences, while above 100 SNP differences, the number of MIRU-VNTR changes increased.

MIRU-VNTR, which involves 24 multiplexed PCRs, is known to be technically challenging to achieve consistent results (14,26,36,37). Indeed there was significant variation in the number of loci amplified in this study (Table 2, Figure 4D), which we attribute to a combination of DNA quantity, quality and the technical difficulties referred to above. Furthermore, other limitations and issues with MIRU-VNTR in relation to the study setting have been discussed in a systematic review (38).

WGS is technically more straightforward and comparable in cost in our hands (~£100 per sample), but with reducing costs and whole genome resolution, it is clearly a superior, more robust method than MIRU-VNTR for strain typing. In addition, WGS can provide added information in the form of identification of markers associated with drug resistance, which could be useful in the context of relapsing cases in a clinical trial. Sequence data is also more amenable to incorporation into other studies and will provide further information on TB evolution as global databases of genome information grow.

Spoligotyping has been widely used for robust division of *M. tuberculosis* into different subtypes (4), but we found that not only were SNPs far more sensitive for determining relapses/reinfections, but they were also more useful for assigning sub-lineages.

Value of WGS in chemotherapy trials

The data from this study in combination with the previous two relapse studies (14,15) allow an evaluation of the relative benefit of using WGS or MIRU-VNTR as a means of determining relapses from reinfections in chemotherapy trials.

The RIFAQUIN trial was in an area of high endemicity, suggesting that reinfections are not likely to be higher elsewhere due to disease prevalence. Thus the data presented in this study and previously (14,15) indicate that the proportion of reinfections is very low compared to relapse, although Guerra-Assunção *et al* (15) suggest that reinfections may rise at later time points after completion of therapy. Furthermore, decisions where isolates are reinfections are more likely to be wrong, as possible errors observed here (processing errors, unrecognized mixed infections) are more likely to suggest a reinfection.

Conclusions

In the pre-genomic era, all post-treatment infections were presumed to be relapses, and it could be argued that due to the low reinfection rates, the increased cost and time required to perform WGS provides only modest gains for the analysis of the primary outcome in a chemotherapy clinical trial of this nature.

Nevertheless, in addition to robust genomic evidence for treatment outcome, the added information that WGS provides is scientifically valuable and will become of greater value as more genome sequence data and more information about genotype-phenotype correlation and its impact on disease and transmission becomes available. Furthermore, future trials for new TB drugs in the development pipeline or novel combination regimens may be in areas of

high TB prevalence where re-infection or mixed infections are more likely, thus making accurate strain discrimination an imperative and WGS should be the method of choice.

Declarations

Ethics approval and consent to participate

The study protocol was reviewed and approved by the ethics review committee at St. George's by medical ethics and regulatory committees representing each of the participating countries, and by the institutional review board of the Centers for Disease Control and Prevention operating in Botswana.

Consent for publication

Not applicable.

Availability of data

Sequence data has been submitted to the ENA database with accession number PRJEB18529. The full analysis pipeline can be downloaded and run from <http://github.com/bugs-bioinf/rifaquin-2016>.

Competing Interests

The authors declare that they have no competing interests.

Funding

The RIFAQUIN trial was funded by the European and Developing Countries Clinical Trials Partnership and the Wellcome Trust; RIFAQUIN Current Controlled Trials number, ISRCTN44153044.

Author Contributions

AW, AB, PP, NS performed the data analysis, DC cultured the isolates, AJ, PB and TM designed the study. All authors contributed to the writing of the manuscript.

Acknowledgements

MIRU-VNTR analyses were carried out by Selina Bannoo, Emma Cunningham, Alice Morgan, Solomon Mwaigwishya and Laura Wright. Sequencing was carried out at UCL Genomics by Tony Brooks and Nipurna Jina.

References

1. Fox W, Ellard GA, Mitchison DA. Studies on the treatment of tuberculosis undertaken by the British Medical Research Council tuberculosis units, 1946-1986, with relevant subsequent publications. *Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis*. 1999 Oct;3(10 Suppl 2):S231-279.
2. Hermans PW, van Soolingen D, Dale JW, Schuitema AR, McAdam RA, Catty D, et al. Insertion element IS986 from *Mycobacterium tuberculosis*: a useful tool for diagnosis and epidemiology of tuberculosis. *J Clin Microbiol*. 1990 Sep;28(9):2051-8.
3. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, et al. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. *J Clin Microbiol*. 2006 Dec;44(12):4498-510.
4. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol*. 1997 Apr;35(4):907-14.
5. Das S, Chan SL, Allen BW, Mitchison DA, Lowrie DB. Application of DNA fingerprinting with IS986 to sequential mycobacterial isolates obtained from pulmonary tuberculosis patients in Hong Kong before, during and after short-course chemotherapy. *Tuber Lung Dis*. 1993 Feb 1;74(1):47-51.
6. Tam CM, Chan SL, Kam KM, Sim E, Staples D, Sole KM, et al. Rifapentine and isoniazid in the continuation phase of a 6-month regimen. Interim report: no activity of isoniazid in the continuation phase. *Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis*. 2000 Mar;4(3):262-7.
7. Benator D, Bhattacharya M, Bozeman L, Burman W, Cantazaro A, Chaisson R, et al. Rifapentine and isoniazid once a week versus rifampicin and isoniazid twice a week for treatment of drug-susceptible pulmonary tuberculosis in HIV-negative patients: a randomised clinical trial. *Lancet Lond Engl*. 2002 Aug 17;360(9332):528-34.
8. Lienhardt C, Cook SV, Burgos M, Yorke-Edwards V, Rigouts L, Anyo G, et al. Efficacy and safety of a 4-drug fixed-dose combination regimen compared with separate drugs for treatment of pulmonary tuberculosis: the Study C randomized controlled trial. *JAMA*. 2011 Apr 13;305(14):1415-23.
9. Jindani A, Nunn AJ, Enarson DA. Two 8-month regimens of chemotherapy for treatment of newly diagnosed pulmonary tuberculosis: international multicentre randomised trial. *Lancet Lond Engl*. 2004 Oct 2;364(9441):1244-51.
10. Walker TM, Monk P, Grace Smith E, Peto TEA. Contact investigations for outbreaks of *Mycobacterium tuberculosis*: advances through whole genome sequencing. *Clin Microbiol Infect*. 2013 Sep;19(9):796-802.

11. Walker TM, Lalor MK, Broda A, Saldana Ortega L, Morgan M, Parker L, et al. Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. *Lancet Respir Med*. 2014 Apr;2(4):285–92.
12. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis*. 2013;13(2):137–146.
13. Satta G, Witney AA, Shorten RJ, Karlikowska M, Lipman M, McHugh TD. Genetic variation in *Mycobacterium tuberculosis* isolates from a London outbreak associated with isoniazid resistance. *BMC Med*. 2016 Aug 16;14(1):117.
14. Bryant JM, Harris SR, Parkhill J, Dawson R, Diacon AH, van Helden P, et al. Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study. *Lancet Respir Med*. 2013 Dec;1(10):786–92.
15. Guerra-Assunção J, Crampin A, Houben R, Mzembe T, Mallard K, Coll F, et al. Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *eLife* [Internet]. [cited 2016 Nov 4];4. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4384740/>
16. Gillespie SH, Crook AM, McHugh TD, Mendel CM, Meredith SK, Murray SR, et al. Four-Month Moxifloxacin-Based Regimens for Drug-Sensitive Tuberculosis. *N Engl J Med*. 2014 Oct 23;371(17):1577–87.
17. Jindani A, Harrison TS, Nunn AJ, Phillips PPJ, Churchyard GJ, Charalambous S, et al. High-Dose Rifapentine with Moxifloxacin for Pulmonary Tuberculosis. *N Engl J Med*. 2014 Oct 23;371(17):1599–608.
18. Kent L, McHugh TD, Billington O, Dale JW, Gillespie SH. Demonstration of homology between IS6110 of *Mycobacterium tuberculosis* and DNAs of other *Mycobacterium* spp.? *J Clin Microbiol*. 1995 Sep;33(9):2290–3.
19. Kozarewa I, Turner DJ. 96-plex molecular barcoding for the Illumina Genome Analyzer. *Methods Mol Biol Clifton NJ*. 2011;733:279–98.
20. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* [Internet]. 2013 Mar 16 [cited 2014 Jul 7]; Available from: <http://arxiv.org/abs/1303.3997>
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009 Aug 15;25(16):2078–9.
22. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinforma Oxf Engl*. 2014 May 1;30(9):1312–3.
23. Coll F, Mallard K, Preston MD, Bentley S, Parkhill J, McNerney R, et al. SpolPred: rapid and accurate prediction of *Mycobacterium tuberculosis* spoligotypes from short genomic sequences. *Bioinformatics*. 2012 Nov 15;28(22):2991–3.

24. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, et al. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat Commun*. 2014 Sep 1;5:4812.
25. Majlessi L, Prados-Rosales R, Casadevall A, Brosch R. Release of mycobacterial antigens. *Immunol Rev*. 2015 Mar 1;264(1):25–45.
26. Casali N, Broda A, Harris SR, Parkhill J, Brown T, Drobniewski F. Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study. *PLOS Med*. 2016 Oct 4;13(10):e1002137.
27. Aber VR, Allen BW, Mitchison DA, Ayuma P, Edwards EA, Keyes AB. Quality control in tuberculosis bacteriology. 1. Laboratory studies on isolated positive cultures and the efficiency of direct smear examination. *Tubercle*. 1980 Sep;61(3):123–33.
28. Burman WJ, Reves RR. Review of False-Positive Cultures for *Mycobacterium tuberculosis* and Recommendations for Avoiding Unnecessary Treatment. *Clin Infect Dis*. 2000 Dec 1;31(6):1390–5.
29. Koser C, M. Bryant J, Becq J, Torok ME, Ellington MJ, Marti-Renom MA, et al. Whole-Genome Sequencing for Rapid Susceptibility Testing of *M. tuberculosis*. *N Engl J Med*. 2013 Jul 18;369(3):290–2.
30. Tufariello JM, Chapman JR, Kerantzas CA, Wong K-W, Vilchèze C, Jones CM, et al. Separable roles for *Mycobacterium tuberculosis* ESX-3 effectors in iron acquisition and virulence. *Proc Natl Acad Sci*. 2016 Jan 19;113(3):E348–57.
31. Hanekom M, Streicher EM, Berg DV de, Cox H, McDermid C, Bosman M, et al. Population Structure of Mixed *Mycobacterium tuberculosis* Infection Is Strain Genotype and Culture Medium Dependent. *PLOS ONE*. 2013 Jul 30;8(7):e70178.
32. Fang R, Li X, Li J, Wu J, Shen X, Gui X, et al. Mixed infections of *Mycobacterium tuberculosis* in tuberculosis patients in Shanghai, China. *Tuberculosis*. 2008 Sep;88(5):469–73.
33. Mallard K, McNerney R, Crampin AC, Houben R, Ndlovu R, Munthali L, et al. Molecular Detection of Mixed Infections of *Mycobacterium tuberculosis* Strains in Sputum Samples from Patients in Karonga District, Malawi. *J Clin Microbiol*. 2010 Dec;48(12):4512–8.
34. Cohen T, Wilson D, Wallengren K, Samuel EY, Murray M. Mixed-Strain *Mycobacterium tuberculosis* Infections among Patients Dying in a Hospital in KwaZulu-Natal, South Africa. *J Clin Microbiol*. 2011 Jan 1;49(1):385–8.
35. Brossier F, Sola C, Millot G, Jarlier V, Veziris N, Sougakoff W. Comparison of a semiautomated commercial repetitive-sequence-based PCR method with spoligotyping, 24-locus mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing, and restriction fragment length polymorphism-based analysis of IS6110 for *Mycobacterium tuberculosis* typing. *J Clin Microbiol*. 2014 Nov;52(11):4082–6.
36. Cowan LS, Mosher L, Diem L, Massey JP, Crawford JT. Variable-Number Tandem Repeat Typing of *Mycobacterium tuberculosis* Isolates with Low Copy Numbers of IS6110 by Using Mycobacterial Interspersed Repetitive Units. *J Clin Microbiol*. 2002 May;40(5):1592–602.

37. Chatterjee A, Mistry N. MIRU–VNTR profiles of three major Mycobacterium tuberculosis spoligotypes found in western India. *Tuberculosis*. 2013 Mar;93(2):250–6.
38. Mears J, Abubakar I, Cohen T, McHugh TD, Sonnenberg P. Effect of study design and setting on tuberculosis clustering estimates using Mycobacterial Interspersed Repetitive Units-Variable Number Tandem Repeats (MIRU-VNTR): a systematic review. *BMJ Open*. 2015 Jan 21;5(1):e005636.

Table 1: Comparison between SNP and MIRU differences

SNP diffs	Number of isolate pairs	MIRU diffs
0	19	0 (n=15) ^a 1 (n=4)
1	7	0 (n=6) ^b , 2 (n=1)
2	1	0
3	2	0
5	3	0 ^c
57 ^e	1	2
737	1	6
1233	1	0 ^d
1294	1	7

^a two samples <10 loci (2, 7);

^b one from only 2 informative loci

^c from only 5 informative loci

^d from only 3 informative loci

^e excluded from further SNP analysis as found to be mixed infection, but re-interpreted as a relapse (see text)

Table 2. Relationship between SNP and MIRU-VNTR differences. The 36 isolates for which WGS was carried out are listed. With the mixed infection, although we concluded it to be a relapse, we could not precisely determine the SNP difference. Treatment arms: Control (2EHRZ/4HR); 4 month (2EMRZ/2P2M2); 6 month (2EMRZ/4P1M1) (for details, see Jindani *et al* (17)). The table is sorted (in order) by number of SNPs, MIRU-VNTR differences, treatment arm, and study number. Isolate 004-2 had previously been shown by DST to be resistant to isoniazid, rifampicin, ethambutol, streptomycin and pyrazinamide, however all other isolates had been determined to be susceptible (17).

Study no.	Location	Treatment arm	SNPs	MIRU-VNTR differences	MIRU-VNTR loci amplified	Prediction
001	Francistown	4 month	0	0	21	Relapse
003	Harare	4 month	0	0	14	Relapse
005	Harare	4 month	0	0	20	Relapse
007	Harare	4 month	0	0	7	Relapse
008	Harare	4 month	0	0	11	Relapse
013	Marondera	4 month	0	0	10	Relapse
014	Marondera	4 month	0	0	11	Relapse
016	Johannesburg	4 month	0	0	14	Relapse
020	Johannesburg	4 month	0	0	17	Relapse
023	Cape Town	4 month	0	0	15	Relapse
029	Cape Town	4 month	0	0	2	Relapse
030	Cape Town	4 month	0	0	15	Relapse
032	Cape Town	4 month	0	0	14	Relapse
017	Johannesburg	6 month	0	0	17	Relapse
034	Cape Town	6 month	0	0	21	Relapse
037	Cape Town	Control	0	1	17	Relapse
011	Harare	4 month	0	1	16	Relapse
021	Johannesburg	4 month	0	1	16	Relapse
028	Cape Town	4 month	0	1	15	Relapse
024	Cape Town	Control	1	0	18	Relapse
033	Cape Town	Control	1	0	15	Relapse
010	Harare	4 month	1	0	18	Relapse
012	Harare	4 month	1	0	19	Relapse
025	Cape Town	4 month	1	0	11	Relapse
027	Cape Town	6 month	1	0	13	Relapse

019	Johannesburg	Control	1	2	4	Relapse
026	Cape Town	4 month	2	0	18	Relapse
002	Harare	4 month	3	0	15	Relapse
006	Harare	4 month	3	0	12	Relapse
018	Johannesburg	Control	5	0	5	Relapse
036	Cape Town	4 month	5	0	17	Relapse
031	Cape Town	6 month	5	0	16	Relapse
015	Johannesburg	Control	1294	7	14	Reinfection
035	Cape Town	4 month	57*	-	-	Relapse
004	Harare	Control	737	6	6	Reinfection
009	Harare	4 month	1233	3	3	Reinfection

* It was not possible to separate the mixed genotypes to precisely determine a SNP difference.

Table 3: Comparison of the use of WGS with MIRU-VNTR for calling relapse or reinfection.

	MIRU-VNTR	WGS
Relapse	32	33
Reinfection	4	3

Table 4: Variants identified in relapse pairs. SNPs between individual pairs predicted to be relapse. S: synonymous SNP; NS: non-synonymous SNP; CHP: conserved hypothetical protein; CP: conserved protein. Function assigned using the Tuberculist database (<http://tuberculist.epfl.ch/>)

Strain pair	Type	Base number *	Gene	Function
002	NS	146316	<i>Rv0120c, fusA2</i>	Translation
	NS	345226	<i>Rv0283, eccB3</i>	Part of ESX-3 (essential, <u>ESX-3 T7SS is implicated in metal homeostasis</u>)
	S	3135592	<i>Rv2827c -109</i>	
	INDEL (TC/TCC)	3600992	<i>Rv3224B</i>	Predicted membrane protein
006	S	1348678	<i>Rv1205 -41</i>	
	S	1370403	<i>Rv1227c</i>	
	S	2828233	<i>hisT</i> down	
010	NS	200390	<i>Rv0170, mce1B</i>	Part of ESX-1, essential for pathogenesis
012	NS	2510502	<i>Rv2237A</i>	CP, non-essential
017	INDEL (GC/GCC)	341124	<i>Rv0281</i>	Possible membrane protein
018	S	783720	<i>fusA1</i>	
	S	783729	<i>fusA1</i>	
	S	783732	<i>fusA1</i>	
	S	1476666	<i>rrl</i>	
	S	4050367	<i>folE</i>	

019	NS	3884906	<i>Rv3467</i>	CHP, non-essential
024	S	848538	<i>PPE12</i>	
025	S	1929374	<i>Rv1703c</i>	
026	NS	1192723	<i>Rv1069c</i>	CP, non-essential
	NS	1690758	<i>Rv1499</i>	CHP, non-essential
027	S	114494	<i>nrp</i>	
031	S	175753	<i>Rv0149</i>	
	S	620981	<i>Rv0530</i>	
	S	1315992	<i>pks4</i>	
	NS	1540497	<i>Rv1367c</i>	CP, non-essential
	S	2788333	<i>plsB2</i>	
033	NS	3618159	<i>Rv3240c, secA1</i>	Protein export, essential
036	S	923816	<i>lysT</i>	
	NS	924229	<i>Rvnt13, pheU</i>	tRNA
	NS	924234	<i>Rvnt13, pheU</i>	tRNA
	S	924263	<i>pheU</i>	
	NS	1476973	<i>Rvnr03, rrf</i>	5S rRNA

*

Table 5: Number of SNP differences between relapse/reinfection paired samples in different studies

Relapse group	Reinfection group	Maximum length of follow up	Study
0-5	>1000	18 m	This study
0-6	>1300	18 m	Bryant <i>et al</i> (14)
0-8	>100	>12 y	Guerra-Assunção <i>et al</i> (15)

Figure 1: Flowchart of pairs of samples studied

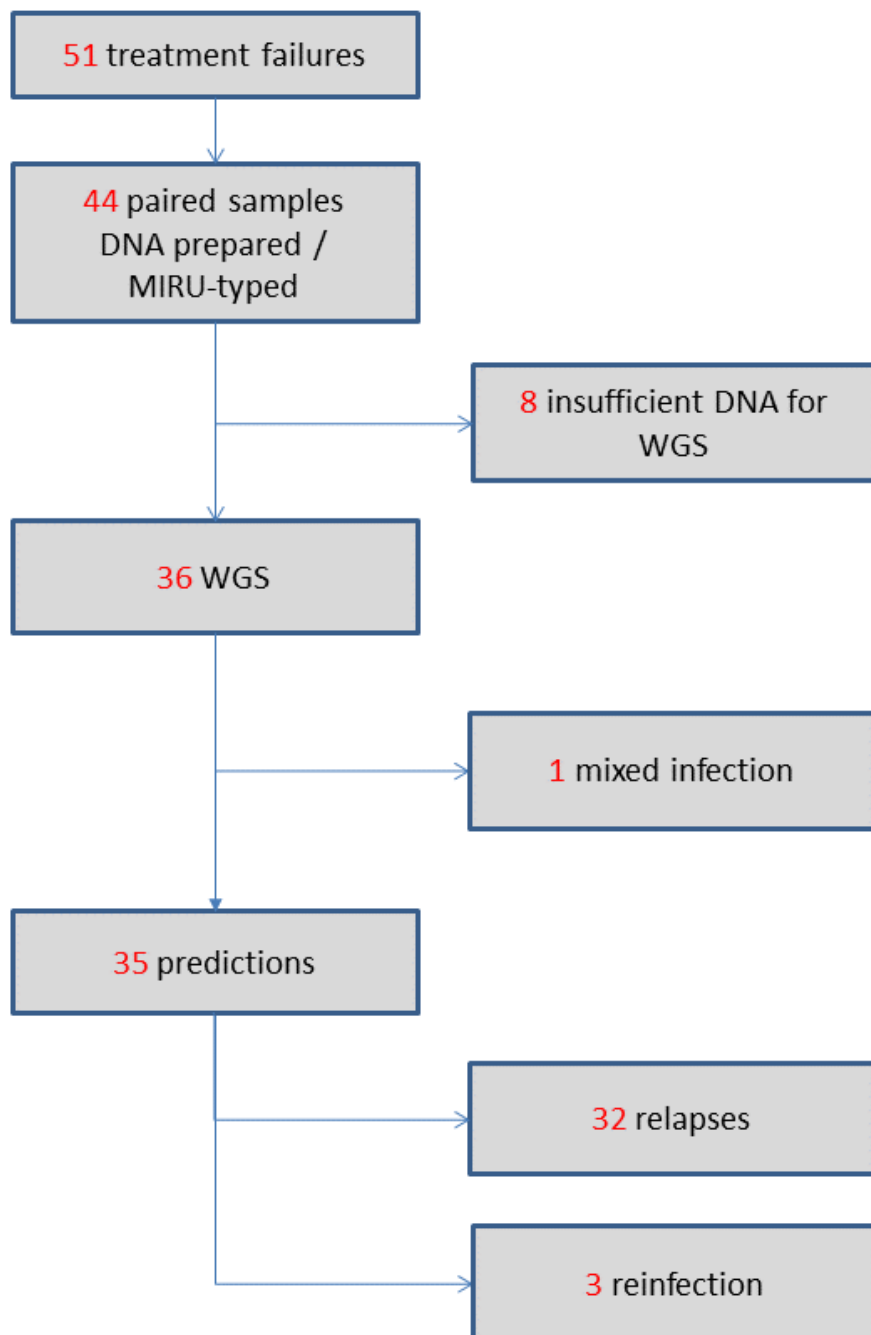
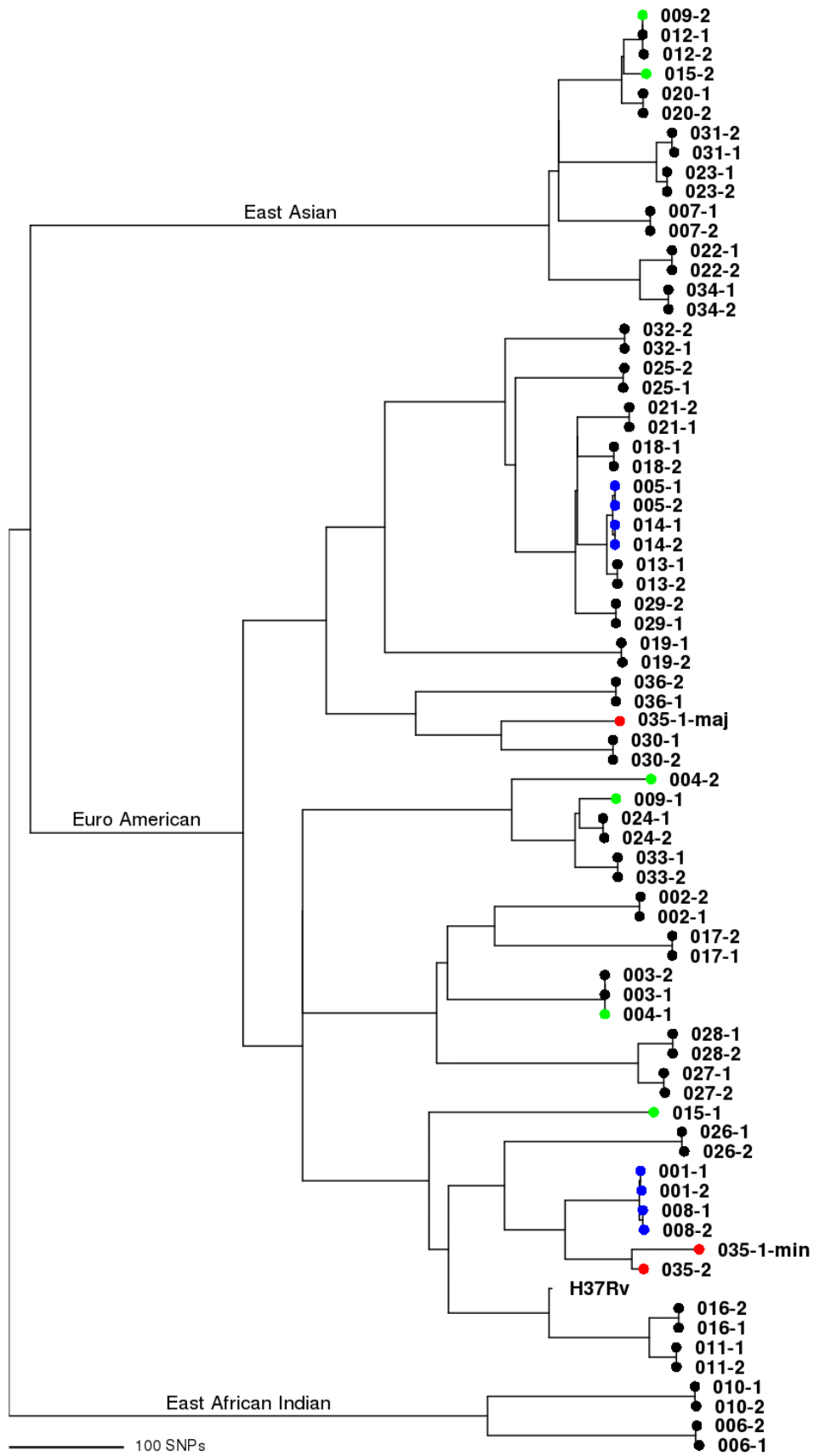


Figure 2: A. Phylogenetic reconstruction of 36 pairs of isolates. (Inferred using 5132 high quality SNPs following the removal of 661,083 low quality sites and the remaining invariant sites. The tree was rooted using the H37Rv reference strain sequence). Relapse, reinfection, and mixed are denoted with black/blue, green and red tips respectively. Blue tip labels are further shown in panels B-E. B-E, branches amplified where unexpected similarity seen; numbers of SNPs between the most divergent samples is given.







B (11 SNPs)	C (4 SNPs)	D (0 SNP)	E (0 SNPs)
 <p>005-1 005-2 014-2 014-1</p>	 <p>008-1 008-2 001-1 001-2</p>	 <p>009-2 012-1 012-2</p>	 <p>003-1 003-2 004-1</p>

Figure 3: Identification of mixed infection. A: Counts of genome sites which are called as a reference base but show a significant proportion of sequence reads also supporting a variant base call (035-1); B: the equivalent plot for an isolate with no mixed infection (035-2). The presence of a second peak in A is suggestive of a mixture with a minority genotype.

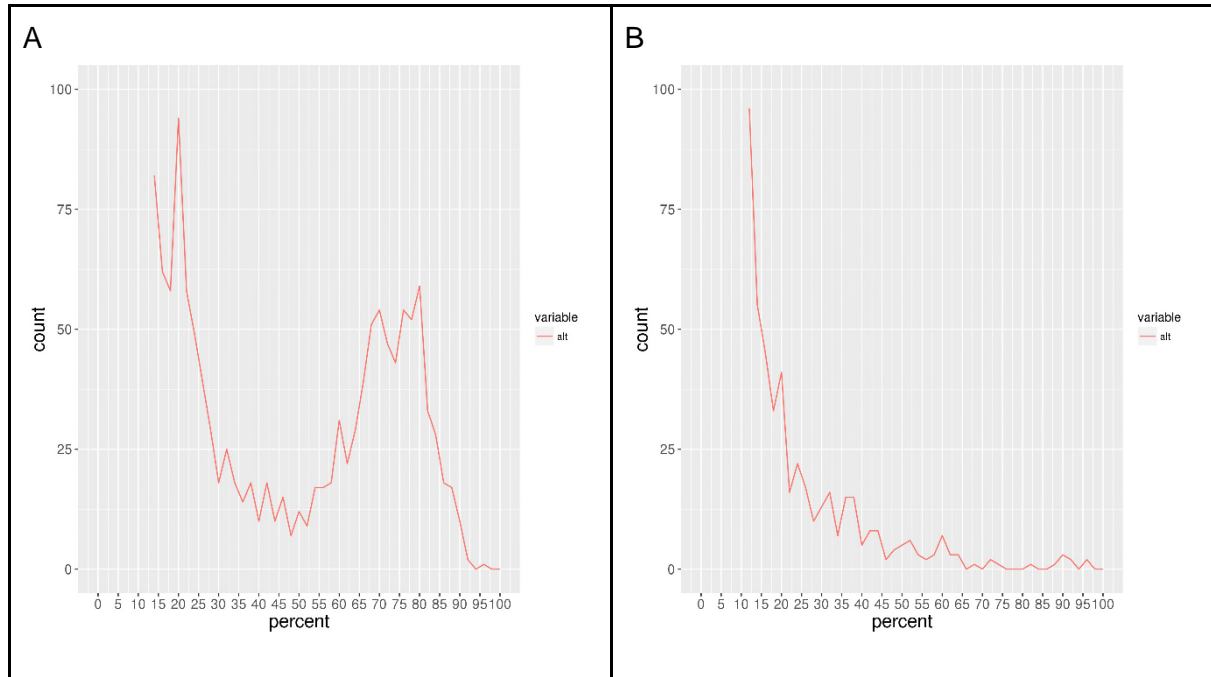


Figure 4: Analysis of SNP and MIRU-VNTR differences between pairs of isolates.

Data is summarised from Tables 1 and 2. A: Number of SNP differences detected between paired isolates; B: Number of MIRU-VNTR differences detected between paired isolates; C: Correlation between SNPs and MIRU differences; D: Number of informative MIRU loci on which differences were based (for each pair of samples, the lower number is shown).

